UNCOVERING HIDDEN PATTERNS OF MOLECULAR RECOGNITION

By

Sebastian Raschka

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biochemistry and Molecular Biology – Doctor of Philosophy
Quantitative Biology – Dual Major

2017

ABSTRACT

UNCOVERING HIDDEN PATTERNS OF MOLECULAR RECOGNITION

By

Sebastian Raschka

It happened in 1958 that John Kendrew's group determined the three-dimensional structure of myoglobin at a resolution of 6 Å. This first view of a protein fold was a breakthrough at that time. Now, more than half a century later, both experimental and computational techniques have substantially improved as well as our understanding of how proteins and ligands interact. Yet, there are many unanswered questions to be addressed and patterns to be uncovered. One of the most pressing needs in structural biology is the prediction of protein-ligand complexes in aiding inhibitor and drug discovery, ligand design, and studies of catalytic mechanisms. Throughout the past few decades, improvements in computational technologies and insights from experimental data have converged into numerous protein-ligand docking and scoring algorithms. However, these methods are still far from being perfect, and only minimal improvements have been made in the past few years. That might be because current scoring functions regard individual intermolecular interactions as independent events in a binding interface.

This thesis addresses existing shortcomings in the conventional view of protein-ligand recognition by characterizing interactions as patterns. Finding that binding rigidifies protein-ligand complexes has led to our design of a robust scoring function that predicts native protein-ligand complexes through the coupling of interactions that rigidifies the protein-ligand interface. Also, the analysis of a non-homologous set of protein-ligand complexes has revealed that binding interfaces are polarized – surprisingly, proteins donate twice as many hydrogen bonds to ligands as they accept, on average, and the opposite is true for ligands. A more in-depth analysis of atom type distributions among H-bond donor and acceptor atoms showed that the discovered trends contain surprisingly strong patterns that are also predictive of native protein-ligand binding. Both the coupling of interactions as well as the distribution of hydrogen bond patterns are currently not

captured by other methods and provide new information for the prediction and design of ligands.

In the absence of the protein receptor structure, our results show that data from experimental assays can be mined to identify functional group patterns on ligands that are predictive of biological activity. Additionally, we present methods to use functional group patterns to improve the success rate of ligand-based virtual screening. Applied to G protein-coupled receptor inhibitor discovery, this approach has led to the discovery of a potent inhibitor that nullifies the biological response and presents the first instance where virtual screening has been used for aquatic invasive species control. Finally, to overcome current challenges in drug discovery for protein-protein interfaces, a new method for identifying small molecules that block protein-protein interactions is presented. We developed and applied an epitope-based virtual screening workflow to find inhibitors of focal adhesion kinase interactions involved in cancer metastasis.

In sum, this work presents both novel insights into the coupling among and trends in intermolecular interactions as well as methods to predict the biological activity of ligands based on patterns of functional groups. Along with the insights gained in this work, computational tools and software for measuring the rigidification that is characteristic of native protein-ligand complexes, analyzing H-bond patterns rigorously, and screening millions of small molecules in hypothesis-driven ligand discovery have been developed and are now being made available to other scientists.

This thesis is dedicated to my mom, dad, and sister. Thank you for always believing in me.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Leslie A. Kuhn, for all her guidance throughout my Ph.D. studies and the opportunity to work on many interesting and exciting projects in her lab. My Ph.D. studies have been the biggest adventure of my life thus far – something I will always remember as one of the most profound experiences of my life.

I would also like to thank my guidance committee, Dr. Michael Feig, Dr. David N. Arnosti, Dr. C. Titus Brown, and Dr. Jian Hu. Having such a great committee really helped me with staying on track throughout the last five years, and I appreciate all the helpful advice, constructive critique, and encouraging words.

Special thanks go to Jessica Lawrence, Becky Conat Mansel, and Jeannine Lee, who helped me navigate through grad school. Knowing that someone was always there to help with various forms and paperwork was invaluable. I would also like to say thanks to Dr. John J. LaPres, Dr. Jon M. Kaguni, and Dr. R. Michael Garavito, who recruited me into the BMS, BMB, and Quantitative Biology programs, and who have done a remarkable job making those programs a very worthwhile experience.

A warm thanks goes to Vahid Mirjalili for being such a good friend throughout these years at MSU: exploring the Michigan rivers on kayaks, traveling to conferences, being witness to my friend's wedding, and collaborating on countless different projects – it was always fun!

I am grateful to my grandparents, uncle, sister, mom, and dad, who have always been there for me with moral support. Although I am sad and very sorry that I missed countless birthday parties and other celebrations, I am very grateful to have such a great family who always understood and always believed in me.

Last but not least, I would like to thank all my friends and fellow students at MSU. Thank you for five amazing years!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Molecular recognition is a keystone of biological function. Understanding the biochemical and biological roles of protein and ligand binding plays an important role in elucidating cellular processes and applications such as therapeutic drug design. Suffice it to say that protein-ligand complexes have been extensively studied using experimental as well as computational techniques. And yet no method exists that can predict protein-ligand binding or model the structures of protein-ligand complexes with high accuracy across vastly different protein and ligand families.

The objective of this dissertation is to derive new insights into principles governing molecular recognition, and the novel contributions of the computational studies presented in this work are the following:

Chapter 2. The objective of this chapter was to test a hypothesis that arose from observations made throughout different inhibitor discovery projects: "proteins favor donating H-bonds to ligands and avoid using groups with both H-bond donor and acceptor capacity." The analysis of a large set of non-homologous proteins bound to their biological ligands revealed strong patterns of chemical group matching preferences for intermolecular hydrogen bonding. In particular, proteins donate twice as many hydrogen bonds than they accept on average, and as a consequence, the opposite is true for their biological ligands. The fact that protein-ligand interactions are dominated by the presence of hydrogen-bond donors may provide the geometric directionality for specificity upon which protein-ligand complexes are formed. Further, the results show that a preference key computed based on the patterns of chemical groups participating in H-bonding is sufficient to predict protein-ligand complexes.

These new insights, and the observed chemical group preferences, can have practical importance in the study and design of protein inhibitors and activators, including the development of therapeutic drugs.

Chapter 3. The testing of the hypothesis that ligand binding and the formation of intermolecular interactions stabilizes the protein-ligand interface provided evidence that ligand poses can be predicted through a computationally quantifiable increase in rigidity upon binding. This computa-

tional study led to the development of SiteInterlock, a new scoring function that has been shown to predict the orientation of a ligand molecule in a protein binding pocket accurately. SiteInterlock measures the rigidity of a binding pocket by considering the cooperativity of the bond network between proteins and their ligand. Experimental results show that the SiteInterlock method can predict near-native ligand poses correctly while providing new information relative to the features measured by pre-existing scoring functions.

Chapter 4. The previous chapters present new information that characterizes protein-ligand interfaces and a method to predict protein-ligand binding poses. However, in many real-world studies of biological processes that are regulated through small molecule binding, the exact structure of the protein binding site is unknown. This imposes additional challenges for the identification of small molecules that can act as either inhibitors or activators. In the absence of structural information about a protein that is involved in regulating biological processes through small molecule binding, this work presents a method, Screenlamp, to identify small molecule inhibitors of biological processes. More specifically, Screenlamp is a new virtual screening framework that allows scientists to test specific hypotheses, such as the importance of particular functional groups in small molecules that are important for activity, to identify potential inhibitors in large databases of millions of molecules efficiently. Screenlamp's approach led to the identification of a potent inhibitor of a GPCR-mediated pheromone signaling pathway for invasive species control. Beyond the computational merits of this work and the development of a toolkit that allows scientists to target specific hypotheses in large-scale virtual screening projects, this work is also the first instance of using ligand-based virtual screening for aquatic invasive species control.

Chapter 5. While many applications of virtual screening are focused on the discovery of small molecules as biological activators or inhibitors, the work presented in this chapter focuses on the mining of the experimental data gathered throughout such studies, to identify the discriminants of biological activity. The motivation behind the analysis of experimental activity data is two-fold: (1) collecting evidence in favor or disfavor of hypotheses about molecular mechanisms involved in

protein-ligand complex formation explaining biological activity, and (2), gaining insights that can be used to drive consequent rounds of small molecule discovery. This includes the development of a machine learning-based analysis pipeline for discovering patterns of functional groups, in small molecules, that are associated with biological activity for a given application. Applied to a dataset of small molecule pheromone inhibitors in a aquatic invasive species control project, the utility of this method was demonstrated by identifying the key functional groups in experimentally tested molecules that were characteristic of active compounds.

Chapter 6. This chapter transfers insights gained through the study of protein-ligand interactions and the discovery of small molecule inhibitors in the previous chapters to identifying small molecule inhibitors of protein-protein interactions. Targeting protein-protein interactions is a notoriously difficult task. One of the reasons is the lack of a cognate small molecule binding partner as a starting point for experiments. Second, the relatively large size of protein-protein interaction sites and their typically largely hydrophobic character create additional difficulties in outcompeting proteins with small molecules that can make fewer interactions. This work presents a new idea and method, called 3D epitope-based virtual screening, that requires as input only a user hypothesis or prior identification of interacting residues on the 3D structure of one side of a protein-protein interface, rather than requiring knowledge of the entire protein-protein complex. The user-defined epitope is used as a 3D structural fragment for volumetric and pharmacophore alignment and scoring with each of 12 million or more commercially available, drug-like molecules in the ZINC database. The goal is to discover mimics of one side of the interface, as potential competitive inhibitors, by assaying the best-scoring small molecules for activity.

Comprehensive introductions and discussions of the related literature are placed at the beginning of each chapter. Beyond the novel scientific contributions presented, all computational methods developed in this work were made freely available on a web-based version control repository, GitHub, under open-source licenses for use by other scientists.

CHAPTER 2

PROTEIN-LIGAND INTERFACES ARE POLARIZED: DISCOVERY OF A STRONG TREND FOR INTERMOLECULAR HYDROGEN BONDS TO FAVOR DONORS ON THE PROTEIN SIDE WITH IMPLICATIONS FOR PREDICTING AND DESIGNING LIGAND COMPLEXES

Adapted with permission from Raschka, Sebastian, Alex J. Wolf, Joseph Bemister-Buffington, and Leslie A. Kuhn. "Protein-ligand interfaces are polarized: Discovery of a strong trend for intermolecular hydrogen bonds to favor donors on the protein side with implications for predicting and designing ligand complexes." *Manuscript submitted for publication.*

2.1 Abstract

Understanding how proteins encode ligand specificity is fascinating and has vast importance for molecular design, similar to deciphering the genetic code. Precise molecular recognition is necessary for all healthy cellular processes and to avoid unregulated growth. For protein-ligand recognition, the combination of an almost infinite variety of interfacial shapes and patterns of chemical groups makes the problem especially challenging. Here we analyze data across non-homologous proteins in complex with small biological ligands to rigorously address observations made in our inhibitor discovery projects: proteins favor donating H-bonds to ligands and avoid using groups with both H-bond donor and acceptor capacity. The results elucidate the code of chemical recognition through the discovery of clear and significant chemical group matching preferences for H-bonds between proteins and native ligands. The trends also provide clear guidance for protein mutagenesis aimed at defining binding sites and ligand interactions. Ligand specificity appears to drive the observed code for binding by disfavoring promiscuous chemical matches and narrowly defining the geometry required to form cognate H-bonds. Together, the chemical and geometric constraints generate a hydrogen bonding lock that can be matched by a ligand bearing the right acceptor-rich key. We demonstrate that measuring an index of preference, based on the atomic chemistry of observed H-bonds, is sufficient to predict protein-ligand complexes. Finally, Hbind and Protein Recognition Index software are provided to rigorously define intermolecular H-bonds by donor/acceptor chemistry and measure the extent to which the H-bonding patterns in a given complex or docking match the preference key.

2.2 Introduction

Across several molecular docking, alignment, screening and crystallographic data analysis projects (Zavodsky et al., 2002; Sukuru et al., 2006; Zavodszky et al., 2009; Van Voorst et al., 2012), we made the following observations:

- Molecules enhanced in chemical groups having both hydrogen bond (H-bond) donor and

acceptor capacity (e.g., hydroxyl groups) tend to lead to false-positive rankings in molecular screening and inaccurate prediction of binding poses for known ligands. This is apparently due to the greater number of potential favorable interactions of donor + acceptor matches (which are augmented by the bond-rotational possibilities for hydroxyl groups), leading to higher protein-ligand interaction counts and overestimated affinity, relative to ligands enhanced in donor-only and acceptor-only groups.

- When analyzing the protein binding sites of a number of protein-small molecule crystal structures, we also noticed that H-bonds tended to be donated from the protein to the ligand, rather than observing an even distribution of donors and acceptors on both sides of the interface.

- While optimizing the docking scoring function for SLIDE (Zavodsky et al., 2002) and the surface alignment scoring function for ArtSurf (Van Voorst et al., 2012) by training on known complexes or site matches, we noted that the terms for matching chemical groups with both donor and acceptor capacity received much smaller weights than donor-only or acceptor-only groups.

An interesting possibility is that nature avoids the presence of chemical groups bearing both H-bond donor and acceptor capacity, such as hydroxyl groups, in the binding sites of proteins or ligands. The many ways of satisfying these groups with H-bond partners could lead to non-selective ligand binding. This hypothesis appears to be supported by the second observation that proteins selectively donate (rather than donate and accept) H-bonds to small molecules. Since those observations were made anecdotally over time and may not hold for protein-ligand complexes in general, the present study was designed to assess whether the above trends (or others) are consistently present in a set of 136 non-homologous proteins bound to a range of biologically relevant small molecules. Selecting this set of proteins with no binding site structural homology between any constituent pair removed any bias towards a given fold, sequence, or function. We then tested whether the resulting statistics of H-bonding trends alone provided enough information

to predict the orientation of ligands relative to their protein partners. The goal was to evaluate whether trends derived from many complexes hold for individual examples well enough to predict the native interactions. Predicting the orientation of ligands for 30 additional complexes also addressed whether the observed trends constitute an essential part of the code for recognition between proteins and ligands.

Advances over the decades in our understanding of protein H-bonding have been well-reviewed in Nittinger et al., 2017. The literature most relevant to the present study falls into two areas: defining energetically favorable H-bonds in terms of geometry (given the integral relationship between favorable geometry and favorable energy) and characterizing H-bond interactions in protein-ligand complexes. Nittinger et al. (Nittinger et al., 2017) analyzed a large number of protein-ligand structures to define preferred H-bond geometries and the extent to which H-bonds observed in experimental structures match theoretically predicted H-bonds based on the valence shell electron pair repulsion model. Their focus was on furthering the accurate modeling and parameterization of H-bonds. As in the work of McDonald & Thornton (McDonald & Thornton, 1994), they found only small energetic differences in out-of-plane H-bonding angles for sp^2 groups such as keto oxygens. This has a key impact on ligand orientational selectivity for donor versus acceptor groups in the present work. Panigrahi and Desiraju (Panigrahi & Desiraju, 2007) also studied protein-ligand H-bonds across a number of diverse, if not necessarily non-homologous, small molecule complexes. Their criteria for defining H-bonds (proton within 3.0 Å of acceptor, resulting in a donor-acceptor distance of up to 4.0 Å, and donor-H-acceptor angle greater than 90°, with 90° reflecting a very weak H-bond) were less stringent than those used here, which could result in the inclusion of relatively low-strength, second-shell (less direct) interactions in their statistics. They defined strong H-bonds as involving polar donor and acceptor atoms, versus weak H-bonds formed by CH donors to oxygen acceptors. They found that N-H—O and O-H—O H-bonds tended to be linear, C-H—O H-bonds to oxygen with Gly and Tyr as donors were ubiquitous in active sites, and that ligands accept twice as often as they donate H-bonds to the protein, consistent with Lipinski's Rule of 5 (Lipinski et al., 1997) as discussed in Section 2.4.8. The current work focuses on identifying

chemical interaction patterns between proteins and their ligands at an atomic chemistry rather than functional group scale, evaluating underlying reasons for such patterns, including ligand selectivity, and testing the extent to which these patterns can predict native interactions.

2.3 Methods

2.3.1 Dataset

A dataset of well-resolved protein complexes with biologically relevant small molecules was constructed based on the intersection between proteins representing different CATH structural folds (Class, Architecture, Topology, Homologous superfamily; http://www.cathdb.info; Dawson et al., 2016) and a set of well-resolved protein structures bound to small organic molecules with known affinity from Binding MOAD (Ahmed et al., 2014; http://bindingmoad.org). This resulted in a dataset of 136 non-homologous protein structures (Table 2.1) from the Protein Data Bank (PDB; http://www.rcsb.org; Berman et al., 2000) with a resolution of 2.4 Å or better (90% at 2.0 Å resolution or better). The protein structures were bound to a diverse set of small ligands (25 peptides, 50 nucleotides, bases and base analogs, and 61 other organic molecules). None of the structures were problematic in ligand fitting or resolution according to the Iridium quality analysis of protein-ligand fitting and refinement (Warren et al., 2012).

Table 2.1: List of all 136 protein-ligand complexes evaluated in this study.

PDB code	Protein description	Ligand code	Ligand category	Lig. chain ID and res. #	Resolution (Å)	R-value work	R-value free
1a9x	Carbamoyl phosphate synthetase	ORN	Peptide-like	A1920	1.8	0.19	-
1af7	Chemotaxis receptor methyltransferase	SAH	Nucleotide-like	A287	2.0	0.20	0.28
1amu	Gramidicin synthetase	PHE	Peptide-like	A566	1.9	0.21	0.25
1awq	Cyclophilin A	Multiple	Peptide-like	B1	1.6	0.34	0.43
1ayl	Phosphoenolpyruvate carboxykinase	OXL	Other	A542	1.8	0.20	0.23
1b4u	Dioxygenase	DHB	Other	D504	2.2	0.16	0.22
1b5e	Deoxycytidylate hydroxymethylase	DCM	Nucleotide-like	B400	1.6	0.19	0.21
1b37	Polyamine oxidase	FAD	Nucleotide-like	A800	1.9	0.20	0.23
1bgv	Glutamate dehydrogenase	GLU	Peptide-like	A501	1.9	0.17	-

Continued on next page

Table 2.1 (cont'd)

PDB code	Protein description	Ligand code	Ligand category	Lig. chain ID and res. #	Resolution (Å)	R-value work	R-value free
1bx4	Adenosine kinase	ADN	Nucleotide-like	A350	1.5	0.19	0.23
1c1d	L-phenylalanine dehydrogenase	NAI	Nucleotide-like	A360	1.3	0.20	0.24
1c96	Aconitase	FLC	Other	A756	1.8	0.23	-
1ccw	Glutamate mutase	TAR	Other	B900	1.6	0.14	0.17
1chm	Creatine amidinohydrolase	CMS	Other	A404	1.9	0.18	-
1cip	Gi-alpha-1 subunit of guanine nucleotide-binding protein	GNP	Nucleotide-like	A355	1.5	0.21	0.24
1cza	Monomeric hexokinase I	G6P	Other	N919	1.9	0.21	0.26
1d0c	Nitric oxide synthase	INE	Other	A760	1.7	0.21	0.26
1d3v	Binuclear manganese metalloenzyme arginase	ABH	Other	A551	1.7	0.16	0.18
1dkx	Molecular chaperone DnaK	Multiple	Peptide-like	B1	2.0	0.21	0.29
1dl5	Protein-l-isoaspartate o-methyltransferase	SAH	Nucleotide-like	A699	1.8	0.18	0.20
1dmh	Catechol 1,2-dioxygenase	LIO	Other	B999	1.7	0.19	0.22
1dtd	Carboxypeptidase A2	GLU	Peptide-like	B300	1.7	0.19	0.23
1e8g	Vanillyl-alcohol oxidase	FCR	Other	A601	2.1	0.22	0.26
1ecm	Chorismate mutase	TSA	Other	A500	2.2	0.19	0.23
1efy	Poly(ADP-ribose) polymerase	BZC	Other	A201	2.2	0.19	0.27
1eu1	Dimethylsulfoxide reductase	MGD	Nucleotide-like	A1001	1.3	0.12	0.15
1evl	Threonyl-tRNA synthetase	TSB	Nucleotide-like	A2002	1.6	0.22	0.23
1eyq	Chalcone isomerase	NAR	Nucleotide-like	A501	1.9	0.24	0.26
1f0l	Diphtheria toxin	APU	Nucleotide-like	A601	1.6	0.19	0.24
1f3l	Arginine methyltransferase	SAH	Nucleotide-like	A529	2.0	0.21	0.26
1f5n	Guanylate-binding protein 1	GNP	Nucleotide-like	A593	1.7	0.23	0.26
1f20	Nitric-oxide synthase	NAP	Nucleotide-like	A1502	1.9	0.19	0.21
1fcy	Retinoic acid nuclear receptor	564	Other	A450	1.3	0.13	0.16
1fk5	Lipid-transfer protein	OLA	Other	A201	1.3	0.14	0.19
1g2l	Coagulation factor x	T87	Other	A1	1.9	0.24	0.27
1g6s	Enzyme 5-enolpyruvylshikimate 3-phosphate synthase	S3P	Other	A601	1.5	0.15	0.17
1g55	DNA methyltransferase homolog	SAH	Other	A392	1.8	0.21	0.25
1g72	Uinoprotein methanol dehydrogenase	PQQ	Other	A701	1.9	0.16	0.19
1gk8	Rubisco	CAP	Other	A1477	1.4	0.15	0.16
1gs5	N-acetyl-L-glutamate kinase	NLG	Other	A1259	1.5	0.21	0.21
1gte	Dihydropyrimidine dehydrogenase	IUR	Nucleotide-like	A1034	1.7	0.18	0.20
1gx5	Hepatitis C virus RNA polymerase	GTP	Nucleotide-like	A1532	1.7	0.19	0.22
1gz8	Cyclin dependent kinase 2	MBP	Nucleotide-like	A1300	1.3	0.15	0.19
1h8e	Mitochondrial F1-ATPase	ADP	Nucleotide-like	A600	2.0	0.21	0.24

Continued on next page

Table 2.1 (cont'd)

PDB code	Protein description	Ligand code	Ligand category	Lig. chain ID and res. #	Resolution (Å)	R-value work	R-value free
1h16	Pyruvate formate-lyase	DTL	Other	A9010	1.5	0.15	0.16
1hfe	Fe-only hydrogenase	CYS	Peptide-like	L432	1.6	0.16	0.18
1hp1	5'-Nucleotidase	ATP	Nucleotide-like	A606	1.7	0.18	0.20
1hqs	Isocitrate dehydrogenase	CIT	Other	A425	1.6	0.20	0.25
1hyo	Fumarylacetoacetate hydrolase	HBU	Other	B1011	1.3	0.18	0.20
1i1q	Anthranilate synthase	TRP	Peptide-like	A1001	1.9	0.22	0.25
1i24	UDP-sulfoquinovose synthase	UPG	Nucleotide-like	A402	1.2	0.19	0.20
1j09	Glutamyl-tRNA synthetase	ATP	Nucleotide-like	A501	1.8	0.20	0.23
1jak	Beta-N-acetylhexosaminidase	IFG	Other	A601	1.8	0.18	0.19
1jc9	Tachylectin	NAG	Other	A270	2.0	0.18	0.20
1jet	Oligo-peptide binding protein	Multiple	Peptide-like	B1	1.2	0.23	0.26
1jhg	Trp repressor	TRP	Peptide-like	A111	1.2	0.13	0.17
1k3y	Glutathione S-transferase	GTX	Other	A5100	1.3	0.14	0.21
1k5n	Major histocompatibility complex molecule HLA-B*2709	Multiple	Peptide-like	C1	1.1	0.12	0.15
1ka1	Halotolerance protein HAL2	A3P	Nucleotide-like	A601	1.3	0.13	0.17
1kek	Pyruvate-ferredoxin oxidoreductase	HTL	Nucleotide-like	A2236	1.9	0.18	0.23
1kgq	Tetrahydrodipicolinate N-succinyltransferase	NPI	Other	A301	2.0	0.18	0.25
1kjq	Phosphoribosylglycinamide formyltransferase 2	ADP	Nucleotide-like	A1	1.1	0.19	0.21
1kmv	Dihydrofolate reductase	LII	Other	A201	1.1	0.13	0.18
1kol	Formaldehyde dehydrogenase	NAD	Nucleotide-like	A1403	1.7	0.17	0.21
1kpf	Protein kinase C interacting protein	AMP	Nucleotide-like	A200	1.5	0.21	0.24
1krh	Benzoate dioxygenase reductase	FAD	Nucleotide-like	A501	1.5	0.24	0.25
1kyf	Alpha-adaptin C	Multiple	Peptide-like	P628	1.2	0.15	0.21
1l5o	Nicotinate-nucleotide–dimethylbenzimidazole phosphoribosyltransferase	2MP	Other	A990	1.6	0.17	0.20
1l8b	Eukaryotic translation initiation factor 4E	MGP	Nucleotide-like	A1000	1.8	0.22	0.25
1lb6	TRAF6 signaling protein	Multiple	Peptide-like	B601	1.8	0.20	0.26
1lri	Beta-elicitin cryptogein	CLR	Other	A99	1.5	0.16	0.19
1ltz	Phenylalanine-4-hydroxylase	HBI	Other	A500	1.4	0.16	0.22
1lug	Carbonic anhydrase	SUA	Other	A1002	1.0	0.12	0.14
1m0w	Glutathione synthase	3GC	Peptide-like	A501	1.8	0.17	0.20
1m15	Arginine kinase	ARG	Peptide-like	A403	1.2	0.12	0.14
1mgp	Hypothetical protein TM84	PLM	Other	A314	2.0	0.20	0.23
1mqo	Beta-lactamase II	CIT	Other	A300	1.4	0.22	0.25
1mrj	Alpha-trichosanthin	ADN	Nucleotide-like	A300	1.6	0.17	-
1msk	Methionine synthase	SAM	Nucleotide-like	A1301	1.8	0.20	0.26

Continued on next page

Table 2.1 (cont'd)

PDB code	Protein description	Ligand code	Ligand category	Lig. chain ID and res. #	Resolution (Å)	R-value work	R-value free
1mxt	Cholesterol oxidase	FAE	Nucleotide-like	A510	1.0	0.11	0.13
1n62	Carbon monoxide dehydrogenase	MCN	Nucleotide-like	B3920	1.1	0.14	0.17
1nd4	Aminoglycoside 3'-phosphotransferase	KAN	Other	A1300	2.1	0.21	0.24
1nki	Fosfomycin resistance protein a	PPF	Other	A5001	1.0	0.15	0.18
1nox	NADH oxidase	FMN	Nucleotide-like	A300	1.6	0.19	0.20
1nvv	Transforming protein p21	GNP	Nucleotide-like	Q1001	2.2	0.21	0.24
1o2d	Alcohol dehydrogenase	NAP	Nucleotide-like	A1800	1.3	0.14	0.17
1o7n	Naphthalene 1,2-dioxygenase	IND	Peptide-like	A505	1.4	0.19	0.20
1o7q	N-acetyllactosaminide alpha-1,3-galactosyl-transferase	UDP	Nucleotide-like	A1374	1.3	0.12	0.15
1oai	Nuclear RNA export factor	Multiple	Peptide-like	B10	1.0	0.15	0.16
1oew	Endothiapepsin	Multiple	Peptide-like	A401	0.9	0.12	0.15
1ouw	Lectin	MLT	Other	D501	1.4	0.15	0.18
1p5d	Phosphomannomutase	G1P	Other	X658	1.6	0.16	0.18
1p6o	Cytosine deaminase	HPY	Other	B410	1.1	0.11	0.15
1p7t	Malate synthase G	ACO	Nucleotide-like	A800	2.0	0.20	0.29
1pfv	Methionyl-tRNA synthetase	2FM	Peptide-like	A553	1.7	0.19	0.20
1pp9	Ubiquinol-cytochrome C reductase complex core protein I	SMA	Other	C2001	2.1	0.25	0.29
1pq7	Trypsin	ARG	Peptide-like	A703	0.8	0.11	-
1puj	Conserved hypothetical protein ylqF	GNP	Nucleotide-like	A501	2.0	0.22	0.25
1pz4	Sterol carrier protein-2	PLM	Other	A200	1.4	0.19	0.23
1q79	Mammalian poly(A) polymerase	3AT	Nucleotide-like	A1000	2.2	0.21	0.24
1qja	14-3-3 Protein zeta	Multiple	Peptide-like	Q7	2.0	0.21	0.28
1qmg	Acetohydroxy-acid isomeroreductase	DMV	Other	A620	1.6	0.20	0.22
1qnf	Photolyase	HDF	Other	A486	1.8	0.20	0.24
1qxy	Methionine aminopeptidase	M2C	Other	A3001	1.0	0.14	0.17
1qz5	Actin	KAB	Other	A500	1.5	0.17	0.19
1r1h	Neprilysin	BIR	Other	A2001	2.0	0.21	0.26
1r4u	Uricase	OXC	Other	A999	1.7	0.16	0.18
1r8s	ADP-ribosylation factor 1	GDP	Nucleotide-like	A401	1.5	0.16	0.17
1rkd	Ribokinase	RIB	Other	A311	1.8	0.22	0.26
1rlz	Deoxyhypusine synthase	NAD	Nucleotide-like	A700	2.2	0.20	0.25
1rqw	Thaumatin I	TLA	Other	A1001	1.1	0.13	0.15
1sox	Sulfite oxidase	MTE	Nucleotide-like	A501	1.9	0.18	0.22
1t2d	L-lactate dehydrogenase	NAD	Nucleotide-like	A323	1.1	0.14	0.15
1tbb	cAMP-specific 3',5'-cyclic phosphodiesterase 4D	ROL	Other	A501	1.6	0.19	0.20

Continued on next page

Table 2.1 (cont'd)

PDB code	Protein description	Ligand code	Ligand category	Lig. chain ID and res. #	Resolution (Å)	R-value work	R-value free
1tl2	Tachylectin-2	NDG	Other	A237	2.0	0.16	0.20
1tw6	Baculoviral IAP repeat-containing protein 7	Multiple	Peptide-like	D1	1.7	0.16	0.17
1tx4	Transforming protein Rhoa	GDP	Nucleotide-like	B680	1.7	0.17	0.22
1u4g	Elastase	HPI	Peptide-like	A800	1.4	0.18	0.20
1ucd	Ribonuclease	URA	Nucleotide-like	A501	1.3	0.20	0.20
1uf5	N-carbamyl-D-amino acid amidohydrolase	CDT	Other	A998	1.6	0.18	0.20
1ufy	Chorismate mutase	MLI	Other	A201	1.0	0.11	0.13
1uio	Adenosine deaminase	HPR	Nucleotide-like	A353	2.4	0.20	-
1unq	Transferase	4IP	Other	A1117	1.0	0.15	0.18
1us0	Aldose reductase	LDT	Other	A320	0.7	0.09	0.10
1usc	Putative styrene monooxygenase small component	FMN	Nucleotide-like	A1179	1.2	0.20	0.22
1uuy	Molybdopterin-bound Cnx1G domain	PPI	Other	A1166	1.5	0.16	0.18
1uw6	Acetylcholine-binding protein	NCT	Other	A1208	2.2	0.22	0.27
1uxy	Uridine diphospho-n-acetylenolpyruvylglucosamine reductase	EPU	Nucleotide-like	A402	1.8	0.20	0.25
1uze	Angiotensin converting enzyme	EAL	Peptide-like	A3002	1.8	0.19	0.21
1v7r	Hypothetical protein PH1917	CIT	Other	A1200	1.4	0.20	0.22
1xva	Glycine N-methyltransferase	SAM	Nucleotide-like	A293	2.2	0.20	0.26
2dpm	Adenine-specific methyltransferase	SAM	Nucleotide-like	A300	1.8	0.24	0.28
2sli	Intramolecular trans-sialidase	SKD	Other	A760	1.8	0.19	0.22
2tct	Tetracycline repressor	CTC	Other	A222	2.1	0.18	-
4ubp	Urease	HAE	Other	C800	1.6	0.15	0.19
5csm	Chorismate mutase	TRP	Peptide-like	A300	2.0	-	

2.3.2 Protonation

Protonation of each protein-ligand complex was performed with the OptHyd method in YASARA Structure (version 16.4.6; http://www.yasara.org; Krieger et al., 2009), retaining interfacial metals and removing bound water molecules, with the goal of assessing direct, strong interactions between proteins and their ligands. During the addition of hydrogen atoms and optimization of the H-bond network using YASARA, heavy atom positions were maintained except for the rotation of the terminal amide groups of asparagine and glutamine side chains through 180° when interchange of the =O and $-NH_2$ groups resulted in improvement in polar interactions. This step disambiguates

the fitting of these side chains into electron density due to the similar density of oxygen and nitrogen atoms at typical crystallographic resolution. YASARA assigns the tautomeric state of the imidazole groups in histidine side chains according to the intra- and intermolecular hydrogen and metal bonding of the histidine and the influence of neighboring polar groups on the pKa of its imidazole ring (Krieger et al., 2012). For nucleotidyl ligands and bases, the results of high-level *ab initio* calculations on protonation states and tautomers, and how they are influenced by H-bonding in complexes, were also considered (Colominas et al., 1996).

2.3.3 Optimization of proton orientation

To minimize steric clashes and optimize the polar interaction network, OptHyd also optimized the orientation of protein and ligand protons (for instance, the hydrogen positions in rotatable NH_3 and OH groups). This method uses the YAMBER force field, a second-generation force field derived from AMBER, which was self-parameterized according to the protonated protein, water molecules, and ions present in the complete unit cells of 50 high resolution X-ray structures (Krieger et al., 2004). All 136 of the complexes in our analysis were checked for agreement between YASARA protonation of the ligand in complex with the protein, relative to protonation of the ligand alone using molcharge in OpenEye QUACPAC (version 1.7.0.2; https://www.eyesopen.com/quacpac; OpenEye Scientific Software, Santa Fe, NM) with the AM1-BCC option (Jakalian et al., 2002). Protonation and ligand valences resulting from YASARA were also visually inspected with PyMOL (version 1.8.2.2, https://www.schrodinger.com/pymol; DeLano, 2002). In cases of ambiguities or differences in protonation state or valence, the chemical literature for the protein-ligand complex and protonation studies for that ligand were consulted, resulting in manual correction relative to the YASARA protonation in a few cases. The protonated ligands provided in PDB format by YASARA were converted to Tripos MOL2 format with the OpenEye OEChem toolkit (version 1.7.2.4; https://www.eyesopen.com/oechem-tk; OpenEye Scientific Software, Santa Fe, NM).

2.3.4 Influence of partial charges

In the Hbind software used to define H-bond interactions (including salt bridges satisfying H-bond criteria, described in the following paragraph), partial charges are only used to assess whether an atom pair can form longer-range salt bridges. The salt bridge assignment requires a higher than 0.3 charge magnitude on the ligand atom interacting with a charged protein or metal atom and a maximum distance of 4.5 Å between the donor and acceptor. These longer range salt bridges are often second-shell interactions and thus were not included in the current analysis. Hence, as expected, charges assigned by either the Merck Molecular Force Field (MMFF94; Halgren, 1996) or AM1-BCC (Jakalian et al., 2002) using molcharge in QUACPAC resulted in the same list of direct H-bond and metal bridge interactions for the 136 complexes.

2.3.5 Hbind software

This software developed in our laboratory (available from GitHub at https://github.com/psa-lab/Hbind) was used to define direct H-bonds and metal bonds with ligands. Pauling wrote, "Only the most electronegative atoms should form H-bonds, and the strength of the bond should increase with increase in the electronegativity of the two bonded atoms ... [Thus] we might expect that fluorine, oxygen, nitrogen and chlorine would possess this ability, to an extent decreasing in this order" (Pauling, 1960). In our software, nitrogen and oxygen atoms are considered as potential donors or acceptors of H-bonds and fluorine and chlorine as potential acceptors. Hbind interprets the donor/acceptor capacity of ligand atoms from information in the MOL2 file detailing the hybridization, the order of covalent bonds with neighboring atoms, and the protonation state of these atoms. The software implicitly evaluates by analytic geometry all orientations of protons in rotatable groups for their ability to satisfy the H-bond criteria defined below, while not altering their coordinates in the PDB or MOL2 file. For instance, protons in X—NH3 and X—OH groups can adopt any sterically admissible position on a circle upon rotation of the X—N or X—O single bond. The H-bond identification criteria are based on those of Ippolito et al. (Ippolito et al., 1990) and McDonald and Thornton (McDonald & Thornton, 1994), all of which must be met:

- Hydrogen to acceptor distance: 1.5-2.5 Å
- Donor to acceptor distance: 2.4-3.5 Å
- Donor-H-acceptor angle (θ): 120-180°
- Pre-acceptor–acceptor–H angle (ϕ): 90-180°

These donor, hydrogen, and acceptor geometries are depicted in Figure 2.1.
Criteria for protein or ligand-bound metals to form a bond with an atom on the second molecule bearing a lone pair of electrons:

- Lone pair atom distance to K or Na: 2.0-2.9 Å
- Lone pair atom distance to Ca, Co, Cu, Fe, Mg, Mn, Ni, or Zn: 1.7-2.6 Å

Hbind calculates and outputs the interaction distance and angles between each protein-ligand atom pair forming an H-bond or metal interaction. Additional command-line options are available to list longer-range salt bridges (up to 4.5 Å between protein and ligand), direct hydrophobic contacts, and the protein-ligand orientation and affinity scores and terms calculated by SLIDE (version 3.4; http://kuhnlab.bmb.msu.edu/software/slide/index.html; Zavodsky et al., 2002).

2.3.6 Identification of ligand H-bonding patterns

This analysis aimed to identify any consistent patterns of nitrogen donor interactions from proteins to ligands in the dataset of 136 non-homologous complexes. When visualizing the complexes with PyMOL, geometrical similarities were apparent in the H-bond networks with nucleotidyl ligands, involving a visually distinctive pattern of protein H-bond donors. To assess this objectively, unsupervised clustering algorithms were used to group and discover common H-bonding patterns and report their occurrence across the residue positions in the protein. The pattern of H-bond interactions within each complex was represented by a binary vector listing the presence (1) or absence (0) of an H-bond to the ligand for each position in the sequence. Because the number

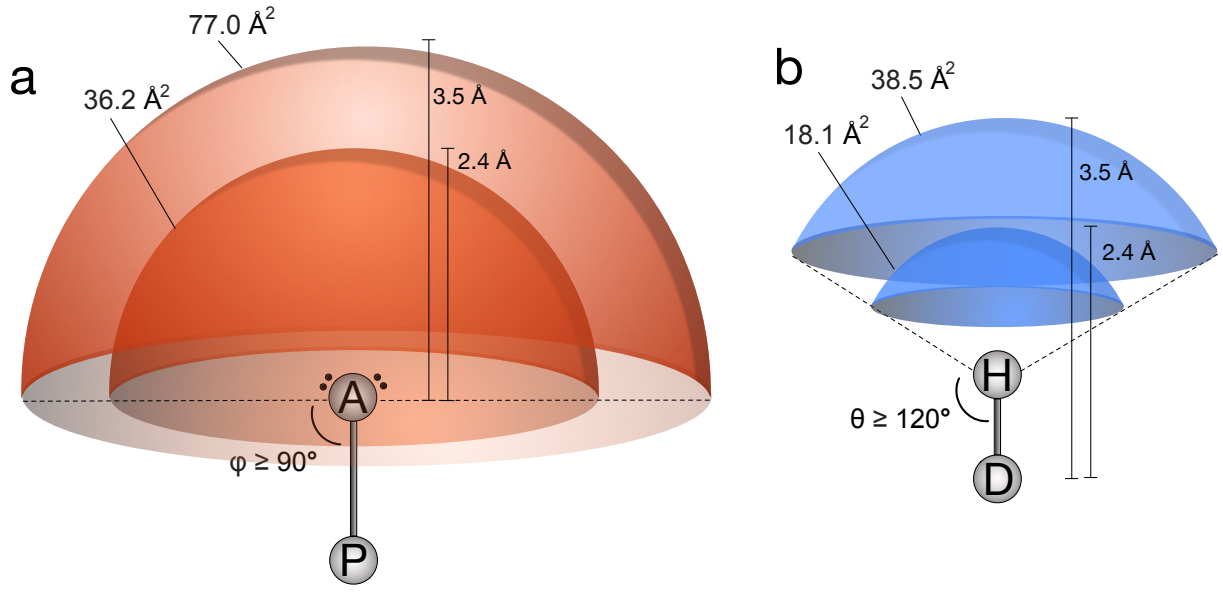


Figure 2.1: Favorable regions for H-bonding partners. The favorable regions are shown for H-bonding partners of (a) an acceptor atom, and (b) a donor atom in the protein or ligand. Based on the geometric criteria described in the text, the outer and inner shells represent the maximum and minimum distances for the H-bond partner atom relative to the acceptor (a) or donor (b). Given that the pre-acceptor–acceptor–H angle (ϕ) can range from 90-180°, the surface area of the outer shell defining the maximum distance at which a donor atom can interact favorably, within 3.5 Å of the acceptor, is 77.0 Å^2. The inner shell at 2.4 Å correspondingly represents the surface of minimum distance for the donor relative to the acceptor. The favorable volume for a protein atom to H-bond with a donor atom (a) or an acceptor atom (b) on the ligand is defined as the volume between the inner and outer shells. Because the pre-acceptor–acceptor–H angle (ϕ) can range from 90-180° for each lone pair on the acceptor (a), while the range for the donor-H-acceptor angle in (b) is narrower (120-180°), the volume in which a ligand proton can favorably bind to a protein acceptor atom (60.6 Å^3) is twice the volume where an acceptor atom can make a favorable interaction with a donor atom (30.3 Å^3).

of possible interaction patterns for protein sequences with hundreds of residues and arbitrary spacing between the H-bonding positions is almost infinite, we chose to focus on the sub-case of identifying local H-bonding sequence patterns with at least three interacting residues and no more than 10 residues intervening between a pair of successive interactions. For each protein, the initial H-bonding vector was then split into non-overlapping sub-vectors (local motifs), such that each sub-vector started and ended with an H-bonding residue and did not contain a contiguous subsequence of more than ten zero-elements (non-interacting residues). For example, an H-bond interaction vector consisting only of nitrogen donors (here, from PDB entry 1f5n; Prakash et al., 2000) appears as follows between the vertical bars, where the initial number (1) is the first residue

number in the sequence, and the last number (1361) is the final residue number:

1| 0 0 0 ... 0 0 0 |1361

Note that these vectors are formatted as binary sequences, where residues forming interfacial H-bonds are labeled with 1's, and residues not involved in H-bonds are set to 0. The following excerpt of a protein H-bond interaction vector shows a region within the above complete vector containing several local interaction vectors:

43| 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 1|69

The extracted sub-vectors or local H-bonding motifs were then:

43| 1 1 1 1 1 |47 and 62| 1 0 0 0 0 0 1 1 |69

Once all sub-vectors were extracted, they were tabulated by protein and concatenated into a dataset containing the local motifs from all 136 complexes. Trailing zeros were added so that all motifs have the same length, facilitating comparison of positions across multiple sequences:

N_1B5E_1_D400| 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0

N_1BX4_1_A350| 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0

N_1CIP_1_A355| 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0

N_1F0L_1_A601| 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0

N_1F5N_1_A593| 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0

N_1F5N_2_A593| 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0

etc.

The first letter in each row denotes whether this subsequence corresponds to a peptide-like (P), nucleotide-like (N), or other organic (O) ligand. Because some protein complexes contained more than one interaction motif, the digit following the underscore after the PDB code indexes the motifs in a given protein. The first character after the next underscore is the PDB chain ID of the protein and ligand analyzed, and the remaining digits specify the residue number of the ligand molecule. Based on the matrix above, the interaction sequences were clustered via average linkage, with Euclidean distance as the metric for the distance between each pair of motifs, by using NumPy (version 1.13.3; http://www.numpy.org; Van Der Walt et al., 2011) and SciPy (version 0.19.1;

https://www.scipy.org; Jones et al., 2001).

2.3.7 Software for statistical analyses

The parsing of Hbind interaction tables and the statistical analyses in this work were carried out in Python using NumPy (version 1.13.3; http://www.numpy.org; Van Der Walt et al., 2011), SciPy (version 0.19.1; https://www.scipy.org; Jones et al., 2001), and Pandas (version 0.20.3; https://pandas.pydata.org; McKinney, 2010). The BioPandas package (version 0.2.2; http://rasbt.github.io/biopandas/; Raschka, 2017a) was used to compute statistics from MOL2 and PDB files.

2.3.8 Visualization and plotting software

All data plots were created using the matplotlib library (version 2.0.2; https://matplotlib.org; Hunter, 2007). The Affinity Designer software (version 1.6.0; https://affinity.serif.com/en-us/designer/) was used to enhance the readability of figure labels as necessary. Structural renderings of molecules were created in PyMOL (version 1.8.2.2; https://pymol.org; DeLano, 2002), and figures depicting geometric properties were drawn in OmniGraffle (version 7.5; https://www.omnigroup.com/omnigraffle).

2.4 Results and Discussion

The output of Hbind with direct intermolecular H-bonds and metal interactions for all 136 complexes (Table 2.1) was the basis for addressing a series of molecular recognition questions presented and discussed in this section. An example for one complex is shown in Figure 2.2.

2.4.1 Are donor groups on proteins preferred in H-bonding to biological ligands?

The interaction tables for 136 complexes were analyzed to count the frequency of protein atoms acting as H-bond acceptors versus donors in direct H-bonds to the ligand and likewise for ligand atoms (Figure 2.3). The preference for the protein to donate H-bonds to a ligand acceptor atom was more than 2:1, with 712 H-bonds donated by the protein to the ligand and 345 H-bonds from

```
--------------------------------------------------------------------------------
PDB code of the protein-ligand complex: 1r8s, chain ID: A, ligand residue number: 401
--------------------------------------------------------------------------------
Hbind (version: 1.0) Protein Structural Analysis & Design Lab, MSU (kuhnlab@msu.edu)

MOL2 file: /home/raschkas/protonated_ligands/1r8s.mol2
PDB file: /home/raschkas/proteins/1r8s.mol2

+++++++++++++++++ Summary ++++++++++++++++++
|
| Protein-Ligand Hydrophobic Contacts:     33
| Protein-Ligand H-bonds             :     16
| Protein-Ligand Salt-bridges        :      4
| Metal-Ligand Bonds                 :      0

++++++++++++++++++++++ Hbind Interaction Table ++++++++++++++++++++++++++
#
#            | Ligand Atom -- Protein  Atom | Bond   D-H-A  Ligand-Protein
#            |  #  type    -- RES   #  type | Dist.  Angle  Interaction
| hbond     1  16  N.am    -- ASP  129  OD1   2.749  173.3  Donor - Acceptor
| hbond     2  18  N.pl3   -- ASP  129  OD2   2.917  165.1  Donor - Acceptor
| hbond     3  22  N.2     -- ASN  126  ND2   3.051  141.5  Acceptor - Donor
| hbond     4  25  O.3     -- LYS  127  NZ    3.221  149.0  Acceptor - Donor
| hbond     5  30  O.2     -- THR   32  N     2.846  150.8  Acceptor - Donor
| hbond     6  30  O.2     -- THR   32  OG1   2.686  178.9  Acceptor - Donor
| hbond     7  31  O.2     -- THR   31  N     2.927  159.3  Acceptor - Donor
| hbond     8  31  O.2     -- THR   31  OG1   2.735  177.4  Acceptor - Donor
| hbond     9  32  O.3     -- LYS  156  NZ    2.757  173.5  Acceptor - Donor
| hbond    10  33  O.3     -- GLY   29  N     3.010  159.5  Acceptor - Donor
| hbond    11  33  O.3     -- LYS   30  N     2.911  160.2  Acceptor - Donor
| hbond    12  33  O.3     -- LYS   30  NZ    2.868  177.9  Acceptor - Donor
| hbond    13  34  O.3     -- GLY   29  N     3.204  123.5  Acceptor - Donor
| hbond    14  39  O.3     -- ALA   27  N     2.850  155.6  Acceptor - Donor
| hbond    15  40  O.2     -- LYS  127  N     3.268  120.7  Acceptor - Donor
| hbond    16  40  O.2     -- ALA  160  N     2.996  131.1  Acceptor - Donor

================================================================================
```

Figure 2.2: Example of Hbind intermolecular direct H-bond and metal interaction output. This Hbind example output shows the intermolecular direct H-bond and metal interactions for chain A of PDB entry 1r8s (Renault et al., 2003) in complex with ligand GDP (chain ID: A, ligand residue number: 401), showing only those interactions meeting the criteria defined in the Methods. The ligand atom number and type are from the MOL2 file definition, and the protein residue number and atom type, bond length between H-bond donor and acceptor atoms, and the donor-hydrogen-acceptor (θ) angle are also listed. The final columns indicate the orientation of the hydrogen bond, i.e., whether the ligand or protein contributed the donor atom, and likewise for the acceptor.

the ligand accepted by the protein, across all 136 complexes. Since H-bonds were analyzed based on atomic interactions, including proton positions, a residue or atom could participate in multiple H-bonds if all the angular and distance criteria were met for each bond.

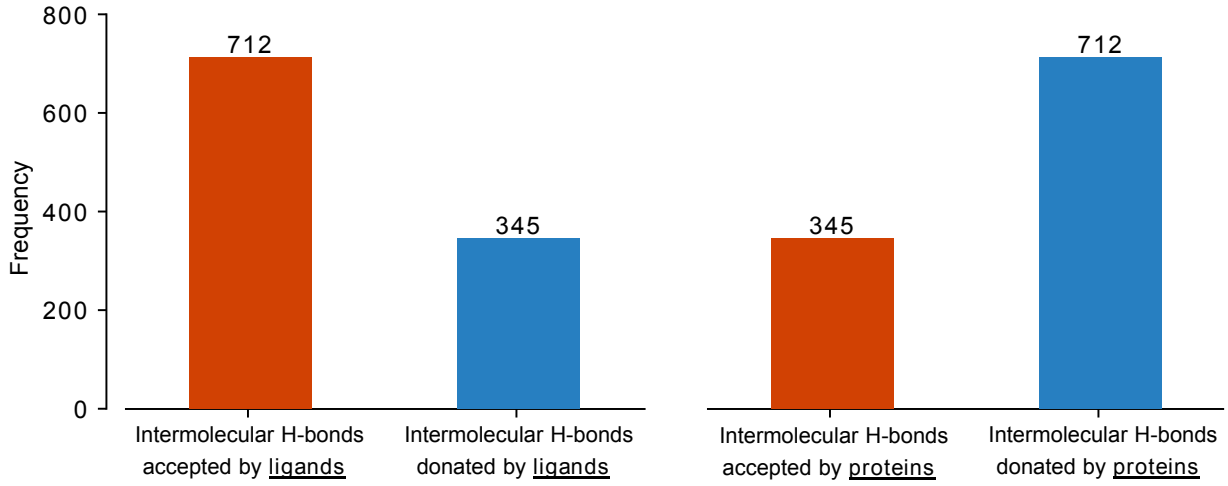


Figure 2.3: Frequency of donated and accepted intermolecular H-bonds across the 136 diverse complexes. The frequency of donated and accepted intermolecular H-bonds across the 136 diverse complexes is shown from the ligand's perspective (two bars on the left) and the protein's perspective (two bars on the right). Throughout the figures, red is used to indicate H-bond acceptors, while blue indicates donors.

When subdivided further according to the patterns of nitrogen and oxygen atoms involved in protein-ligand H-bonds, an interesting trend came to light: the majority (70%) involved both a nitrogen atom donor and an oxygen acceptor (Table 2.2), with a full 76% of intermolecular H-bonds donated by a nitrogen atom. The second most prevalent case paired a hydroxyl group donor with an oxygen acceptor (24%). Other possibilities for native ligand H-bonding were rare, particularly nitrogen atoms acting as H-bond acceptors, whether on the protein or ligand side. The tendency of hydroxyl groups to contribute only one-quarter of all protein-ligand H-bonds despite having two lone pairs and one proton, all of which can form H-bonds, can be rationalized by the resulting reduction in ligand selectivity. A ligand group with either good donor or acceptor geometry could both interact with that hydroxyl group, bringing the risk of misrecognition. This could have been the basis for negative selection during functional evolution.

Table 2.2: Intermolecular NH versus OH hydrogen bond donor frequencies for oxygen and nitrogen acceptors.

H-bond donor molecule	H-bond type	Frequency	H-bond acceptor molecule
Protein	N-H $\cdots$ O	524	Ligand
Protein	N-H $\cdots$ N	53	Ligand
Protein	O-H $\cdots$ O	127	Ligand
Protein	O-H $\cdots$ N	6	Ligand
Ligand	N-H $\cdots$ O	219	Protein
Ligand	N-H $\cdots$ N	1	Protein
Ligand	O-H $\cdots$ O	124	Protein
Ligand	O-H $\cdots$ N	1	Protein

2.4.2 Can the observed trends in interfacial polarity, with H-bonds tending to be formed by donors on the protein side of the interface interacting with acceptors on the ligand side, be explained by the prevalence of binding-site protons versus lone pairs?

To answer this question, the binding site was defined as all protein residues with at least one heavy atom within 9 Å of a ligand heavy atom. This set of potentially interacting atoms is typically used for interfacial analysis or scoring. All the previously mentioned criteria were then applied to identify intermolecular H-bonds, namely, meeting the 2.4-3.5 Å range for donor-acceptor distance and satisfying both the donor-H-acceptor and preacceptor-acceptor-H angular criteria. An example binding site and intermolecular H-bond network for one of the complexes appears in Figure 2.4. For each binding site or ligand atom with H-bonding potential, the number of protons available to donate and the number of lone pairs available to accept H-bonds were tabulated and summed over the 136 complexes. The results (Figure 2.5) show that acceptor lone pairs are significantly more prevalent than donor protons in the ligand binding sites of proteins (approx.16,000 lone pairs: approx. 10,000 protons available to donate), with a similar excess of lone pairs found in the ligands (approx. 15,000 lone pairs: approx. 9,000 donor protons). Thus, if formation of intermolecular H-bonds were primarily driven by the prevalence of protons and lone pairs, the protein would be expected to accept H-bonds 1.6 times more often than it donates them. Given that the observed trend is in the opposite direction (a 2:1 tendency to donate H-bonds to the ligand; Figure 2.3), there appears to be an underlying strong chemical or evolutionary preference for proteins to act as donors

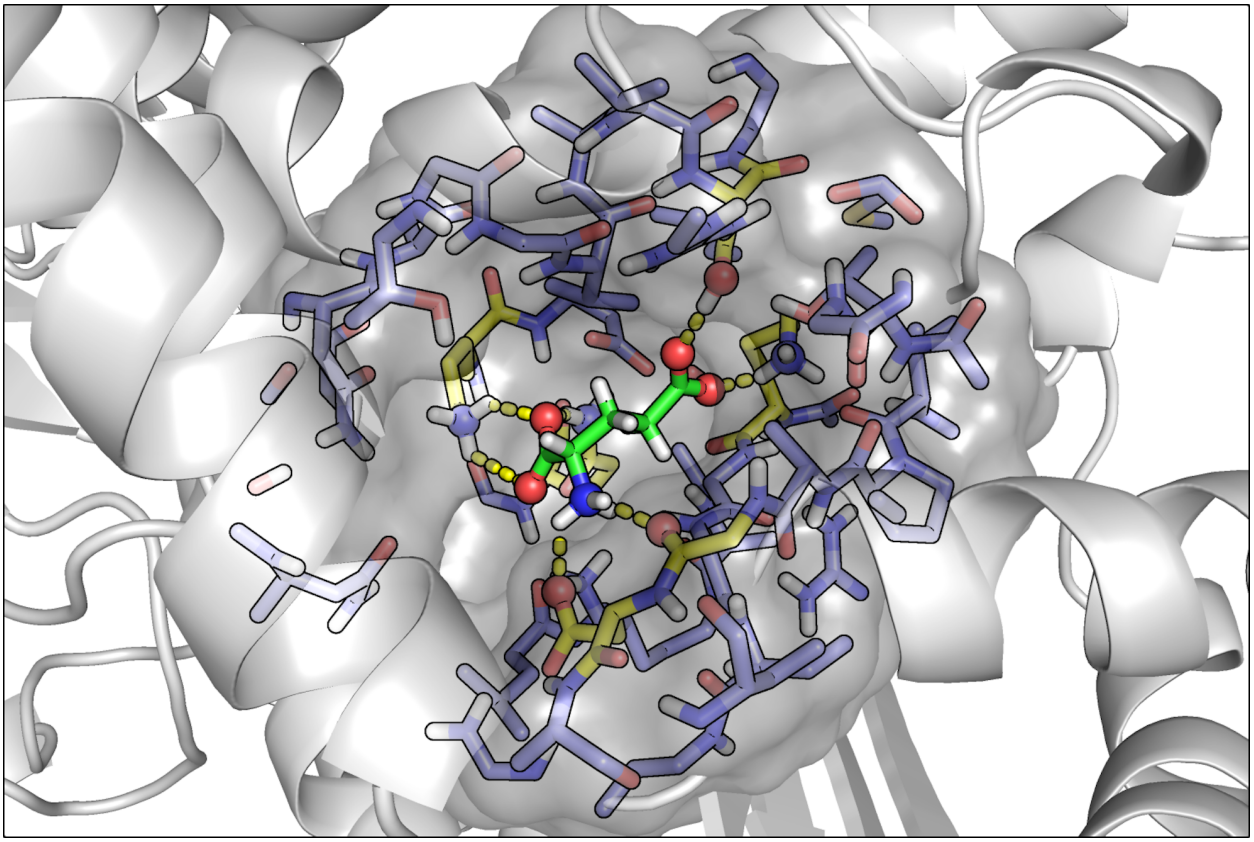


Figure 2.4: Binding site definition for glutamate hydrogenase interacting with a glutamic acid ligand. This figure shows the 9 Å binding site definition for glutamate hydrogenase interacting with a glutamic acid ligand (PDB entry 1bgv; Stillman et al., 1993). The gray solvent-accessible molecular surface envelops the ligand binding pocket defined as all protein atoms within 9 Å of the ligand's heavy atoms (green tubes). The binding site residues H-bonding to the ligand are shown with carbon atoms in yellow, and all other binding site residues' carbon atoms are colored in purple. Protein-ligand H-bonds as defined by Hbind are shown as yellow dashed lines.

when binding cognate ligands.

2.4.3 Do certain residues predominate in the observed preference for proteins to donate H-bonds to ligands?

The statistics of donor and acceptor atoms participating in interfacial H-bonds (Figure 2.3) were further analyzed by atom type (Figure 2.6b). Panel (a) shows that amines, especially the terminal NH groups in Arg, Asn, Gln, and Lys, are the dominant donors of H-bonds to ligands, relative to hydroxyl groups. This cannot be explained by their prevalence in the binding sites: When the number of H-bonds formed is divided by the number of binding site occurrences, the H-bonding

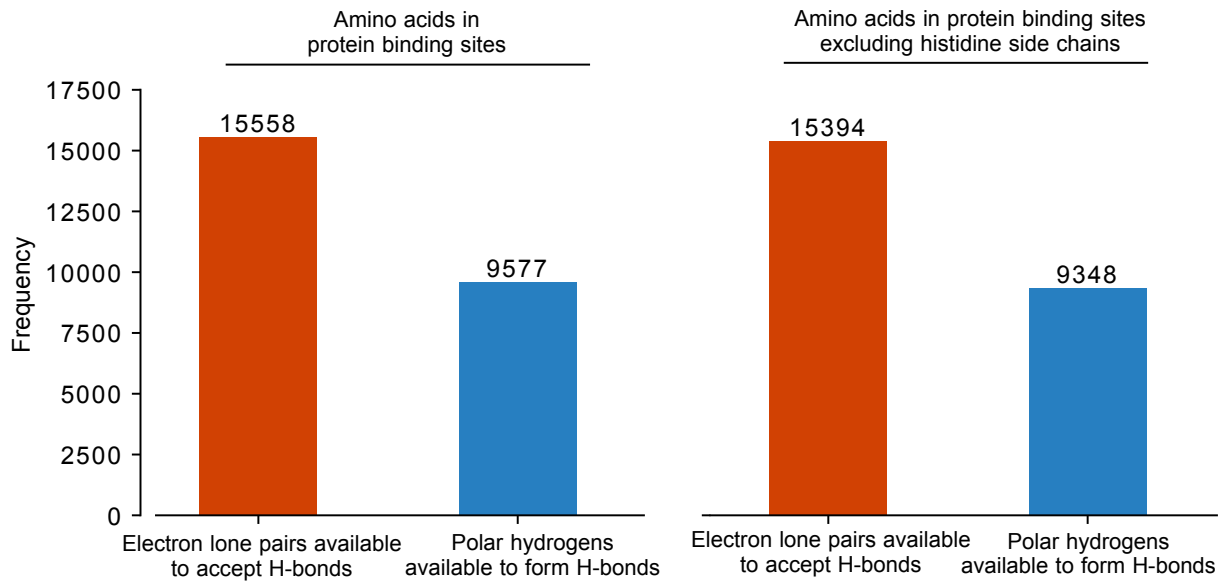


Figure 2.5: Statistics across 136 non-homologous complexes of the number of electron lone pairs in the protein's binding site available to act as H-bond acceptors compared with the number or protons available to be donated. The observed frequencies indicate that ligand binding sites have a significant excess of lone pairs relative to protons that can participate in H-bonds. The analysis was performed with histidine side chains (two bars at left) and without (two bars at right), because this residue's protonation state is more difficult to define. However, the histidine residues in the 136 complexes are primarily involved in metal interactions (in which the nitrogen lone pairs form bonds with cationic metals). Consequently, the statistics are substantially similar with and without histidine.

of terminal amines, especially in lysine, only becomes more pronounced (Figure 2.6b). This is interesting, because Lys pays a higher entropic cost in lost degrees of bond-rotational freedom when H-bonding to ligands (due to having 4 side chain single bonds), relative to Arg (3 side chain single bonds) and especially Ser or Thr (2 single bonds). Lys, Ser, and Thr can each potentially form up to three H-bonds with ligands, relative to Arg, which can form up to five. This also does not explain the preference for Lys. It could be that the greater flexibility and length of Lys and its rotatable proton positions (relative to the rigid and planar Arg guanidinium group) allow this side chain to better optimize H-bonds with ligands. Overall, the most prevalent H-bond donors and acceptors to ligands are the charge-bearing side chain atoms in Arg, Asp, Glu, and Lys, followed by the polar amine groups in Asn and Gln. The Asn and Gln NH_2 groups form about 3 times as many ligand H-bonds as their terminal keto oxygens, despite having the capacity to form the same number of H-bonds per group.

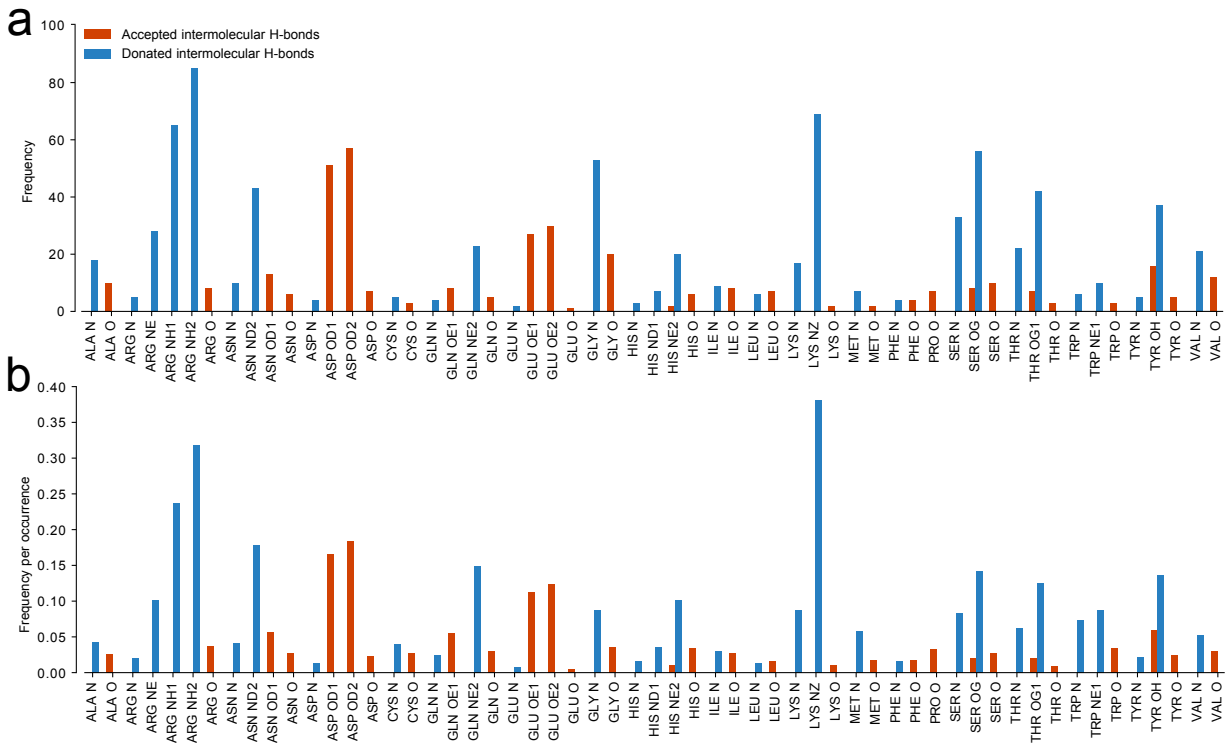


Figure 2.6: Intermolecular H-bonds formed by each amino acid atom type in ligand binding sites. (a) The frequency of H-bonds to ligand by atom type in 136 protein complexes. (b) The frequency per binding site occurrence of H-bonds to ligand. Pro N is omitted, because it lacks an amide proton to donate.

2.4.4 When protein and ligand atoms are categorized according to their chemistry, are H-bonding preferences between proteins and ligands fundamentally similar or different?

Protein atoms forming H-bonds with ligands were divided into main chain versus side chain categories (Figure 2.7), and their H-bonds were tabulated according to atomic chemistry for keto oxygens (O), hydroxyl groups (OH), carboxylate oxygens (COO^-), and amine nitrogens (NH and NH_2). Amine donors were found to dominate in the total number of H-bonds formed with ligands, with almost equal representation from main and side chain amines (Figure 2.7a). However, when normalized by the number of binding site occurrences, side chain amines were found to form 16 times as many ligand H-bonds as main chain amines (Figure 2.7b). Hydroxyl groups donate a meaningful, though lesser, number of H-bonds to ligands (about one-fourth as many as amine groups donate) and rarely act as acceptors.

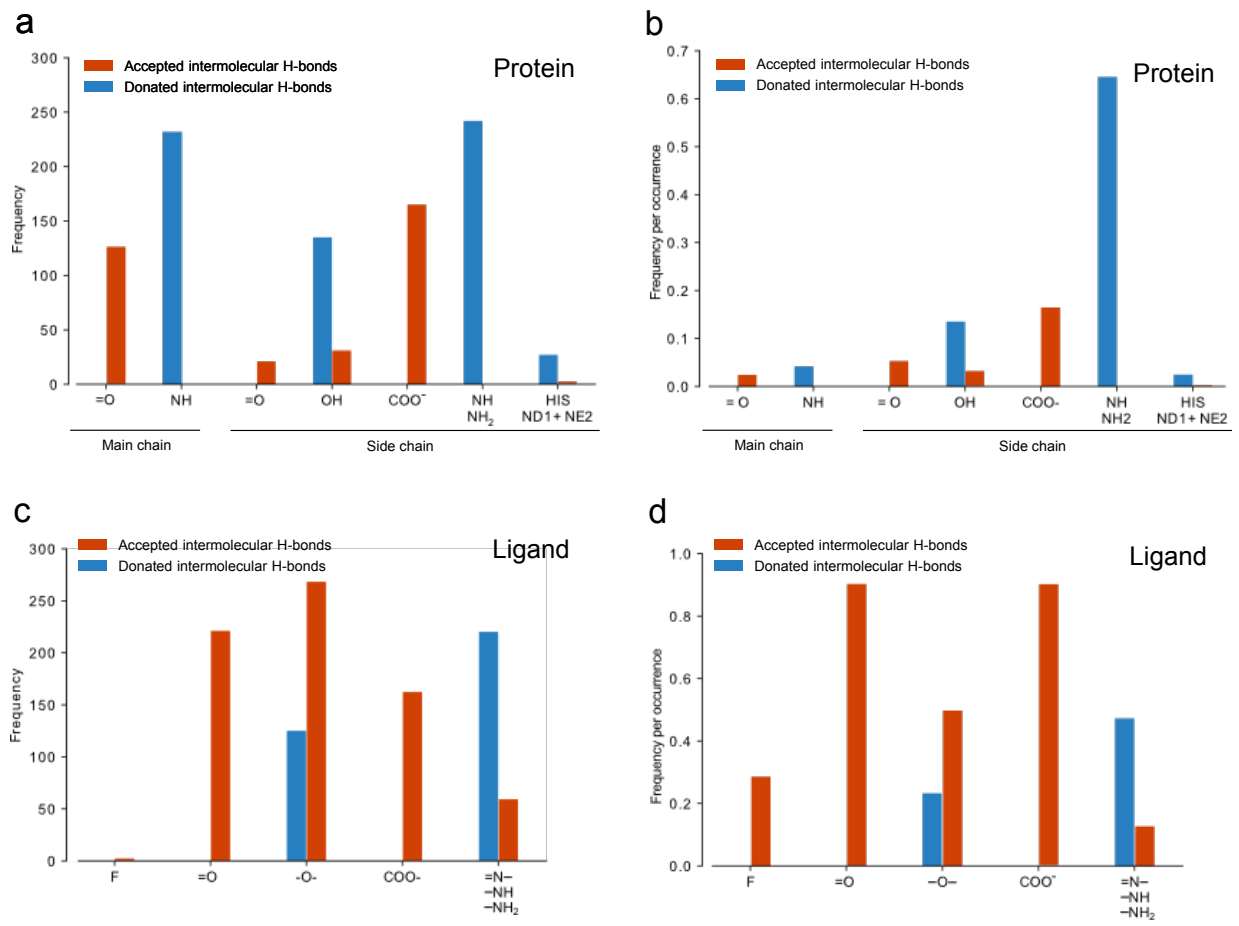


Figure 2.7: Comparison of the chemistry and prevalence of atoms forming intermolecular H-bonds, by protein versus ligand side of the interface. The bar plot in (a) shows the frequency of protein atoms participating in H-bonds to ligands in the 136 complexes, while (b) shows the same data normalized by the number of binding site occurrences, yielding the average number of H-bonds to ligand per atom type. The bar plots (c) and (d) show the same data from the ligands' perspective. In panels (c) and (d), F indicates fluorine. The label =O includes O.2 (sp^2-hybridized) oxygen atoms; –O– includes O.3 (sp^3-hybridized) hydroxyl and ester oxygens; COO^- includes O.co2 oxygen atoms in carboxylate, sulfate, and phosphate groups; and =N, –NH, and $–NH_2$ includes N.2, N.3, N.am, N.ar. For COO^- groups, each terminal oxygen was tabulated separately.

Surprisingly, the trends for H-bond donors and acceptor chemistry in ligands are quite different, and less influenced by the charge magnitude (Figure 2.7c and 2.7d). Neutral keto (=O) and ester + hydroxyl oxygen acceptors (–O–) of H-bonds predominate in the total number of H-bonds formed to proteins (Figure 2.7c), followed by carboxylate oxygens. Carboxylate oxygens are as important as keto oxygen acceptors for protein H-bonds when normalized by the total number of atom occurrences in the ligands, followed by fluorine atoms, then amines (which seldom act as acceptors). Consistent with the observed strong trend for ligands to accept rather than donate H-bonds to proteins, ligand hydroxyl and amine donors only form one-third as many protein H-bonds per occurrence when compared to oxygen acceptors.

These results are also consistent with results from an earlier analysis of water molecules forming H-bonded bridges between proteins and ligands (A. Cayemberg and L. A. Kuhn, unpublished results) in a set of 20 non-homologous complexes (Raymer et al., 1997). There, without defining donor or acceptor roles, we discovered that water molecules H-bonding directly to both the protein and ligand interacted with oxygen atoms on the ligand 74% of the time and nitrogen atoms only 25% of the time (with Cl atoms representing the final 1%). The same was true for water molecules forming di-water bridges between protein and ligand, with a 76% preference for interacting with oxygen on the ligand.

2.4.5 Do different classes of ligand differ in their tendency to accept versus donate H-bonds?

For this analysis, the 136 complexes were considered from the ligand perspective, with the 25 peptidyl, 50 nucleotide-like, and 61 other small organic ligands analyzed as individual sets (Table 2.1). The 2:1 ratio for ligands to accept rather than donate H-bonds to cognate proteins was seen for both nucleotidyl and other organic ligands (Figure 2.8). Peptidyl ligands, on the other hand, showed no strong preference for donating versus accepting H-bonds. This is expected, because of the fundamental chemical and evolutionary parity between the peptides and proteins in these complexes: both cannot act primarily as donors and still make sufficient intermolecular H-bonds. The more polar, often charged, nucleotidyl ligands formed 50% more H-bonds with proteins than

the other organic molecules. This is in line with the observation (Figure 2.6) that charged protein side chains play a more important role than neutral side chains in H-bonding to ligands. The strength of an H-bond also increases with the magnitude of the complementary charge on the participating atoms (Shan & Herschlag, 1996). However, the greater number of H-bonds for nucleotidyl ligands could also reflect their greater number of heavy atoms, 31.7 +/- 11.2 on average, relative to other organic molecules, 17.8 +/- 10.8. The average number of heavy atoms for peptidyl ligands was 27.8 +/- 21.1. These results indicate that strong H-bonds involving charged groups are common in

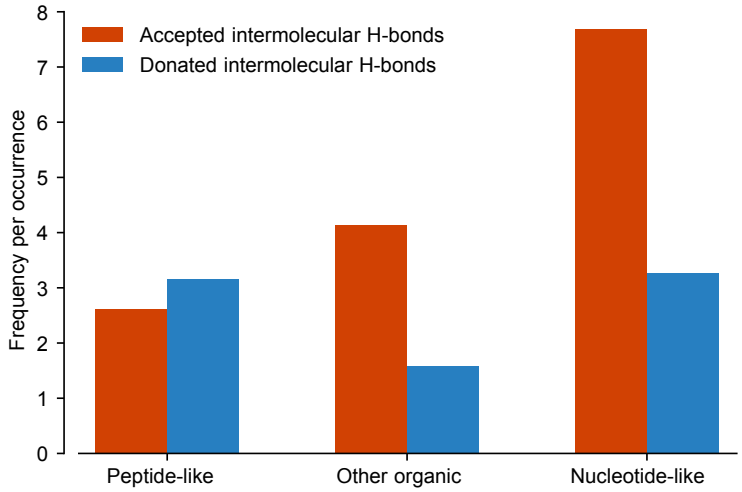


Figure 2.8: The average number of H-bonds donated or accepted for each ligand type. The bar plot shows the average number of H-bonds donated (blue) or accepted (red) for each ligand type: peptidyl, nucleotide-like, and other organic.

cognate protein-ligand complexes. The prevalence of strong H-bonds involving very polar groups is not necessarily expected, given that ligands need to be released from their proteins as part of the enzymatic, signaling, or transport cycles. Strong H-bonds also contribute to formation of the catalytic transition state between enzymes and their biological ligands (Shan & Herschlag, 1996).

Given the visual observation of dense protein networks of nitrogen H-bond donors interacting with the nucleotides, cluster analysis was performed on the vectors representing H-bond patterns along the sequence to discern any similar patterns of ligand H-bonding across the 136 complexes. Nucleotidyl ligands were the only ligand class for which a clear local pattern of H-bonding appeared, involving at least 3 nitrogen H-bond donor groups separated by no more than 10 residues (Figure

2.9). Three to five H-bond donors occurred within six residues of the amino acid sequence (positions 1-6 in Figure 2.9) in 18 of the 24 cases (rows 2-19). Thirteen of the 18 patterns involved nucleotidyl ligands, with the sequence pattern **Gly-Lys-(Ser,Thr)-(Thr,Ser,Tyr,**Cys,Ala) found in 7 cases. (Boldface indicates the dominant residue type(s) and regular font indicates other allowed residues.) This structural motif turns out to be the P-loop nest for phosphate binding (Bianchi et al., 2012), a strong and particularly geometrically ordered example showing the tendency for proteins to donate H-bonds to ligands (Figure 2.10). The program for creating these PyMOL H-bond interaction views from Hbind tables, as shown in this figure, is freely available to researchers at https://github.com/psa-lab/Hbind-interaction-viz.

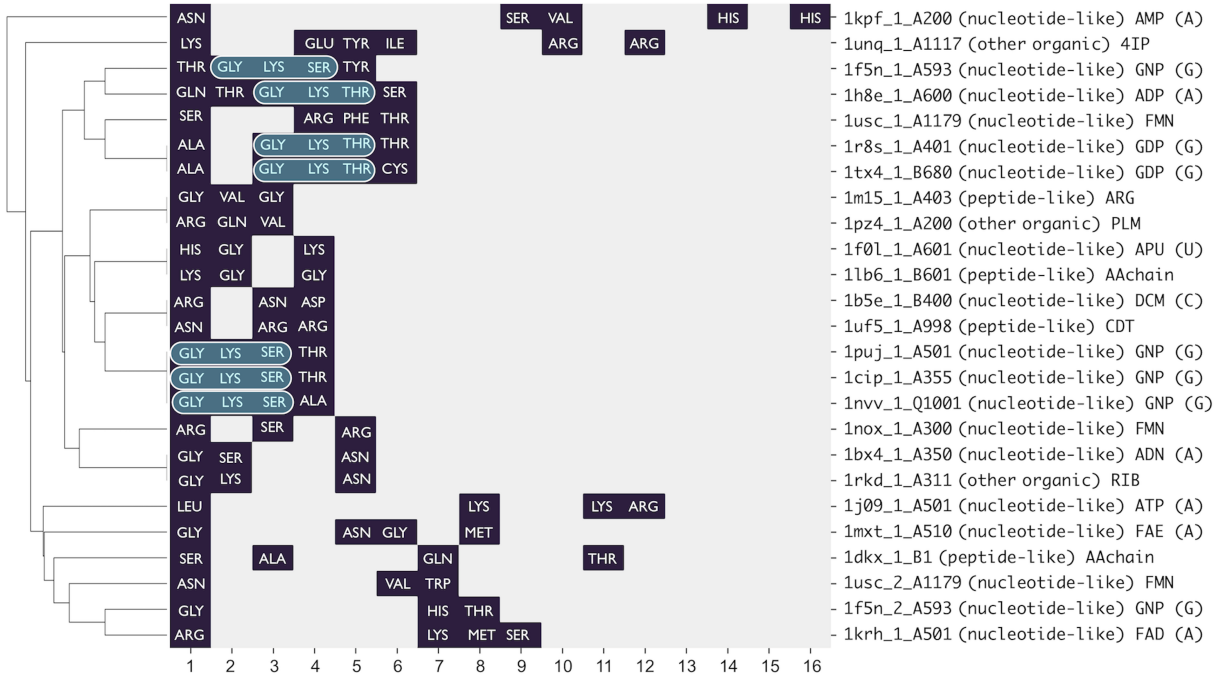


Figure 2.9: Clustered patterns of H-bonds to ligands that are localized in the protein sequence and involve three or more nitrogen donors. The x-axis indexes from the first to the sixteenth position in all amino acid sequences with no more than 10 residues between adjacent H-bond donors to ligand. The label in the rightmost column provides the PDB code and index of the H-bond pattern (1, 2, etc.) in a given protein, the chain ID and residue number of the ligand in the PDB structure file, the ligand category (nucleotide-like, peptidyl, or other organic), and the 3-letter ligand name in the PDB. Where appropriate, the base (adenine, A; guanine, G; or C, deoxycytidine) present in the nucleotide-like ligands is provided at the end of the label. Highlighted in blue are the Gly-Lys-Ser/Thr motifs found hydrogen-bonding to phosphate groups in seven of the nucleotidyl ligands. Clustering shown on the left indicates the degree of similarity in the pattern of H-bonds, with each difference in presence/absence of an H-bond (not amino acid identity) counting as 1 distance unit.

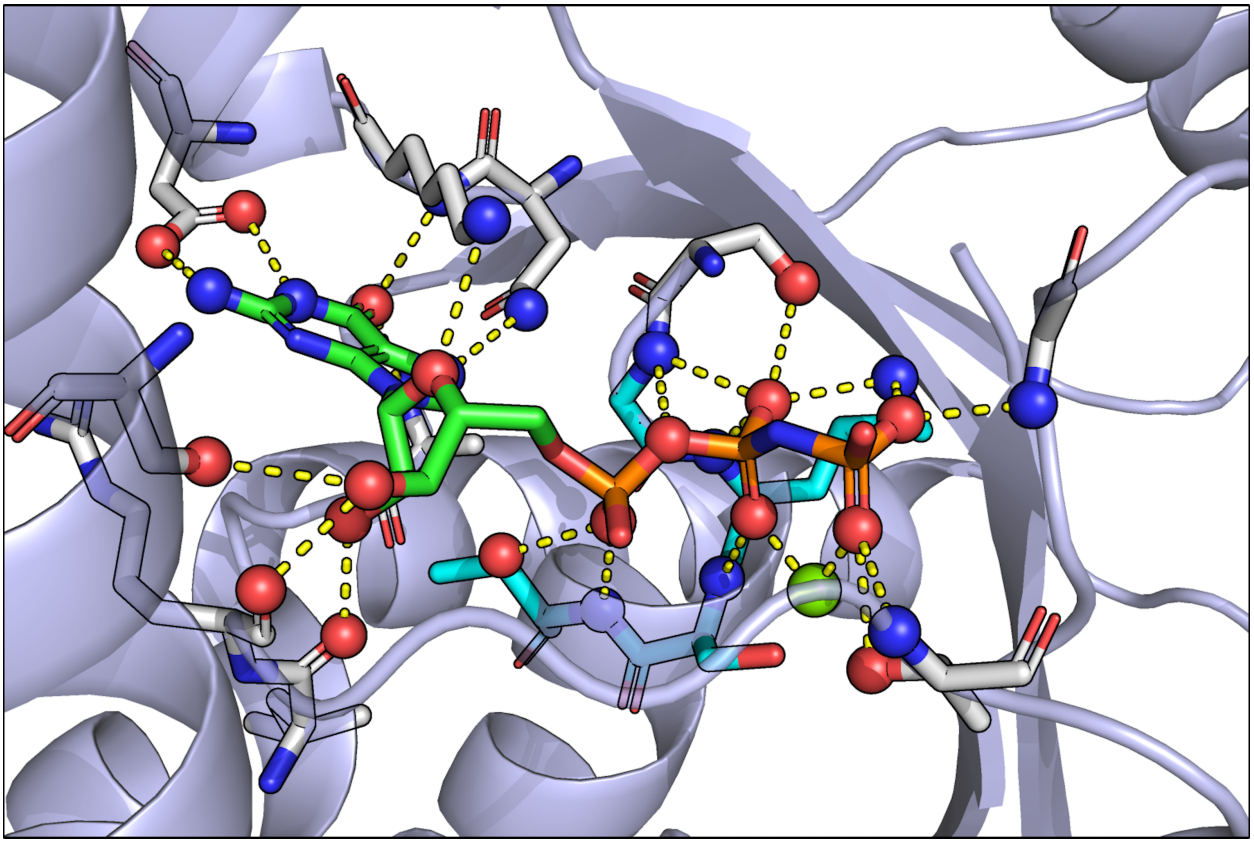


Figure 2.10: P-loop nest motif Gly-Lys-Ser-Thr for phosphate binding. This figure shows an example of an observed P-loop nest motif donating a local network of protein H-bonds to the oxygen-rich triphosphate group (carbon atoms are shown in cyan). This example is from a high-resolution G protein structure in complex with GTP (PDB entry 1cip; Coleman & Sprang, 1999). H-bonds forming the P-loop nest interaction are shown as yellow dashed lines, and polar atoms participating in these interfacial interactions appear as red spheres for oxygen atoms, blue spheres for nitrogen atoms, and a green sphere for the bound Mg^{2+}. For clarity, hydrogen atoms are omitted.

2.4.6 Can orientational selectivity of the biological ligand explain the preference for proteins to donate H-bonds to ligands?

Here we evaluate whether geometrical aspects of H-bond interactions, in particular the angular dependence of H-bonds, can provide a ligand-selectivity advantage in proteins that donate H-bonds to ligands more often than accepting them. Underlying protein-ligand binding is a 3D code defined by structure and chemistry that determines which ligands can bind to a protein, as well as avoiding binding to ligands that inappropriately alter activity.

An interesting example of how strong selectivity for a molecular partner can confer a functional

advantage is the observation that narrow-spectrum (more selective) antibiotic ligands avoid drug resistance much more effectively than broad-spectrum antibiotics (Palumbi, 2001). Narrow-spectrum antibiotics form interactions that are highly tuned for their protein target, which means that the protein must accumulate more mutations to abrogate binding by a narrow-spectrum antibiotic, relative to broad-spectrum antibiotics. The same effect, in the absence of any mutations, allows proteins that form many ligand-selective interactions to prevent misrecognition and binding to the wrong partners.

The simplest case supporting a hypothesized preference for proteins to use more chemically or geometrically selective interactions in ligand binding is the observed 3:1 preference for proteins to use amines relative to hydroxyl groups in ligand H-bonds (797 amine-involving H-bonds versus 258 hydroxyl-involving H-bonds; Table 2.2). This is despite the potential of the hydroxyl group to accept two H-bonds and donate one, which allows about 1.5 times as many H-bonds to the ligand relative to the most common protein amine groups (NH and NH_2). In general, protein lone pairs available to accept H-bonds are 1.6 times as prevalent as protons available to donate (Figure 2.5). However, the hydroxyl group is less selective in its interactions, allowing both donor and acceptor groups on ligand partners, which may result in insufficient selectivity for the correct ligand relative to the thousands of alternative molecules in the cell.

Ligand selectivity can also be conferred by the difference in geometrical constraints on donor versus acceptor interactions. To quantitate examine how selectivity relates to the 3D geometry of interaction, the favored angular and donor-acceptor distance ranges are shown for H-bond acceptor and donor atoms (Figure 2.1). A favorable donor–H$\cdots$acceptor angle θ range of 120–180° in well-resolved crystal structures, in combination with a favorable donor–acceptor separation of 2.4-3.5 Å (McDonald & Thornton, 1994; Ippolito et al., 1990), results in a significantly smaller volume (30.3 $Å^3$) in which a ligand acceptor atom can favorably interact with a protein donor atom, in comparison with the volume in which a ligand proton can favorably interact with lone pairs on a protein acceptor (60.6 $Å^3$). This is partly due to the more permissive pre-acceptor–acceptor–H angle (ϕ) of 90-180° (relative to the θ constraint on donor–H$\cdots$acceptor angle), and also due to the

presence of two lone pairs on the majority of H-bond acceptor atoms in proteins (oxygens). The two lone pairs create a large, continuous volume in which a proton can H-bond with the acceptor atom. The observed distribution of donor atoms relative to oxygen acceptor atoms in well-resolved protein X-ray structures (McDonald & Thornton, 1994) indicates there are few constraints on out-of-plane interactions with acceptor lone pairs, resulting in an almost isotropic, hemispheric shell of donor proton positions relative to the acceptor.

An evolutionary emphasis on matching large volumes of favorable interaction around acceptors on the protein might well result in too little selectivity for the cognate ligand. While proteins, with their current amino acid content, cannot avoid the presence of oxygen atoms on the surface, nor do proteins entirely avoid ligand interactions with acceptors, we hypothesize that cognate protein-ligand interactions may have evolved to favor the use of donor groups on the protein to create small volumes that the arrangement of acceptor atoms on cognate ligands must uniquely match. This is supported by the enhancement of oxygen atoms on small molecule ligands (Figure 2.7c). It is also supported by an observation of Taylor et al. (Taylor & Kennard, 1984): though the majority of intramolecular N–H$\cdots$O H-bond angles are in the 100–140° range, intermolecular N–H$\cdots$O angles are typically much more linear (170–180°), corresponding to stronger H-bonds as well as a narrow tolerance to be met in recognizing the cognate ligand. Donation of H-bonds to the ligand is, of course, one component of recognition. Shape complementarity, hydrophobic surface matching, interfacial ion binding, and additional H-bonds including water-mediated interactions (Raymer et al., 1997; Taylor & Kennard, 1984; Arkin & Wells, 2004; Kuhn et al., 1995) complete the selection of and enhanced affinity for the native ligand.

Another selective advantage that could drive the evolution of strong donor patterns (rather than mixed donor-acceptor patterns) for ligand binding, is to disfavor aberrant protein-protein interaction. Binding site donor geometries that evolved to match a small molecule ligand could not easily be satisfied by other proteins, which on average also favor binding site donor patterns that would tend to repel interaction with other sets of donors. Finally, the finding that asymmetry in packing of the peptide amide dipole results in larger positive than negative regions in proteins (Gunner et al., 2000)

would tend to enhance the preference for proteins to interact with more electronegative, lone-pair bearing atoms.

2.4.7 Do protein-bound metal ions contribute significantly to ligand binding, and how does their bond chemistry relate to observed trends in H-bonding?

When protein-bound metal ions were found in the ligand interface, they were included in the analysis. Table 2.3 provides detailed statistics, while Figure 2.11 summarizes ligand interactions per occurrence for the 8 metal types observed in the 136 complexes. Mg^{2+} was by far the most common, with 24 occurrences, followed by Mn^{2+} with 14 occurrences. All other metal types were present 7 or fewer times. Ni^{2+}, Mg^{2+}, Cd^{2+}, Mn^{2+}, Co^{2+}, and Na^{+} each accounted for 1-2 direct ligand bonds per occurrence (using bond-length criteria listed in Methods), while Fe (exhibiting various oxidation states in the different complexes) and Zn^{2+} averaged half an interaction per occurrence. Metal interactions with lone pairs on electronegative atoms within bonding distance, as measured here, are almost covalent in strength. This makes them significant contributors to the enthalpy change upon complex formation. Because these metals are positively charged, the trend in polarity of the interface is like the dominant H-bond classes observed above, with a positively charged group on the protein side forming a bond with a lone pair of electrons on the ligand.

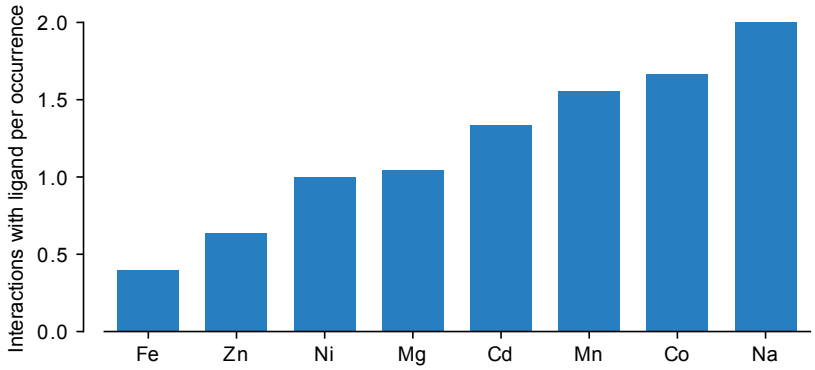


Figure 2.11: The average number of bonds to ligand formed per occurrence by protein-bound metal ions in the 136 complexes.

Table 2.3: Statistics of ligand-metal interactions.

Metal ion	Ligand Interactions	Interfacial Occurrences	Interactions/Occurrence
CD	4	3	1.33
CO	5	3	1.67
FE	2	5	0.4
MG	24	23	1.04
MN	14	9	1.56
NA	2	1	2
NI	2	2	1
ZN	7	11	0.64

2.4.8 How do these results relate to Lipinski's rule of 5 for drug-likeness in small molecules?

In the late 1990s, Lipinski and colleagues at Pfizer undertook a study of 2,245 small molecules from the World Drug Index considered to have superior physicochemical properties, based on meeting solubility and cell permeability criteria required for entry into Phase II clinical trials (Lipinski et al., 1997). The drug-like criteria for small molecules derived from their analysis of the 2,245 compounds are known as the Rule of 5. Poor absorption or permeability tends to occur for a compound matching any of the Rule of 5 features: more than 5 H-bond donors, more than 10 H-bond acceptors, a molecular weight greater than 500 Da, or a calculated logP value of greater than 5, defined as the logarithm of the partition coefficient between n-octanol and water. These criteria remain widely used for selecting sets of molecules for virtual or high-throughput experimental screening, and as ideal physicochemical ranges to match when redesigning lead compounds to bind with higher affinity or better bioavailability.

The Rule of 5 criteria were not derived to predict molecules as effective protein ligands. However, most drugs do target proteins, and thus the Rule of 5 criteria may select for the ability to bind proteins as well as enter the cell. In fact, the maximum H-bond acceptor to donor ratio in the Rule of 5 (10:5) matches the trend found here: twice as many H-bonds being accepted by ligands (5 on average) as donated (2.5 on average; Figure 2.3). The two-fold preference for ligand acceptors relative to donors in H-bonding may therefore be a molecular mechanism underlying the drug-like criteria in the Rule of 5. Additionally, the ability of H-bond acceptor and donor numbers to predict drug-likeness suggests that the trends identified in this paper can also be useful for predicting ligand

interactions.

2.4.9 Can the observed H-bonding trends be used to predict protein-ligand interactions?

We addressed the question of whether the observed prevalence of H-bond acceptors and donors in the 136 complexes, tabulated by PDB atom type for protein binding site atoms (e.g., Arg O, N, NE, NH1, and NH2) and by MOL2 atom type for ligand atoms (e.g., O.2, O.3, N.2, N.3, etc.), can be used to predict the cognate protein-ligand orientation from a series of dockings of the small molecule. To test this, we used 10 ligand dockings on average in each of 30 protein-small molecule complexes that were recently used in a comparison of docking scoring functions and do not overlap with the 136 complexes (Raschka et al., 2016a; Table 2.4). The crystallographic binding pose was not included, because the correct pose is unknown in a predictive study and therefore never exactly sampled. Secondly, many scoring methods can readily detect the crystallographic pose as the global optimum due to their parameterization, suggesting excellent accuracy when the crystal pose is included; a much more realistic assessment of their real-world performance is the identification of near-native poses. The best-sampled ligand docking poses here ranged from 0.1-1.4 Å RMSD relative to the crystallographic position across the 30 complexes, as shown by the green cumulative distribution curve in Figure 2.12. The goal of this analysis of docked positions was not to develop a new scoring function, but to assess whether the H-bond interaction statistics accumulated across 136 structures capture the essential molecular recognition features that occur within individual structures sufficiently well to discriminate native or near-native interactions.

For the protein H-bond component of the scoring function, the frequency scores of all protein atoms observed to make an H-bond with the ligand were summed, based on the raw data compiled across the 136 complexes, with sample data shown below. In the first entry,

{Acceptor: 0, Donor: 18},

indicates that in the 136 complexes, alanine main chain nitrogen atoms accepted H-bonds from the ligand 0 times and donated H-bonds to ligands 18 times.

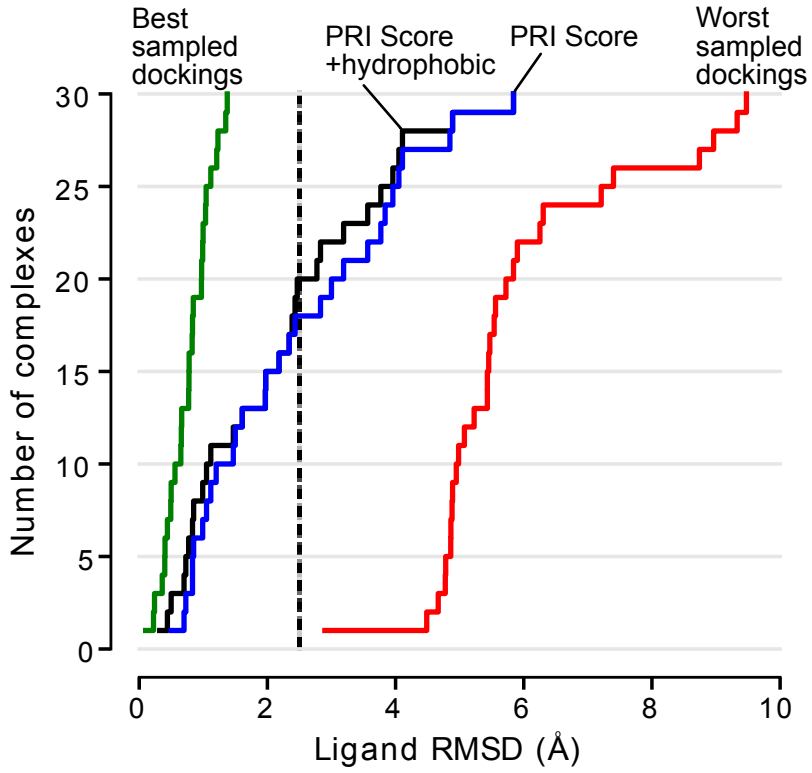


Figure 2.12: Enrichment plot comparing the Protein Recognition Index to other scoring functions. The enrichment plot shows the degree of native-likeness (RMSD relative to crystallographic position) of docking, showing the highest Protein Recognition Index for all 30 complexes. The ligand orientation for each complex was predicted according to the highest H-bond PRI value (blue trace) or the highest PRI+ hydrophobic contact score (black trace) among all the ligand orientations. All 30 complexes' best-scoring ligand orientations were then compiled, and their RMSD values relative to the crystallographic position were sorted from best (closest to 0 Å) to worst RMSD (4-5 Å). These RMSD values were then plotted as a cumulative distribution function of the number of ligand orientations selected to within X Å RMSD of the crystallographic position. For instance, all ligand orientations selected by either PRI or PRI + hydrophobic scoring that appear to the left of the dashed black vertical line at ligand RMSD = 2.5 Å were within 2.5 Å RMSD of the crystallographic position. This was true for 18 of the PRI scored complexes and 20 of the PRI + hydrophobic scored complexes. The result that would be obtained by the best-possible scoring of ligand orientations (selecting the best-sampled docking of the ligand for each complex) is shown by the green trace. The result from selecting the worst docking (highest RMSD position) of each ligand across the 30 complexes is shown by the red trace.

Table 2.4: Thirty protein-ligand complexes analyzed for predicting the native binding mode using H-bond statistics. The dataset consists of 19 complexes for which the holo conformation was used for docking, and 11 complexes for which the apo conformation of the protein was used. More details are available in (Raschka et al., 2016a).

PDB entry (holo/apo)	Protein	Ligand	Resolution (Å)	Holo-apo binding site RMSD (Å)
1a9x / -	carbamoyl phosphate synthetase	L-ornithine	1.8	-
1amu / -	gramidicin synthetase 1	L-phenylalanine	1.9	-
1b5e / -	deoxycytidylate hydroxymethylase	deoxycytidylic acid	1.6	-
1bgv / -	glutamate dehydrogenase	L-glutamate	1.9	-
1bx4 / -	adenosine kinase	adenosine	1.5	-
1c96 / -	mitochondrial aconitase	citrate anion-iron/sulfur cluster	1.81	-
1cbs / -	retinoic acid binding protein	retinoic acid	1.8	-
1cbx / -	carboxypeptidase A	L-benzylsuccinic acid	2	-
1ccw / -	glutamate mutase	D-tartaric acid	1.6	-
1chm / -	creatine amidinohydrolase	carbamoyl sarcosine	1.9	-
1com / -	chorismate mutase	prephenic acid	2.2	-
1coy / -	cholesterol oxidase	dehydroepiandrosterone	1.8	-
1cps / -	carboxypeptidase A	sulfodiimine	2.25	-
1did / -	D-xylose isomerase	2,5-dideoxy-2,5-imino-D-glucitol	2.5	-
1hwr / -	HIV-1 protease	Xk216	1.8	-
1rx1 / -	dihydrofolate reductase	NADP+	2	-
3ks9 / -	metabotopic glutamate receptor	Z99	1.9	-
3odu / -	G-protein-coupled chemokine receptor	IT1t	2.5	-
7tim / -	triosephosphate isomerase	phosphoglycolohydroxamic	1.9	-
10gs / 16gs	glutathione S-transferase	L-cysteine amide	2.20 / 1.90	0.27
1ahb / 1ahc	alpha-momorcharin	formycin-5'-monophosphate	1.90 / 2.00	0.75
1aj2 / 1ajz	dihydropteroate synthase	pterin diphosphate	2.20 / 2.00	0.64
1gmr / 1gmq	ribonuclease	guanosine-2'-monophosphate	1.77 / 1.80	0.46
1kel / 1kem	sulfide oxidase antibody	methylphosphonic acid	1.90 / 2.20	0.68
1nsc / 1nsb	influenza B neuraminidase	O-sialic acid	1.70 / 2.20	0.32
1swd / 1swa	streptavidin	biotin	1.90 / 1.90	0.52
3tmn / 1tli	thermolysin	tryptophan	1.70 / 2.05	0.69
1tmt / 1vr1	alpha-thrombin	D-phenylalanine	2.20 / 1.90	0.66
1ydb / 1ydc	carbonic anhydrase II	acetazolamide	1.90 / 1.95	0.3
5sga / 2sga	proteinase A	acetyl group	1.80 / 1.50	0.19

ALA:

N: {Acceptor: 0, Donor: 18}

O: {Acceptor: 10, Donor: 0}

ARG:

N: {Acceptor: 0, Donor: 5}

NE: {Acceptor: 0, Donor: 28}

NH1: {Acceptor: 0, Donor: 65}

etc.

So, for instance, if you were to score a ligand orientation accepting an H-bond from the main chain

N in Ala and H-bonds from both Arg NE and Arg NH1, the protein H-bond score for that binding mode would be:

18 + 28 + 65 = 111

The higher the score, the more the docking reflects the known preferences in the 136 complexes for H-bonds donated or accepted by the protein. This Protein Recognition Index, or PRI-prot, differs from the typical scoring of H-bonds in protein-ligand docking, because here the contribution of each H-bond is weighted according to the prevalence of intermolecular H-bonds involving this protein atom type in crystal complexes. Scoring is performed the same way for the ligand side of the interaction, leading to a PRI-lig value. Standardization is then performed on the PRI-lig values across the dockings for a given complex, rescaling such that the score distribution has a mean value of 0 and a variance of 1. This converts the PRI-lig to a Z-score measured in standard deviations above or below the mean (more favorable or less favorable, statistically). The same standardization is performed for protein PRI-prot values across the dockings, putting the ligand and protein PRI values on the same scale. PRI-prot and PRI-lig values can then be summed (reflecting the simplest possible weighting, giving even importance to the protein and ligand side of the interface), to yield what we call the PRI. High PRI values reflect that the H-bond groups linked between protein and ligand in the current ligand orientation match the H-bond preferences found in the 136 unrelated complexes.

For a series of ligand dockings in a given protein, the docking with the highest PRI is predicted as the most native-like complex. This process was performed for all 30 complexes, and the results are summarized in Figure 2.12. To consider the extent to which hydrophobic contacts add information for defining the cognate ligand orientation, we created a variant of PRI that includes an equal-weighted, standardized hydrophobic contact term (PRI+hydrophobic). The hydrophobic term counts the number of carbon-carbon and carbon-sulfur contacts (atom centers within 4 Å) between the protein and ligand, as reported in the Hbind software output (Figure 2.2). Software we used to compute the Protein Recognition Index (and its PRI-prot and PRI-lig) components is being made available at https://github.com/psa-lab/PRI-protein-recognition-index. We envision

this software will be a broadly useful tool for assessing the native-likeness of designed or predicted protein-ligand interfaces, as well as for guiding protein mutagenesis to identify ligand binding residues and predict ligand binding sites (using the PRI-prot component alone) or assessing ligand physicochemical suitability for a protein target (using the PRI-lig component alone).

The results of the ligand orientation prediction enrichment plot (Figure 2.12) clearly show that the statistical information encoded in the Hbind hydrogen-bonding preferences of different atom types is able to identify near-native ligand orientations, selecting an orientation within 2.5 Å RMSD of the crystallographic position in two-thirds of the complexes. Adding a hydrophobic contact term leads to a slight improvement in prediction, while the H-bonding preferences account for most of the predictive power. Measuring the Pearson linear correlation coefficient (r) between the PRI values, PRI+hydrophobic values, and two commonly used docking scoring functions, AutoDock Vina (version 1.1.2; http://vina.scripps.edu; Trott & Olson, 2010) and DSX (also known as DrugScore X; version 0.88; http://pc1664.pharmazie.uni-marburg.de/drugscore; Neudert & Klebe, 2011) across 300 dockings for the 30 complexes, show that the PRI value is almost uncorrelated with the scores from AutoDock Vina (r = -0.26) and DSX (r = -0.19), despite these scoring functions also including H-bond interaction terms. This indicates that PRI provides new information that has high predictive value on its own, while also easily being combined with existing protein-ligand scoring metrics. Weighting H-bonds according to their statistical prevalence by atom type measures a chemical aspect of protein-ligand recognition that is both predictive of native interactions and not reflected in the other measures.

2.5 Conclusions

To address the question that motivated this work – whether proteins tend to donate rather than accept H-bonds when binding biological small molecules – a utility called Hbind was developed to label the donor/acceptor capacity of each atom, and characterize each H-bond in terms of its atomic chemistry and geometry. Making this software available allows such data to be generated readily and analyzed for a range of other interesting questions with the vast crystal structure data now

available. Handling both protein and ligand chemistry at the atomic rather than coarser functional group or side chain levels allowed an in-depth analysis of the trends and potential underlying mechanisms in ligand recognition by proteins. Our conclusions were:

- Across 136 non-homologous protein complexes including a mix of nucleotide-like, peptidyl, and other organic ligands, the proteins were found to donate twice as many H-bonds as they accepted from ligands.

- Lone pairs available to accept H-bonds are actually 1.6 times as prevalent as protons available to donate, both on the protein and ligand side of the interface. Thus, the relative availability of donor and acceptor groups does not explain the trend for proteins to preferentially donate H-bonds to their ligands.

- A corresponding, strong preference for ligands to accept H-bonds from proteins suggests that focusing on the prevalence and positioning of H-bond acceptors in both designed ligands and molecules assessed in screening (that is, a more detailed, structural measure of "drug-likeness") is likely to result in ligands that better match the protein-encoded determinants for binding. The Protein Recognition Index (PRI) software was designed for this purpose.

- Nitrogen atoms served as donors for 76% of the intermolecular H-bonds and hydroxyl groups in 24%, considering both protein and ligand donors together. This suggests that amine nitrogens are much more effective donors in biological complexes than hydroxyl groups, providing another straightforward way to enhance molecular design.

- The side chains in proteins most likely to donate H-bonds to ligands are Arg and Lys, with Asn and Gln being about half as important. Asp and Glu are the side chains most likely to accept H-bonds from ligands. Polar H-bonds are apparently favored in the underlying code of molecular recognition. These results suggest focusing on these side chains when predicting binding sites or carrying out experiments to identify key H-bonding groups within a site.

- Metals bound in protein ligand-binding sites are not a dominant feature. Most metal ions in binding sites account for 1-1.7 bonds to the ligand, on average, with Fe and Zn accounting for fewer ligand interactions (0.5, on average) in the 136 complexes. While these bonds are occur less frequently, their almost-covalent strength makes them important contributors to affinity.

- These trends, analyzed from all angles, indicate a surprising degree of interfacial polarity for non-peptidyl organic molecule complexes with proteins, favoring donors on the protein side and acceptors on the ligand side, with amine donors and oxygen acceptors pairing in the vast majority of intermolecular H-bonds.

- By developing software to calculate a Protein Recognition Index (PRI), measuring the similarity between H-bonding features in a given complex (predicted or designed) and the characteristic H-bond trends from crystallographic complexes (Figure 2.7), we show that the cognate orientation between protein and ligand can be predicted from this information alone. The PRI for a set of protein or ligand atoms can also be calculated, to discern the extent to which their H-bonding groups match the favored distribution of donor and acceptor atom types in known complexes.

- The 2:1 acceptor to donor ratio observed here for ligand atoms forming H-bonds to proteins appears to be a structural explanation for the 2:1 ratio of the number of ligand H-bond acceptor atoms to donor atoms in Lipinski's Rule of 5. We anticipate the Protein Recognition Index may prove similarly useful in guiding protein and ligand design to design more selective and tighter-binding complexes.

- The trend for proteins to donate H-bonds to their cognate ligands, especially via amine donor groups, may have evolved as a ligand selectivity determinant. Amine donors have relatively narrow angular constraints and volumes in which an acceptor group can form an energetically favorable H-bond. Two acceptor lone pairs are present on the oxygen atoms in proteins, and a consequence is that the lone pairs present a broad surface and volume for favorable interaction

with donor atoms (twice that of an NH donor interacting with an acceptor group). Molecular evolution is expected to favor a narrow selection of ligand partners due to the potential for misrecognition if many ligands could easily match H-bonding groups in a protein pocket. The relative orientation and spacing of these groups is also an extremely important aspect of the code for matching H-bonds between protein and ligand.

2.6 Acknowledgements

This research was supported by funding from the Great Lakes Fishery Commission (Project ID: 2015_KUH_54031). We gratefully acknowledge OpenEye Scientific Software (Santa Fe, NM) for providing academic licenses for the use of their of QUACPAC (molcharge) and OEChem software. We also thank these lab graduates for their early motivations or contributions to this research: Dr. Maria Zavodszky (now at GE Global Research Center), who observed that hydroxyl-rich ligands tended to result in false positives in screening, Dr. Amy Cayemberg McQuade (now at Carroll University) for carrying out the statistical analysis of protein-water-ligand bridges, and Dr. Jeffrey VanVoorst (now at Veritas Technologies, LLC) for developing the non-homologous dataset of 136 protein-small molecule complexes.

CHAPTER 3

DETECTING THE NATIVE LIGAND ORIENTATION BY INTERFACIAL RIGIDITY: SITEINTERLOCK

Adapted with permission from

Raschka, Sebastian, Joseph Bemister-Buffington & Leslie A. Kuhn. 2016. "Detecting the native ligand orientation by interfacial rigidity: SiteInterlock."

Proteins: Structure, Function, and Bioinformatics 84(12). 1888–1901.

3.1 Abstract

Understanding the physical attributes of protein-ligand interfaces, the source of most biological activity, is a fundamental problem in biophysics. Knowing the characteristic features of interfaces also enables the design of molecules with potent and selective interactions. Prediction of native protein-ligand interactions has traditionally focused on the development of physics-based potential energy functions, empirical scoring functions that are fit to binding data, and knowledge-based potentials that assess the likelihood of pairwise interactions. Here we explore a new approach, testing the hypothesis that protein-ligand binding results in computationally detectable rigidification of the protein-ligand interface. Our SiteInterlock approach uses rigidity theory to efficiently measure the relative interfacial rigidity of a series of small-molecule ligand orientations and conformations for a number of protein complexes. In the majority of cases, SiteInterlock detects a near-native binding mode as being the most rigid, with particularly robust performance relative to other methods when the ligand-free conformation of the protein is provided. The interfacial rigidification of both the protein and ligand prove to be important characteristics of the native binding mode. This measure of rigidity is also sensitive to the spatial coupling of interactions and bond-rotational degrees of freedom in the interface. While the predictive performance of SiteInterlock is competitive with the best of the five other scoring functions tested, its measure of rigidity encompasses cooperative rather than just additive binding interactions, providing novel information for detecting native-like complexes. SiteInterlock shows special strength in enhancing the prediction of native complexes by ruling out inaccurate poses.

3.2 Introduction

3.2.1 Stabilization of protein complexes by ligand binding

Experimental methods that probe the relationship between protein order, stability, and ligand binding have proven increasingly useful in structure determination and ligand screening. For instance, thermal shift assays such as differential scanning fluorimetry (DSF) and calorimetry

measure the temperature at which a protein gains or loses structural integrity. Taking advantage of the tendency for ligand binding to shift the unfolding equilibrium toward the native state and for ligand binding to increase the melting temperature (Niesen et al., 2007; Brandts & Lin, 1990) DSF has become important for high-throughput drug discovery (Pantoliano et al., 2001) and the discovery of ligands that stabilize proteins for structure determination (Ericsson et al., 2006; Vedadi et al., 2006). Nuclear magnetic resonance (NMR) studies have also shown that many intrinsically disordered protein domains adopt stable structures upon binding to their targets (Wright & Dyson, 1999). Theoretical models of protein folding indicate that proteins with greater thermal stability tend to have fewer major internal motions and less flexibility overall at constant temperature (Tang & Dill, 1998). These principles have been used to design proteins with high-affinity, pre-specified ligand binding, by focusing on the principles of "energetically favorable hydrogen-bonding and van der Waals interaction with the ligand..., high overall shape complementarity to the ligand, and ... structural pre-organization in the unbound protein state, which minimizes entropy loss upon ligand binding (Tinberg et al., 2013)."

However, experiments have revealed that designing ligands by maximizing the number of noncovalent interactions in the binding interface does not always improve the affinity between a protein and its binding partner (Velazquez-Campoy et al., 2000; Chodera & Mobley, 2013). Theory tells us that the net enthalpic gain of newly designed interactions may be overcome by the entropic cost of losing bond-rotational degrees of freedom due to the additional noncovalent constraints. Similarly, assuming the additivity and dominance of enthalpic contributions can be oversimplifications (Dill, 1997). However, neither of these considerations rules out the possibility of localized rigidification being a typical feature of the site of interaction between the protein and ligand, which may be accompanied by compensatory flexibility elsewhere in the molecules. In this work, we test whether such a measure of interfacial rigidity, involving protein atoms close to the ligand, contains sufficient information to predict their binding mode.

3.2.2 Computational probes of protein rigidity and flexibility

Two computational approaches for identifying rigid (stable) and flexible regions in proteins based on their intramolecular contacts or bond networks, rather than force field calculations by methods such as molecular dynamics (MD), have become widely used in recent years. The aim of these methods is to simplify the analysis of coupled motions and access larger-scale, biologically relevant conformational changes. The pioneering atomistic elastic network models for proteins evolved into faster, residue-based Gaussian network models (Bahar et al., 1997, 1998). These network models use normal mode analysis to identify the principal directions and amplitudes of motion at different frequencies within an oscillating spring system representing the protein, in which the spring force constants reflect the strength of noncovalent forces between atoms or residues. In contrast, ProFlex (initially named FIRST) evaluates protein flexibility by counting the bond-rotational degrees of freedom on a three-dimensional graph of the covalent and noncovalent bond network (Jacobs et al., 2001). This approach evolved from structural rigidity theory developed in the 1800s by James Clerk Maxwell for analyzing the distribution of flexible, rigid, and strained regions in bridges and other truss-work, based on the number and configuration of the struts (Maxwell, 1864). Instead of struts, bonds are used to represent the covalent and noncovalent interactions in proteins, including hydrophobic contacts, strong hydrogen bonds, and salt bridges. The 3D constraint counting search on the graph representing the protein covalent and noncovalent bond network results in a decomposition of the protein structure into spatial subsets: regions that are overconstrained by bonds and are rigid; cooperatively flexible regions that are formed by a coupled network of rotatable and nonrotatable bonds; and entirely flexible regions, such as side chains and main-chain termini that do not interact with other groups (Jacobs et al., 2001; Jacobs & Hendrickson, 1997). The temperature dependence of flexibility and the spatial hierarchy of flexible regions within a protein can also be evaluated with ProFlex (Hespenheide et al., 2002; Rader et al., 2002). The use of ProFlex by a number of research groups has shown its ability to reproduce main-chain crystallographic temperature factors and flexible regions identified by NMR for a number of proteins (Jacobs et al., 2001; Hespenheide et al., 2002; Zavodszky et al., 2004), as well as subtle

long-range changes in flexibility, including accurately predicting how flexibility redistributes upon ligand binding in the Ras/Raf and HIV protease complexes (Jacobs et al., 2001; Gohlke et al., 2004). Interestingly, despite taking less than a second of computing time per protein on a standard desktop computer, ProFlex results substantially agree with the flexible regions identified by elastic network models (Rader et al., 2002) and computationally more expensive MD simulations (Gohlke et al., 2004). For HIV protease (Figure 3.1), ProFlex reproduces NMR, crystallography, and MD results (Jacobs et al., 2001; Goodman et al., 2000; Korn & Rose, 1994; Gerstein & Krebs, 1998), indicating that the flaps above the binding pocket rigidify upon ligand binding and that chemical asymmetry within a ligand induces asymmetry in the flexibility of the monomers forming the active site.

3.2.3 Computational detection of protein-ligand interfacial rigidification

Given the experimental support for a protein-stabilizing effect of ligand binding in many cases, and the availability of ProFlex, a tool uniquely suited to define the rigid and flexible regions in a protein-ligand complex, we tested the hypothesis that native ligand binding results in rigidification of the protein-ligand interface through cooperative interactions. Interfacial rigidification has not previously been evaluated theoretically or computationally as a predictor of protein-ligand binding. In the majority of cases, the ProFlex-based SiteInterlock rigidity measure can predict the native complex given a series of sampled conformations and orientations of the ligand. SiteInterlock also provides new information to combine with existing protein-ligand scoring potentials, given that it is not highly correlated with scoring functions that have been trained to predict the interaction energy. Rather than being trained with a particular set of proteins to predict a response variable such as $\Delta G_{\text{binding}}$, SiteInterlock directly evaluates the change in rigidity of the interfacial bond network upon complex formation.

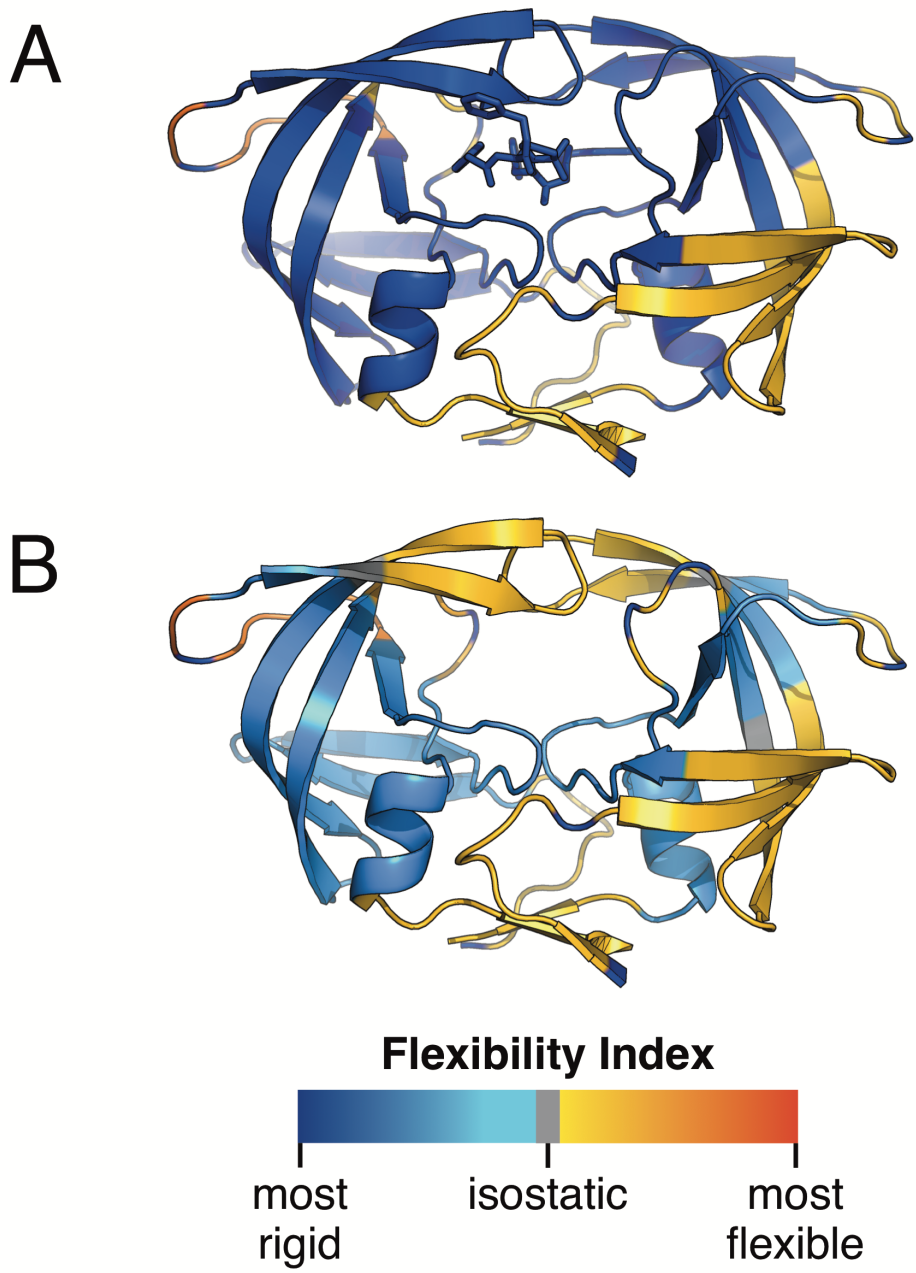


Figure 3.1: ProFlex assessment of the change in HIV protease flexibility upon inhibitor binding. A: X-ray crystal structure of HIV-1 protease (PDB entry 1htg; Jhoti et al., 1994) in complex with a penicillin-derived, asymmetric inhibitor. The protein structure is shown in cartoon representation, with the ligand in stick representation in the central binding pocket. The protein main-chain and the ligand heavy atoms are colored according to the flexibility indices measured by ProFlex. Note that the inhibitor has induced an asymmetry in flexibility between the two chains of HIV-1 protease, observed in the flexible beta strands to the right, while both halves of the dimer interface are similarly flexible (bottom center). B: The same PDB structure was analyzed with the ligand removed (while reflecting ligand-induced conformational changes in the protein), indicating that interactions with the ligand in (Å) are responsible for rigidifying the beta hairpin flaps (top center) over the ligand, while the flaps become flexible in the absence of the ligand (B).

3.3 Materials and Methods

The SiteInterlock analysis can be summarized in three main steps: (1) sampling all low-energy conformations of each ligand by using a tool such as OMEGA (version 2.3.2; OpenEye Scientific Software, Inc., Santa Fe, NM; http://www.eyesopen.com; Hawkins et al., 2010; Hawkins & Nicholls, 2012) if this is not already done by the ligand docking/orientational sampling tool, (2) sampling and saving a variety of sterically allowed orientations of ligand conformations in the protein site by using SLIDE (http://kuhnlab.bmb.msu.edu/software/slide/index.html; Zavodsky et al., 2002) or another docking tool without using the docking scoring function to filter the orientations, and (3) analyzing the structural rigidity of the protein-ligand binding interface for all docked ligand orientations with SiteInterlock, which employs ProFlex rigidity analysis (Jacobs et al., 2002).

3.3.1 Protein-ligand complexes analyzed

To test the efficacy of SiteInterlock in predicting native-like complexes, a set of 30 diverse protein complexes was prepared, including 25 enzymes and five receptors (Table 3.1 and Figure 3.2). All are determined at crystallographic resolution of 2.5 Å or better and are not listed as problematic structures in a quality analysis of protein-ligand fitting and refinement (Warren et al., 2012). Water molecules, hydrogen atoms, ligands, and nonprotein atoms were removed from the Protein Data Bank (PDB, Warren et al., 2012) files prior to docking; however, metal ions were retained if they were part of the ligand binding pocket. The 30 protein targets can be distinguished further as holo or apo structures. Eleven apo structures, in which a ligand-free structure of the protein was used for docking, were included to represent the additional challenge of not knowing the precise conformation of the protein bound to the ligand. For these 11 apo cases, the corresponding ligand-bound structures were available as separate PDB entries and used to provide an initial conformation of the ligand and also to validate the accuracy of the SiteInterlock-selected complex. For the 19 holo and 11 apo structures, the ligand of interest was extracted from the protein binding site and then conformationally sampled to reflect the realistic situation of not knowing the bioactive

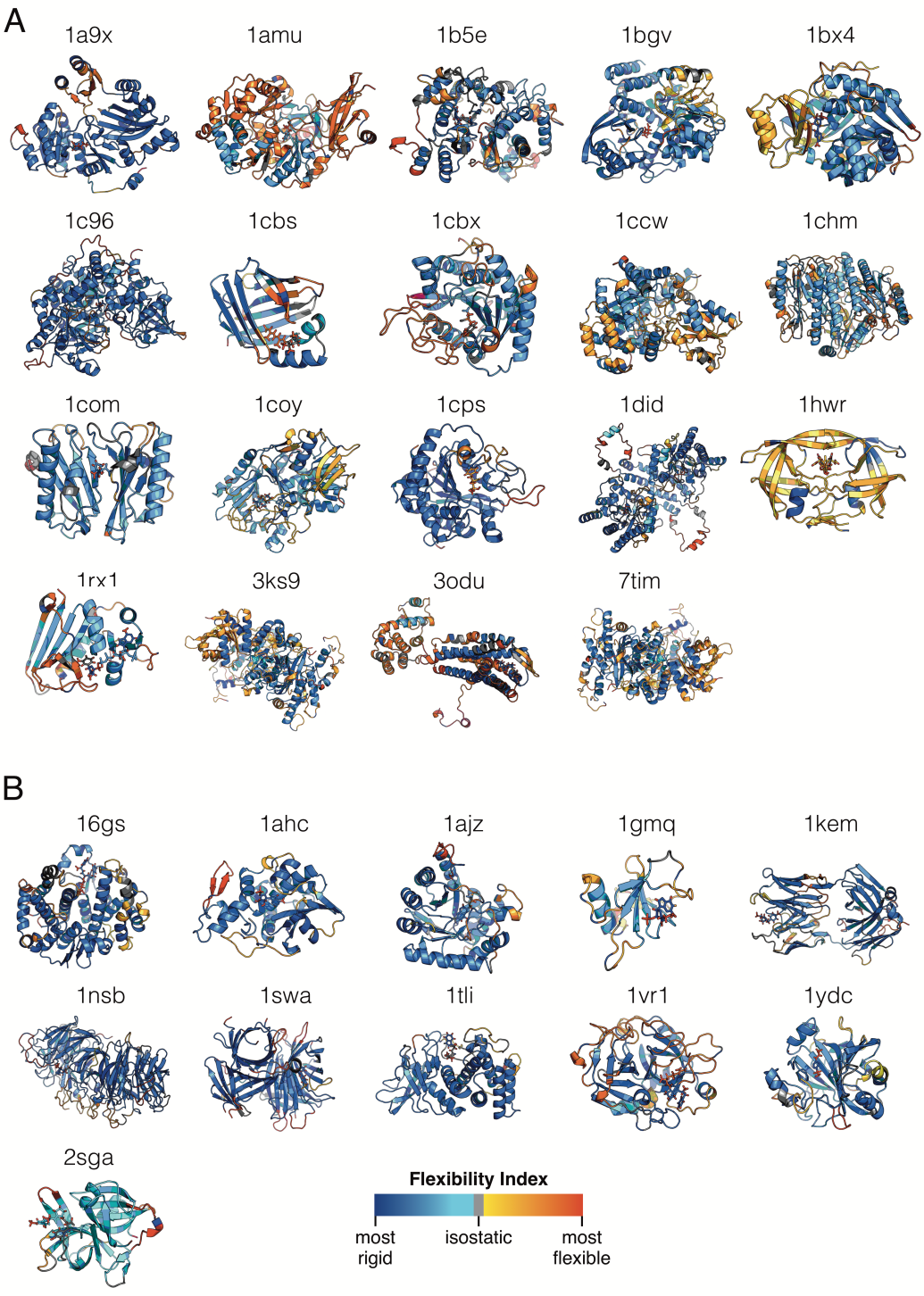


Figure 3.2: Flexible and rigid regions in 30 diverse protein crystal structures used to evaluate SiteInterlock and other scoring methods for their ability to detect the native ligand binding orientation. A: Crystal structures of the 19 complexes in the holo structure set. B: Crystal structures of the 11 apo protein structures. The protein structures (cartoon representation) and ligands (stick representation) are colored to reflect the degree of structural flexibility defined by ProFlex and SiteInterlock, as shown by the color spectrum below.

conformation of the ligand. The exact crystallographic ligand conformation was not included in docking for any of the 30 cases. This results in the "needle in the haystack problem" of having a large number of imperfect complexes (due to many orientations and many conformations of the ligand, plus protein conformational inaccuracies), challenging the scoring method to identify the most native-like.

Table 3.1: Protein-ligand complexes analyzed.

PDB entry (holo/apo)	Protein	Ligand	Resolution (Å)	Holo-apo binding site RMSD (Å)
1a9x / -	carbamoyl phosphate synthetase	L-ornithine	1.8	-
1amu / -	gramidicin synthetase 1	L-phenylalanine	1.9	-
1b5e / -	deoxycytidylate hydroxymethylase	deoxycytidylic acid	1.6	-
1bgv / -	glutamate dehydrogenase	L-glutamate	1.9	-
1bx4 / -	adenosine kinase	adenosine	1.5	-
1c96 / -	mitochondrial aconitase	citrate anion-iron/sulfur cluster	1.81	-
1cbs / -	retinoic acid binding protein	retinoic acid	1.8	-
1cbx / -	carboxypeptidase A	L-benzylsuccinic acid	2	-
1ccw / -	glutamate mutase	D-tartaric acid	1.6	-
1chm / -	creatine amidinohydrolase	carbamoyl sarcosine	1.9	-
1com / -	chorismate mutase	prephenic acid	2.2	-
1coy / -	cholesterol oxidase	dehydroepiandrosterone	1.8	-
1cps / -	carboxypeptidase A	sulfodiimine	2.25	-
1did / -	D-xylose isomerase	2,5-dideoxy-2,5-imino-D-glucitol	2.5	-
1hwr / -	HIV-1 protease	Xk216	1.8	-
1rx1 / -	dihydrofolate reductase	NADP+	2	-
3ks9 / -	metabotopic glutamate receptor	Z99	1.9	-
3odu / -	G-protein-coupled chemokine receptor	IT1t	2.5	-
7tim / -	triosephosphate isomerase	phosphoglycolohydroxamic	1.9	-
10gs / 16gs	glutathione S-transferase	L-cysteine amide	2.20 / 1.90	0.274
1ahb / 1ahc	alpha-momorcharin	formycin-5'-monophosphate	1.90 / 2.00	0.752
1aj2 / 1ajz	dihydropteroate synthase	pterin diphosphate	2.20 / 2.00	0.641
1gmr / 1gmq	ribonuclease	guanosine-2'-monophosphate	1.77 / 1.80	0.465
1kel / 1kem	sulfide oxidase antibody	methylphosphonic acid	1.90 / 2.20	0.676
1nsc / 1nsb	influenza B neuraminidase	O-sialic acid	1.70 / 2.20	0.323
1swd / 1swa	streptavidin	biotin	1.90 / 1.90	0.523
3tmn / 1tli	thermolysin	tryptophan	1.70 / 2.05	0.691
1tmt / 1vr1	alpha-thrombin	D-phenylalanine	2.20 / 1.90	0.655
1ydb / 1ydc	carbonic anhydrase II	acetazolamide	1.90 / 1.95	0.302
5sga / 2sga	proteinase A	acetyl group	1.80 / 1.50	0.192

3.3.2 Sampling complexes by molecular docking

After the ligands were extracted from their Protein Data Bank complexes (Table 3.1), hydrogen atoms and partial charges were assigned via partial semi-empirical AM1 geometry optimization with bond charge correction (Jakalian et al., 2002) by using molcharge (version 1.3.1) from the QUACPAC package (version 1.6.3.1; OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com). Up to 50,000 conformations were sampled for each ligand with OpenEye OMEGA

(version 2.3.2; Hawkins et al., 2010; Hawkins & Nicholls, 2012), and the most energetically favorable conformations (up to 200 conformers) were kept for docking. SLIDE, which docks ligands by exhaustive three-point pharmacophore matching between each conformer and the binding site and performs minimal protein side-chain and ligand single-bond rotations to allow van der Waals collision-free docking, was then used to sample a range of dockings for each complex. SLIDE version 3.4 was modified to output all sterically allowed orientations of each ligand, given the OMEGA conformers as input. To assess the goodness of a docking, the root-mean-square deviation (RMSD) between nonhydrogen atom positions was calculated between each docking and the crystallographic ligand pose. Starting with this large set of ligand dockings labeled by RMSD, a series of dockings was selected to span the RMSD range between 0 and 3 Å (relative to the crystallographic position), representing a range of sterically feasible, near-native but otherwise unscored dockings. For each complex, this series included the best-sampled docking (closest to 0 Å RMSD), the docking closest to 3 Å RMSD, and an average of 8 additional dockings distributed semi-uniformly in the 0-3 Å RMSD range. Ligand dockings in the range of 3-6 Å RMSD were also sampled, and several dockings with different RMSD values in that range were also kept for each complex as examples of poor dockings. For seven of the complexes, dockings in the 6-10 Å RMSD range were also observed and included. Ideally, evenly separated dockings would be selected over a specified RMSD interval for all targets (for example, ligand dockings with 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2 Å RMSD, and so forth, relative to the crystallographic position). However, the RMSD space of possible dockings is remarkably restricted by the size, geometry, and flexibility of the particular ligand as well as by the binding site geometry. This is found even with thorough ligand conformational sampling prior to docking. For each complex, the crystallographic ligand conformer was not included in pose prediction, because the bioactive conformation is not known a priori in a real world application. For all 30 complexes, the set of docking poses (reflecting both conformational and orientational sampling) and corresponding protein conformations (which may include SLIDE-rotated side chains) were presented to SiteInterlock and the other five scoring functions. All resulting protein and ligand structural figures were rendered by PyMOL (version

1.5.0.4; Schroedinger, LLC; http://pymol.org; DeLano, 2002).

3.3.3 Evaluating correlation between scoring functions

To assess the degree of monotonicity between two scoring functions (the extent to which they rank dockings in the same order), Spearman's rank correlation coefficient, ρ, was calculated as

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

where d_i is the difference in the ranks of two poses x_i and y_i for scoring functions x and y, and n is the number of docking poses. Spearman's ρ takes values in the range between -1 and 1, where a perfect monotonic relationship in ranks between two scoring functions exists when $\rho = 1$, and a perfect inverse relationship exists when $\rho = -1$. A complete absence of correlation in ranking is indicated by $\rho = 0$.

3.3.4 Rigidity analysis

To prepare the series of dockings for rigidity analysis by ProFlex version 5.2 (http://www.kuhnlab.bmb.msu.edu/software/proflex/index.html), hydrogen atoms were added to the protein structures via Reduce (Word et al., 1999), and the coordinates of the ligand poses were converted to PDB format. The ligand atom hydrogen-bond donor and acceptor assignment was automated for each docking analyzed by ProFlex, based on the intermolecular interactions identified by SLIDE for that docking. This is more accurate than assigning hydrogen-bonding roles prior to docking. For instance, a hydroxyl group could potentially act as a hydrogen-bond donor and/or an acceptor. SLIDE determines whether one or both occur, based on evaluation of interaction distances and angles between the protein and ligand for a given ligand orientation (Zavodsky et al., 2002). The steps in SiteInterlock (Figure 3.3) were designed to test whether a ligand docking close to the known crystallographic orientation and conformation can be detected based on exhibiting greater protein-ligand interface rigidity than is found for incorrect dockings. The first step in the procedure is to select an energy for ProFlex rigidity analysis of the protein structure, determining which

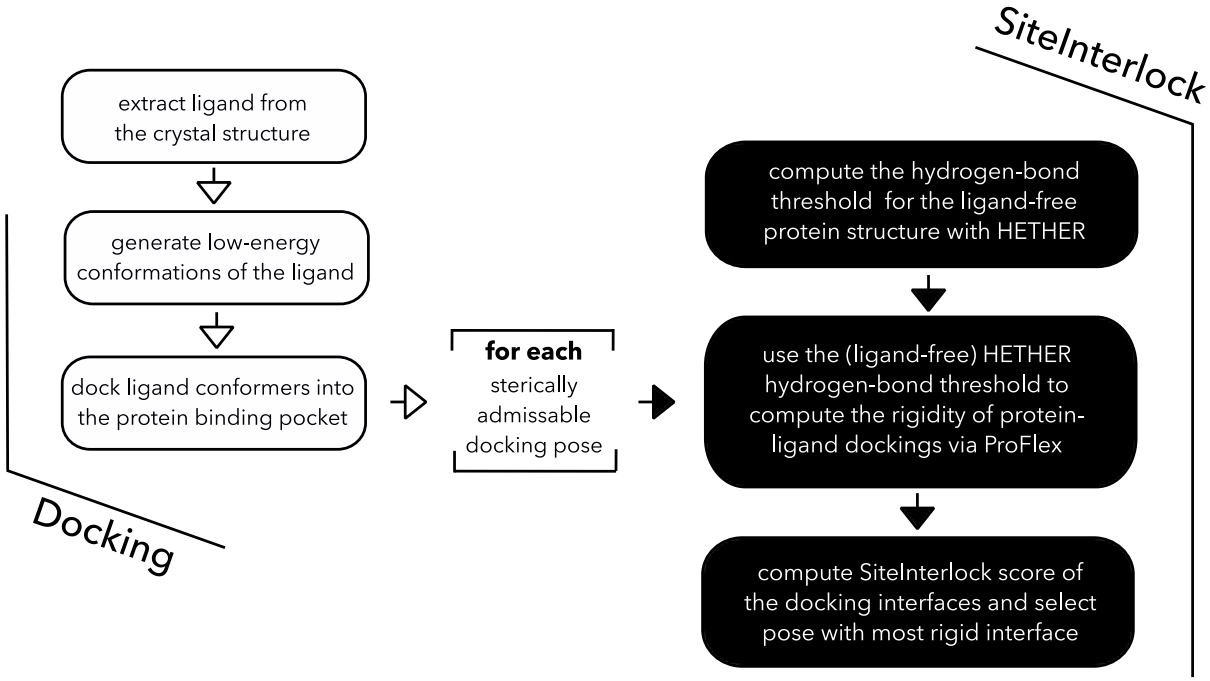


Figure 3.3: Flowchart of the preparation of input structures for SiteInterlock. Illustrated are the ligand preparation steps for SiteInterlock analysis (left) and the SiteInterlock analysis itself (right): HETHER selection of the ProFlex hydrogen-bond energy threshold for the protein in absence of the ligand, ProFlex analysis of protein- ligand interfacial flexibility/rigidity for each docking, and selection of the docking pose with the greatest interfacial rigidity.

hydrogen bonds and salt bridges will be included in the bond network based on their energy values, which are measured as a function of atom type, distance, and angle. This selection of a suitable hydrogen-bond/saltbridge energy threshold adjusts for the fact that protein structures in the PDB are solved at different temperatures and pressures, in different solvents, and with different fitting and refinement software, all of which affect the prevalence of noncovalent interactions that meet a given set of distance and angle criteria. The native state of most proteins is poised near the rigid to flexible transition energy (Rader et al., 2002), where the main-chain remains structurally stable (mostly rigid) while also exhibiting some flexible regions, which are often relevant to ligand binding (Zavodszky et al., 2004; Jacobs et al., 2002).

The HETHER (Hydrogen-bond Energy ThresHold Estimator for Rigidity analysis) software module developed here (included in the SiteInterlock distribution; https://github.com/psa-lab/

siteinterlock) is designed to identify that native-like energy threshold. HETHER reads the results of the hydrogen-bond dilution function in ProFlex that mimics the thermal denaturation of a protein (Hespenheide et al., 2002). HETHER analyzes changes in the regions of the protein main-chain that remain either independently rigid (able to move as rigid bodies relative to each other), mutually rigid, or flexible, as the ProFlex hydrogen-bond energy (temperature) increases. As the energy increases, noncovalent interactions break, and regions that were rigid become flexible or less coupled to each other. ProFlex reports every energy value at which the size or number of rigid regions in the protein main chain has changed, as well as the noncovalent interactions included in that bond network. From the series of energy values at which main-chain rigidity differences were observed, HETHER selects the lower energy value (the more rigid state) between the two adjacent energy values (structural states) at which the number of independent rigid regions changed the most. This is called the energy threshold (or cutoff) for HETHER and SiteInterlock analysis. This energy threshold detects the point at which the protein is rapidly changing from a rigid to a flexible state (Rader et al., 2002), when the protein is also sensitive to changes in the interfacial bond network upon ligand interaction. For instance, if there are two independent rigid regions at one energy value, and four at the next higher energy (due to rigid regions breaking apart upon the loss of noncovalent interactions), then the increase in the number of rigid regions is two. If this is the greatest change in the number of rigid regions between any two consecutive energy values, then the bond network of the system with two independent rigid regions will be chosen by HETHER for SiteInterlock analysis of the protein-ligand complex. HETHER only considers the range of energy values at which the main-chain is between 25% and 90% rigid (leaving out totally rigid or mostly flexible states), and HETHER defines rigid regions as those containing at least three alpha carbons to avoid including trivial rigid regions such as dipeptides containing proline as the second residue. The rigid-to-flexible transition energy threshold is identified by HETHER for the apo or de-ligated holo version of each protein complex, and then the same energy threshold is used to analyze each docked ligand complex of that protein. An example of a hydrogen-bond dilution plot and illustration of the energy threshold chosen by HETHER for SiteInterlock analysis is shown in

Figure 3.4.

To quantify the degree of structural rigidity in a protein-ligand complex, we used the continuous flexibility index f_i, which ProFlex computes for each atom i. For atoms in rigid regions, the flexibility index quantifies the degree of rigidity of each atom based on the larger number of constraints in that region relative to the number needed for the region to be just barely rigid; this total-number-of-constraints value for the region is divided by the total number of bonds in that region to define the flexibility index for each atom in the region. The same calculation is done for atoms in flexible regions, which show fewer constraints than are needed for the region to be rigid. Following the rigid region decomposition by ProFlex, each atom i is also assigned a rescaled flexibility score f_i' in the range from 0 to 100, where a value of 50 indicates that the atom belongs to an isostatically (just barely) rigid region, and atoms with a flexibility index below 50 or above 50 are part of a rigid region or a flexible region, respectively. This rescaling is done for the convenience of writing flexibility data in the crystallographic temperature factor column of PDB files, typically for 3D visualization with a color spectrum.

It should be noted that ProFlex is sensitive to the stereochemical quality of the protein structure being analyzed, particularly the main-chain bond lengths and angles, because they are critical for defining the rigidity of the protein structure as a whole. Thus, we recommend using structural validation tools such as PROCHECK (Laskowski et al., 1993), MolProbity (Davis et al., 2007), and SWISS-MODEL (Bordoli et al., 2008). Structural assessment to evaluate the stereochemical quality of any protein structure before using it as the basis for ProFlex or SiteInterlock analysis. An example of a structure which is borderline in suitability for ProFlex analysis is a second PDB entry for HIV-1 protease bound to a different inhibitor (relative to that shown in Figure 3.1). At the end of the third line of the ProFlex results on holo structures in Figure 3.2, this second HIV-1 protease structure is assessed as mostly flexible at the ProFlex energy threshold selected by HETHER for use in SiteInterlock. To understand the basis for this unexpected flexibility relative to the other 29 proteins analyzed, PROCHECK was run. It showed this PDB entry to have a main-chain (ϕ, ψ) angle value distribution that is "unusual" for structures solved at this (1.8 Å) resolution, and its

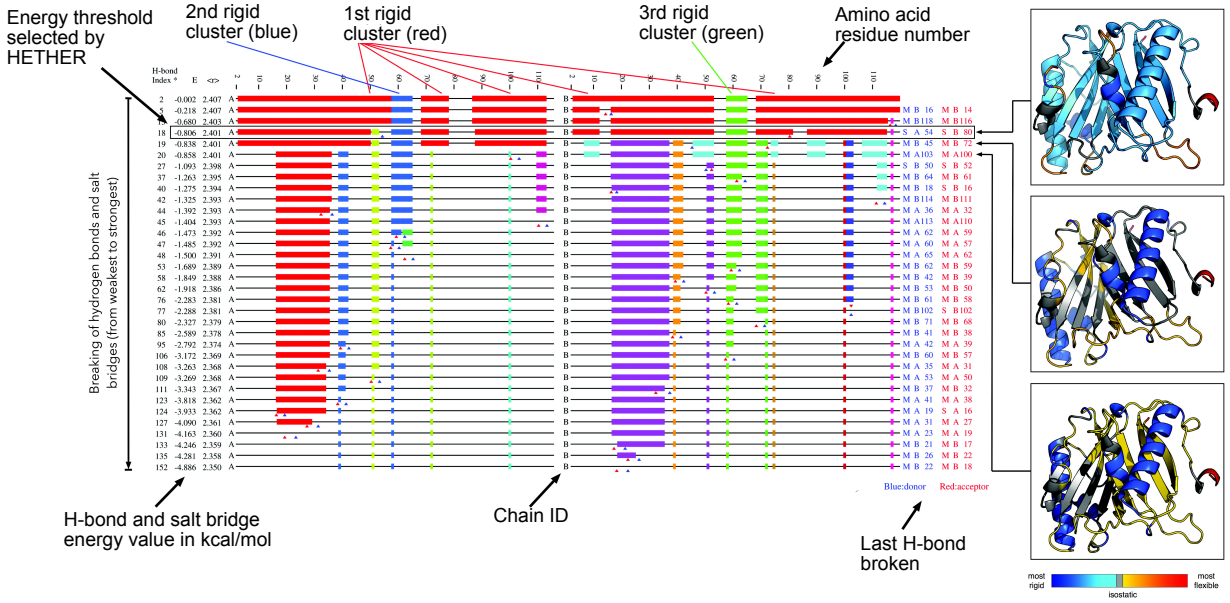


Figure 3.4: ProFlex hydrogen-bond dilution plot of the de-ligated protein structure of a monofunctional chorismate mutase. The analysis of monofunctional chorismate mutase from Bacillus subtilis (PDB code: 1com; Chook et al., 1994), is showing the transition from mostly rigid to mostly flexible as hydrogen bonds and salt bridges are broken with increasing energy. The distinct lines in this plot show the rigid and flexible regions of the protein at different energy values, with successive lines representing increasingly flexible states of the protein as the energy level (temperature) increases. Residues of the protein chain are numbered from left to right at the top of the plot. At a given energy value, the thick, colored blocks in each row indicate the rigid clusters of the protein main-chain, with a different color used for each independently rigid cluster of atoms. The thin, black lines correspond to intervening flexible regions observed in the protein bond network at that energy. A rigid region may be comprised of residues that are not contiguous in sequence; thus, blocks of residues with the same color indicate residues belonging to the same mutually-rigid region. The energy value for each row is listed in the second column from the left. The first row shows the predicted state of the protein when all hydrogen bonds and salt bridges are included in the bond network. The third column shows the average number of bonds to each atom (averaged over all atoms in the protein) at that energy level, including covalent single and double bonds, bond-coordination constraints (constraining sp^3 and sp^2 centers in the correct geometry), hydrophobic tethers, hydrogen bonds, and salt bridges. For instance, the second row, at an energy value of -0.218 kcal/ mol, shows the rigid and flexible regions in the protein when all hydrogen bonds and salt bridges with an energy of -0.218 kcal/mol or stronger are included in the bond network. Moving down the rows of the plot, the energy values increase and hydrogen bonds and salt bridges are incrementally broken (from weakest to strongest), resulting in an overall increase of flexible regions in the protein structures indicated by the intervening, black lines and fragmentation of rigid regions.

Figure 3.4 (cont'd). The energy value selected by HETHER is highlighted by the black frame shown at -0.806 kcal/mol, in which the main-chain is mostly rigid (comprised by the large rigid region shown in red, plus two approximately 10-residue independent rigid regions colored in blue and green, and a very short rigid region in lime green appearing at residue 50). This state shows some residual flexibility that is sensitive to native-like ligand interactions, as described in the results in the main text. The rigid and flexible regions mapped onto the corresponding, ligand-free protein structure at different energy levels are shown at the far right, now colored by flexibility index (with colors defined in the spectrum bar shown beneath the structures). At the next energy step (-0.838 kcal/mol) above that chosen by HETHER, the protein structure decomposes into eight rigid clusters (red, yellow, blue, green, cyan, orange, lime green, and dark blue), which results in a structure with about one-third of the main-chain being flexible. Thus, HETHER selected the last substantially stable state of the protein structure, as intended.

main-chain bond angle and Ω (peptide bond planarity) angle distributions are "highly unusual." ProFlex is appropriately sensitive to main-chain stereochemistry, because the main-chain hydrogen bond network is essential for maintaining overall structural integrity. While the SiteInterlock ligand orientation results are reasonable for this protein, as detailed below, in general we would recommend considering an alternative PDB structure with better stereochemistry.

3.3.5 SiteInterlock interfacial rigidity score

The protein rigidity metric P ("ProteinAvg") was computed as the average flexibility index f' of all protein atoms n (including hydrogens) within 9 Å of one or more heavy atoms in the docked ligand,

$$P = \frac{1}{n} \sum_{i=1}^{n} f_i'.$$

Here, f_i' is the flexibility index of the ith protein atom in the protein binding site. Similarly the ligand rigidity metric L ("LigandAvg") was calculated as the average flexibility index of all m ligand atoms in the current docking,

$$L = \frac{1}{m} \sum_{j=1}^{m} f_j'.$$

As for protein interfacial atoms, the ligand atoms' flexibility index values are influenced by the changes in noncovalent interactions as well as ligand and protein conformational differences between

the different dockings. The final SiteInterlock score was calculated as the average of protein scores (P) and ligand scores (L),

$$\text{Siteinterlock score} = \frac{1}{2}\Big(P' + L'\Big),$$

where P' and L' are rescaled versions of P and L, respectively. To rescale P and L to fall on the same scale for computing the SiteInterlock score, Z-score standardization was used based on the mean score (μ) and standard deviation (σ) of P and L values across the docking poses of a target,

$$x' = \frac{x - \mu}{\sigma},$$

where x presents a single score to be rescaled. Thus, the SiteInterlock score is an equal weighting of interfacial protein atoms's average rigidity (or flexibility) and interfacial docked ligand atoms's average rigidity (or flexibility), in units of standard deviations above or below the mean value for that set of dockings. This measure of rigidity considers any reorganization of protein and ligand groups upon docking, reflecting the cooperativity of the bond network in the interface. The workflow of the SiteInterlock software, including preparatory steps that may be done with user-preferred tools, and the roles of HETHER and ProFlex, is outlined in Figure 3.3. The HETHER, ProFlex, and SiteInterlock software modules are available to academic researchers at https://github.com/psa-lab/siteinterlock under GNU General Public License and to commercial entities by making licensing arrangements.

3.3.6 Other scoring functions

Scoring functions for comparison with SiteInterlock were used with their respective default settings, unless noted otherwise. Values for the docking scoring function X-Score were computed by using X-Score version 1.3, which outputs binding affinities in pKD units of the different ligand poses as the average of the X-Score scoring functions HPScore, HMScore, and HSScore (Wang et al., 2002). DrugScore (DSX) version 0.88 was used (Neudert & Klebe, 2011). LigScore was executed from the IMP package (version 2.2; Russel et al., 2012), using the PoseScore module for ranking ligand

orientations (Fan et al., 2011). Protein PDB and ligand MOL2 files were prepared for DOCK6 Amber Score (DOCK6 version 6.3; Allen et al., 2015b) via their `prepare_amber.pl` script, using the recommended parameter set in http://dock.compbio.ucsf.edu/DOCK_6/tutorials/amber_score/dock.in. For scoring protein-ligand complexes via AutoDock Vina (version 1.1.2), protein and ligand files were prepared by using the `prepare_ligand4.py` and `prepare_receptor4.py` in the AutoDockTools utilities from the MGLTools package (version1.5.6; Morris et al., 2009).

3.4 Results and Discussion

3.4.1 Detecting structural rigidification upon protein-ligand complex formation

To assess whether the native ligand orientation results in a discernible rigidification of the protein-ligand interface, 30 different protein-ligand complexes were analyzed with SiteInterlock (Table 3.1; Figure 3.2). Nineteen of the cases were holo protein structures solved in complex with a ligand (Figure 3.2A). The native ligand was deleted from the crystal structure, HETHER energy-based selection of hydrogen bonds was performed on the de-ligated structure, and rigid region decomposition was performed by ProFlex on each of the docked complexes at the same energy threshold.

First, our analysis focused on whether the native (crystallographic) complex exhibited greater rigidity in the protein-ligand interface with the ligand present versus absent. This tested whether there is a consistent trend toward rigidification upon complex formation for the ideal case with no significant conformational or orientational inaccuracies in the ligand or protein structure. To quantify the rigidity of a structure, the SiteInterlock score was computed as the equally weighted sum of the averaged flexibility indices of ligand atoms and interfacial protein atoms (those within 9 Å of non-hydrogen atoms in the ligand). In the majority (17 out of 19) of the holo complexes, interfacial protein atoms were found to become more rigid in the presence of the ligand presented in the crystallographic binding mode (Figure 3.4), due to cooperativity of the noncovalent bond network between the molecules. This is consistent with a previous analysis of protein-ligand complexes showing that 71% of protein atoms within 8 Å of ligand atoms in the holo structures

have decreased mobility (lower crystallographic temperature factors) relative to their apo states (Yang et al., 2005). This phenomenon is illustrated in Figure 3.1 for HIV protease (Jacobs et al., 2001; Goodman et al., 2000; Korn & Rose, 1994; Gerstein & Krebs, 1998). In two cases, the protein interface in the complex was equally rigid with and without the ligand (Figure 3.4). In one of these

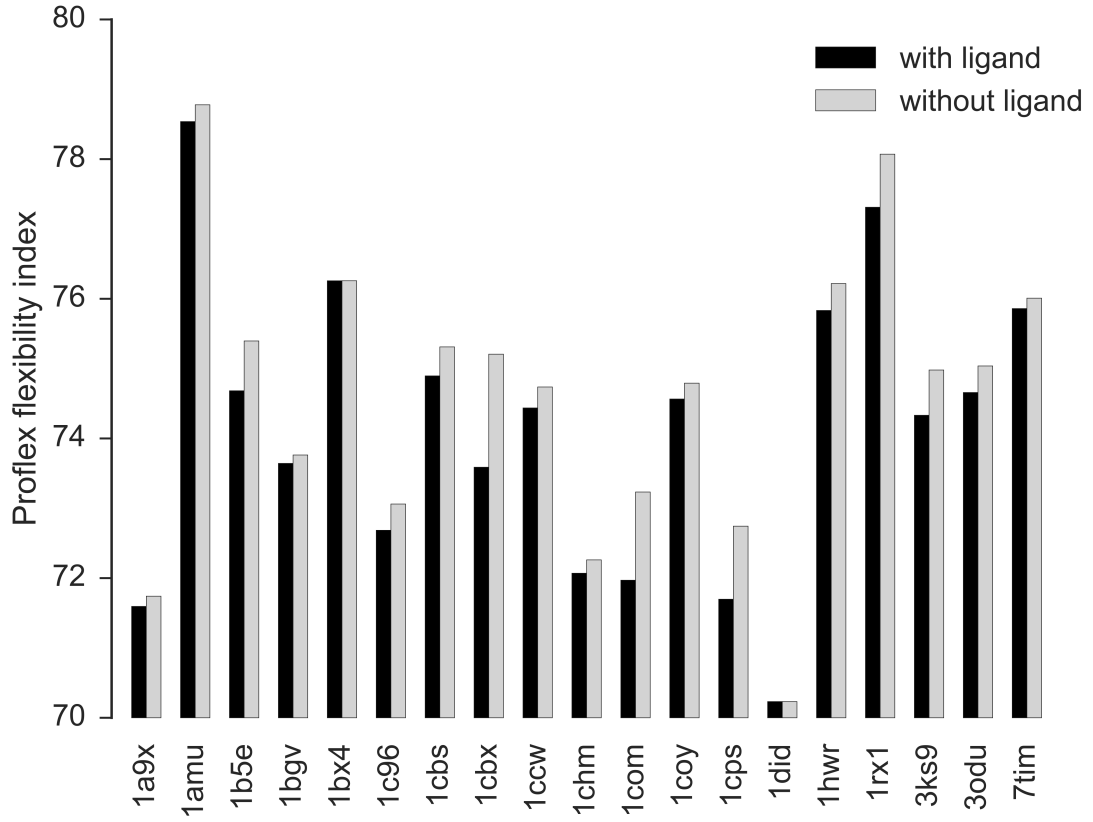


Figure 3.4: Rigidity of interfacial protein atoms. The rigidity of atoms in the protein binding site (within 9 Å of ligand heavy atoms) is shown in the presence (black bars) and absence (gray bars) of the crystallographic ligand pose for the 19 holo structures. Lower ProFlex values indicate greater rigidity. For 17 cases, the protein interface is more rigid in the presence of the ligand, and for 2 cases (PDB entries 1bx4 [Mathews et al., 1998] and 1did [Collyer & Blow, 1990]), it is equally rigid.

cases, adenosine kinase (PDB entry 1bx4; Mathews et al., 1998), p:p or p:cation interactions with the adenosine ring system in the ligand were not assigned as strong noncovalent interactions by ProFlex, suggesting an area for improvement. The possibility of an equally rigid protein site in the presence and absence of ligand also suggested that the role of ligand rigidification in complex formation be considered. The SiteInterlock score, which includes the LigandAvg component as

well as ProteinAvg, was therefore used for analyzing docked complexes. This combination scoring also has the practical advantage of breaking ties in rigidity values between different protein-ligand dockings that could be observed when using ProteinAvg or LigandAvg alone.

An example of SiteInterlock rigidity analysis of the crystallographic binding mode versus an inaccurately docked pose is shown for chorismate mutase (Figure 3.5). The protein backbone and ligand are colored by rigidity, and it is evident that both the protein and ligand are more rigid in the near-native (0.36 Å ligand RMSD) complex (Figure 3.5A) than in the 3.56 Å RMSD ligand docking (Figure 3.5B). Reorganization of protein side chains and ligand flexible groups to accommodate the mispositioned ligand yielded decreased rigidity of the protein binding site and flanking beta sheet, while the ligand remained flexible due to few stabilizing interactions. Across all 30 complexes, it was observed that a net decrease in flexibility of the combination of protein and ligand atoms at the interface (the SiteInterlock score) is a signature of native or near-native complexes, rather than both the protein and ligand individually becoming more rigid.

The SiteInterlock approach was then tested for the ability to discriminate and predict the native binding pose from a series of docked poses with increasing RMSD relative to the crystallographic position. Favorable ligand conformations from OMEGA were used as the input to sample a variety of binding poses with SLIDE for the 19 holo protein structures. Only sterically permissible dockings were retained, with no filtering of poses based on docking scores. To reflect the real-world case of protein complex prediction in which the ligand conformation and orientation and the conformations of interfacial protein side chains upon binding are all unknown, apo crystal structures for 11 proteins were also used as the basis for docking. The corresponding holo structures (Table 3.1) were used to provide the ligand structure as input to conformational sampling for docking and to assess the accuracy of the apo structure dockings selected by SiteInterlock and the other scoring methods.

For chorismate mutase, the range of sampled poses and corresponding SiteInterlock scores appears in Figure 3.6A, showing a funnel-like profile in which the protein-ligand interface becomes increasingly rigid as the ligand RMSD approaches 0 (the crystallographic pose). The prephenic acid ligand pose with the most rigid SiteInterlock score falls within 0.4 Å of the crystallographically

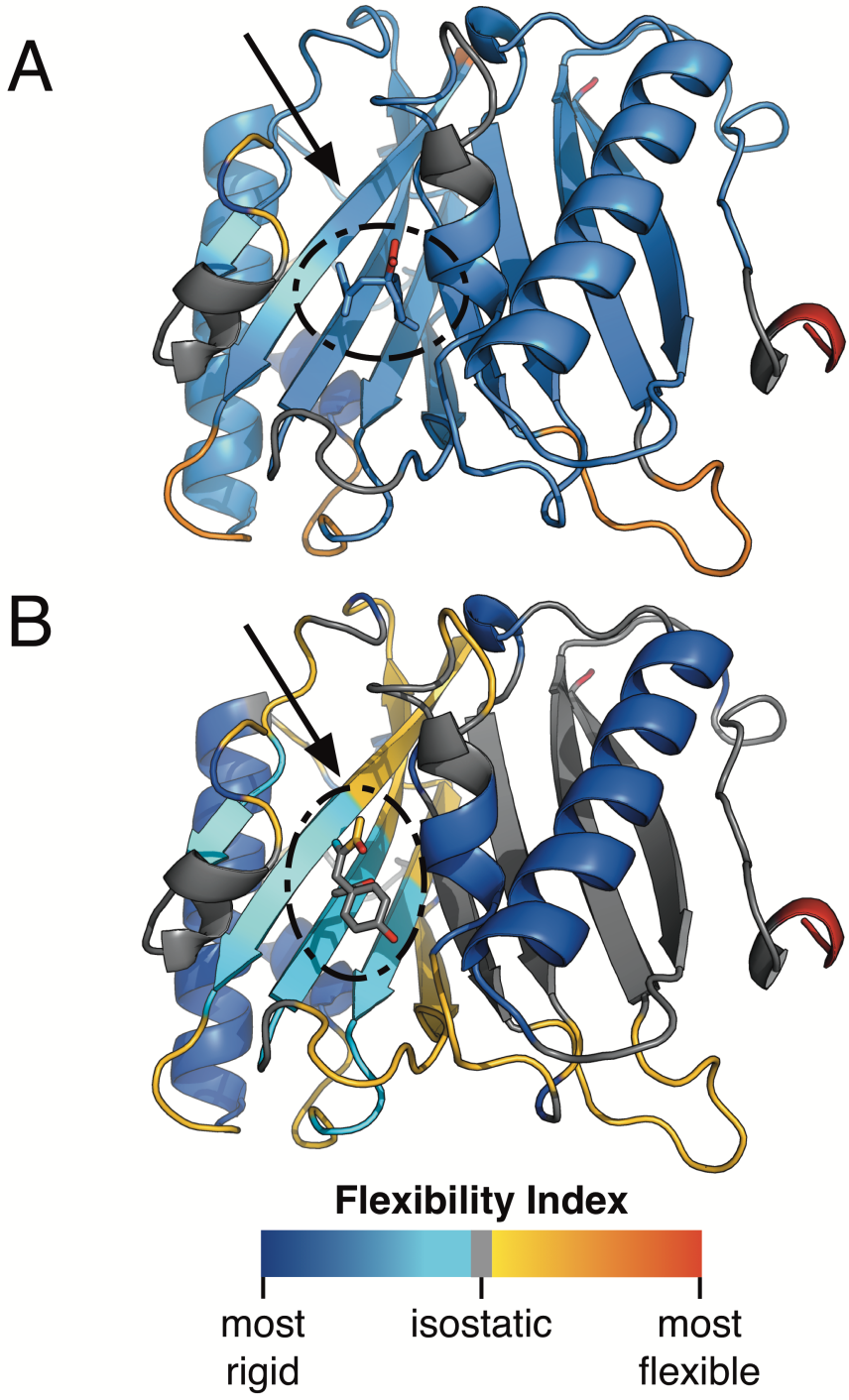


Figure 3.5: Comparing changes in structural flexibility between crystal structures and dockings. Shown are the changes in structural flexibility of the complex of monofunctional chorismate mutase with its enzymatic product, prephenic acid, of native-like (PDB entry 1com; Chook et al., 1994) versus non-native dockings of the ligand. Arrows point to prephenic acid in the binding site. A: Near-native docking pose (ligand RMSD 0.36 Å). B: Inaccurate docking pose (ligand RMSD 3.56 Å). Note the enhanced rigidity of both the binding site and the ligand in the native pose relative to the misdocked pose.

observed position. The ability of SiteInterlock score to rank the docking poses from lowest to highest RMSD was then tested for all the complexes. A positive correlation was found between decreasing RMSD and greater rigidity (more negative SiteInterlock score) for 25 out of the 30 cases, which is also apparent when all the dockings are pooled (Figure 3.6B). The Spearman rank correlation coefficient (median value of 0.55 across the 30 complexes) between the SiteInterlock score and the docked ligand RMSD indicates that SiteInterlock is well behaved in discriminating among poses across a broad RMSD range.

For predicting the native protein-ligand complex, when the ligand pose with the most rigid SiteInterlock value is identified for each of the 30 complexes (Figure 3.7), it is found to be within 0.5 Å RMSD of the best-sampled pose for 14 of the complexes and within 1.5 Å RMSD for 11 others. A poor docking was identified only for the glutamate dehydrogenase complex (3.9 Å ligand RMSD; PDB entry 1bgv; Stillman et al., 1993). SiteInterlock inclusion of both protein and ligand interfacial rigidity for identifying native-like dockings clearly outperforms using the protein interfacial rigidity value alone (ProteinAvg), especially in avoiding low-accuracy dockings (Figure 3.7).

SiteInterlock was then compared with five commonly used methods for evaluating ligand binding to proteins – PoseScore, AutoDock Vina, DSX, DOCK6 Amber Score, and X-Score – which reflect a spectrum of commonly used knowledge-based, empirical and force field scoring functions. SiteInterlock performs competitively with the better of these methods (Figure 3.8), performing particularly well in predicting most protein-ligand complexes to within 1-2.5 Å ligand RMSD. SiteInterlock also avoided selecting suboptimal dockings for all but one of the 30 complexes (PDB entry 1bgv, 3.9 Å RMSD; Stillman et al., 1993). SiteInterlock also shows strength in avoiding inaccurate (high RMSD) ligand orientations when docking into an apo structure, where the protein is not pre-conformed to bind that ligand (Figure 3.8B). Four of the other scoring functions selected poor-accuracy (5.4-9.3 Å RMSD, Table 3.2). poses for between one and three of the apo cases, possibly because they were parameterized to favor interaction geometries found in holo structures. However, all scoring functions performed well on the holo structure set (Table 3.2). These results

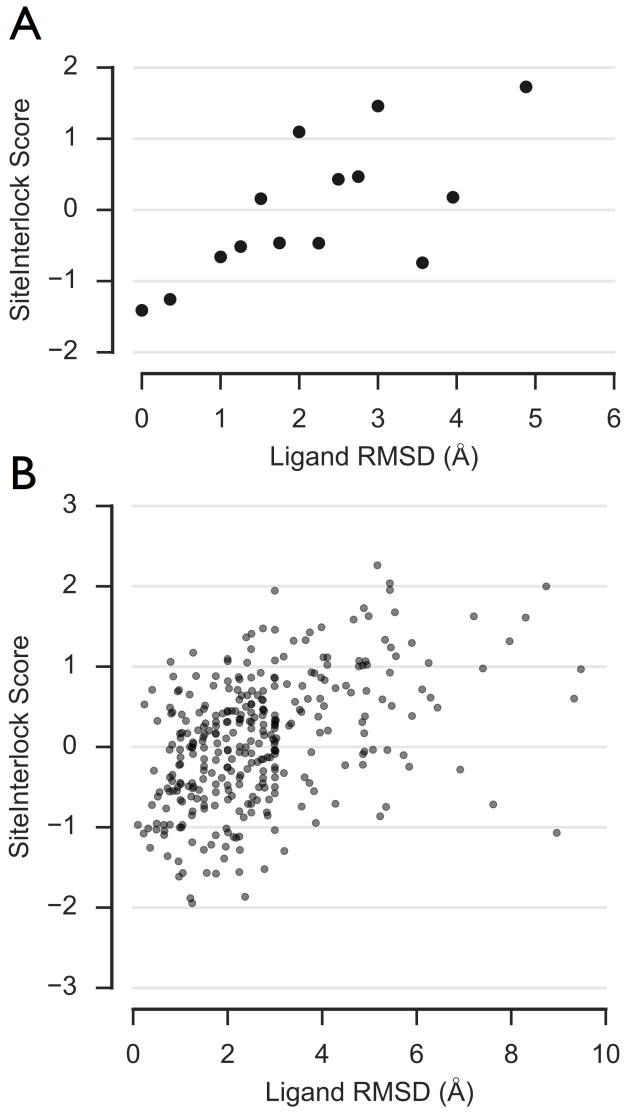


Figure 3.6: Relationship between the SiteInterlock score and ligand RMSD relative to the crystallographic pose. A: dockings spanning the RMSD range of 0-5 Å for prephenic acid in complex with chorismate mutase (PDB entry 1com; also see Figure 3.5) for SiteInterlock results on two of these poses). B: 331 dockings from all 30 protein-ligand complexes. A funnel-like tendency is seen that discriminates more native-like dockings (closer to 0 Å RMSD) based on these dockings having more negative (rigid) SiteInterlock scores, particularly for dockings with RMSD values of $\leq$ 3 Å.

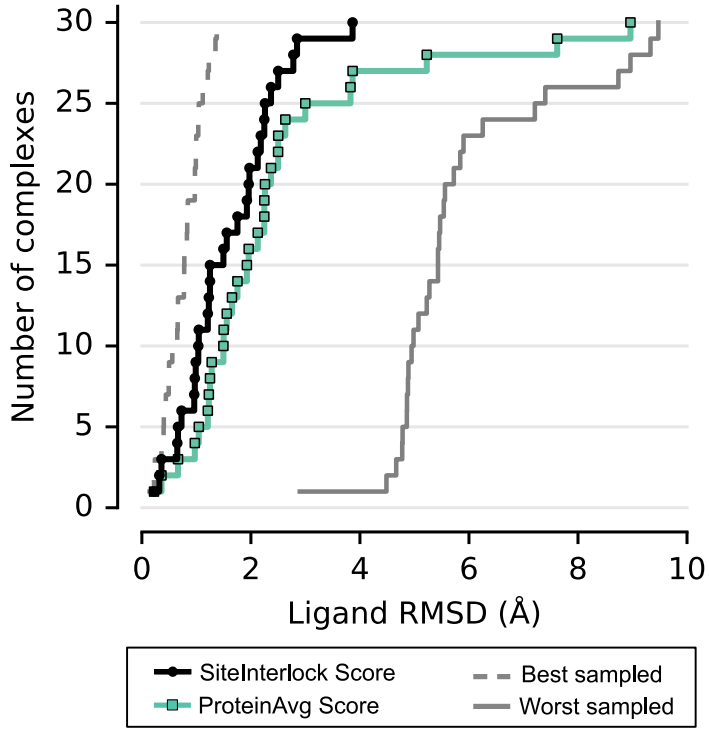


Figure 3.7: Enrichment plot comparing the SiteInterlock score with the ProteinAvg score for selecting near-native docking poses for the 30 targets. Here, the y axis value shows the number of complexes for which the best-scoring pose selected by SiteInterlock (black curve) and ProteinAvg (green curve) is within the ligand RMSD value shown on the x axis. For example, we see that the best-scoring ligand pose selected by SiteInterlock is under 3 Å RMSD in 29 of the 30 cases. The combination of protein and ligand interfacial rigidity in the SiteInterlock score is apparently a better predictor of native-like poses than protein rigidity alone (ProteinAvg). The gray dashed line indicates the best scoring performance possible, if the best-sampled pose were selected for each complex, and the solid dashed line indicates the worst possible performance, based on selecting the worst-RMSD pose for each complex.

suggest not only that SiteInterlock performs robustly on its own in selecting near-native dockings across a wide range of protein and ligand types, but also that it has unique strengths in ferreting out decoy poses.

3.4.2 Interfacial rigidity as a signature of native protein-ligand interaction

To assess the relationship between SiteInterlock and other scoring function rankings of the same ligand poses, scatter plots were made to compare all pairs of scoring function values (SiteInterlock, PoseScore, AutoDock Vina, DSX, DOCK6 Amber Score, and X-Score) for the same 331 dockings

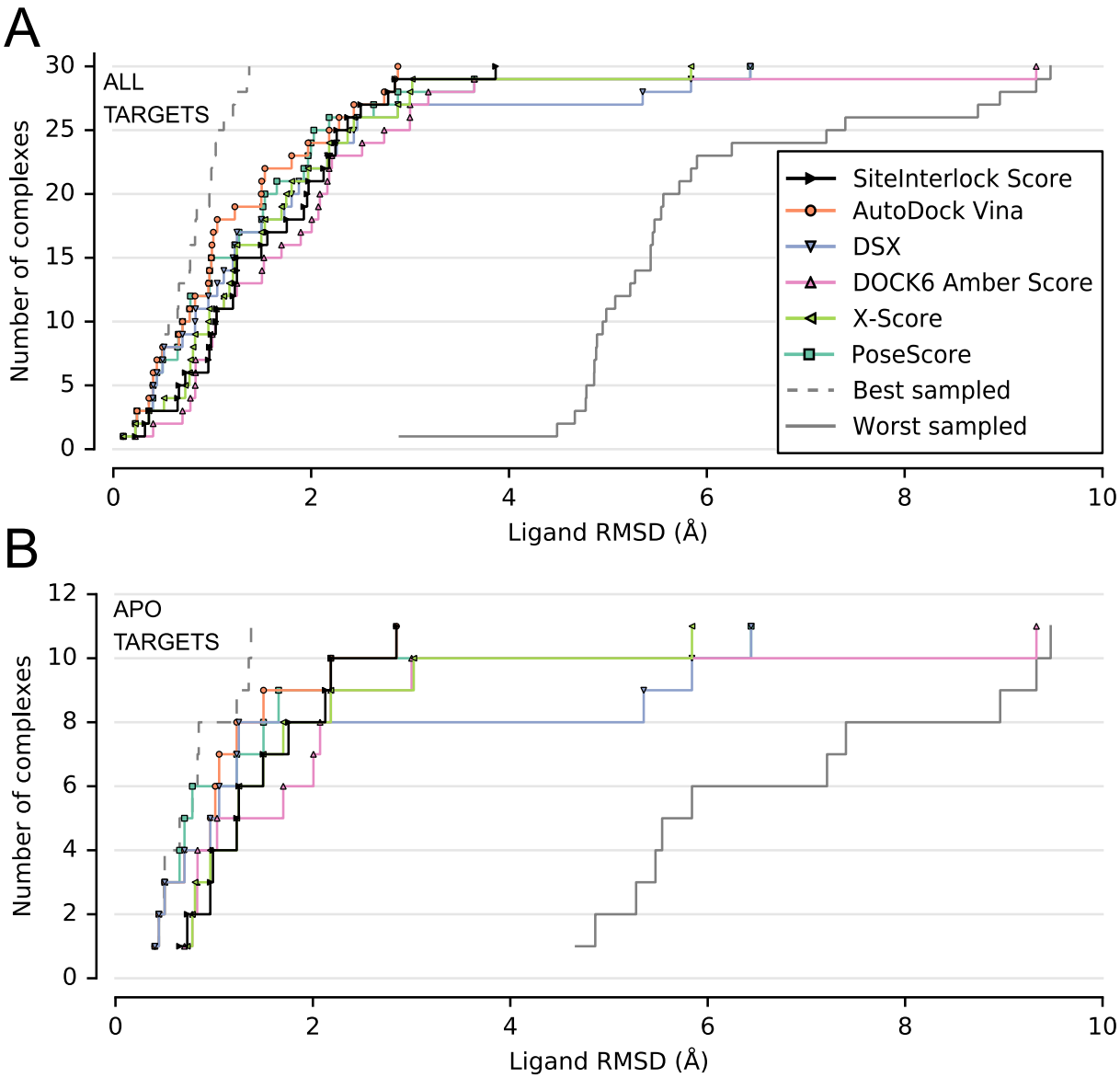


Figure 3.8: Enrichment plot comparing the accuracy of pose selection of SiteInterlock with five different docking scoring functions. The accuracy of pose selection of SiteInterlock is shown as black line with square symbols, which is compared with five different docking scoring functions (see color legend on plot), bounded by the curves showing the best-sampled (dashed gray line) and worst sampled (solid gray line) poses for the complexes. A: Performance for all 30 protein targets. B: The 11 apo protein cases only, showing that four of the other scoring functions select poor-accuracy (5.4-9.3 Å RMSD) poses for between one and three of the apo cases, possibly because they were parameterized to favor interaction geometries found in holo structures.

Table 3.2: Ligand RMSD values (in Å) of the best predicted docking poses.

PDB entry	Holo/apo	SiteInterlock Score	PoseScore	AutoDock Vina	DSX	DOCK6 Amber Score	X-Score
1a9x	holo	0.66	3.65	0.66	1.88	3.65	3.00
1amu	holo	2.37	0.4	0.40	0.40	0.40	2.37
1b5e	holo	1.93	1.93	2.28	1.12	1.12	1.12
1bgv	holo	3.87	2.03	2.43	2.43	3	2.43
1bx4	holo	0.32	0.10	0.10	0.10	2.21	0.10
1c96	holo	1.04	2.88	2.88	2.88	1.04	2.88
1cbs	holo	2.25	1.27	1.00	2.47	1.50	1.75
1cbx	holo	0.97	0.97	0.97	2.26	1.52	0.97
1ccw	holo	1.97	2.63	0.83	0.83	2.09	0.83
1chm	holo	1.04	1.97	1.97	1.97	1.90	1.97
1com	holo	0.36	1.51	0.36	0.36	1.00	0.36
1coy	holo	1.96	0.24	0.24	0.51	3.19	0.51
1cps	holo	2.26	1.53	1.53	1.73	0.97	1.53
1did	holo	2.78	0.97	1.80	1.80	2.52	1.80
1hwr	holo	1.56	0.77	0.77	0.83	0.83	1.17
1rx1	holo	0.22	0.22	0.22	0.22	0.22	0.22
3ks9	holo	1.21	2.00	2.74	1.21	2.74	1.21
3odu	holo	2.50	0.99	0.99	2.16	2.16	2.16
7tim	holo	1.25	0.66	1.50	1.50	1.25	0.77
16gs	apo	1.75	0.78	1.05	1.05	0.78	0.78
1ahc	apo	1.25	1.50	1.50	1.25	3.00	1.25
1ajz	apo	2.85	6.44	2.85	6.44	9.33	3.02
1gmq	apo	1.23	1.23	1.23	1.23	2.07	1.23
1kem	apo	0.99	0.44	0.44	0.44	2.01	0.73
1nsb	apo	0.70	0.70	0.70	0.70	0.70	1.496
1swa	apo	2.13	0.50	0.50	0.50	1.70	1.70
1tli	apo	0.65	0.65	1.01	5.84	0.83	5.84
1vr1	apo	0.96	1.65	0.96	0.96	0.83	0.96
1ydc	apo	2.18	2.18	2.18	5.35	2.18	2.18
2sga	apo	0.73	0.40	0.40	0.40	1.03	0.81

for the full set of complexes (Figure 3.9). A narrow, linear or flame-like pattern in a plot of scoring function x versus scoring function y values for the dockings indicates that the two scoring functions rank the dockings similarly, whereas a diffuse (globular or more scattered) pattern indicates that the two scoring functions measure different features of the complexes and rank the dockings only partly similarly. The similarity in trends of two scoring functions across the dockings can be summarized by a single number, the nonparametric Spearman rank correlation coefficient, ρ, as shown in Figure 3.9. Unlike the Pearson linear correlation coefficient, the Spearman ρ does not assume a linear relationship between the scoring methods being compared. If two scoring methods rank all the dockings in the same order, a Spearman ρ of 1 will be assigned, whereas a value of 21 indicates the methods rank the dockings in exactly the opposite order, and a value of 0 indicates no correlation in their ranking.

Most pairs of scoring functions evaluated here have a Spearman ρ value in the range of 0.5-0.8 (Figure 3.9), while the correlation between SiteInterlock and other scoring functions is lower,

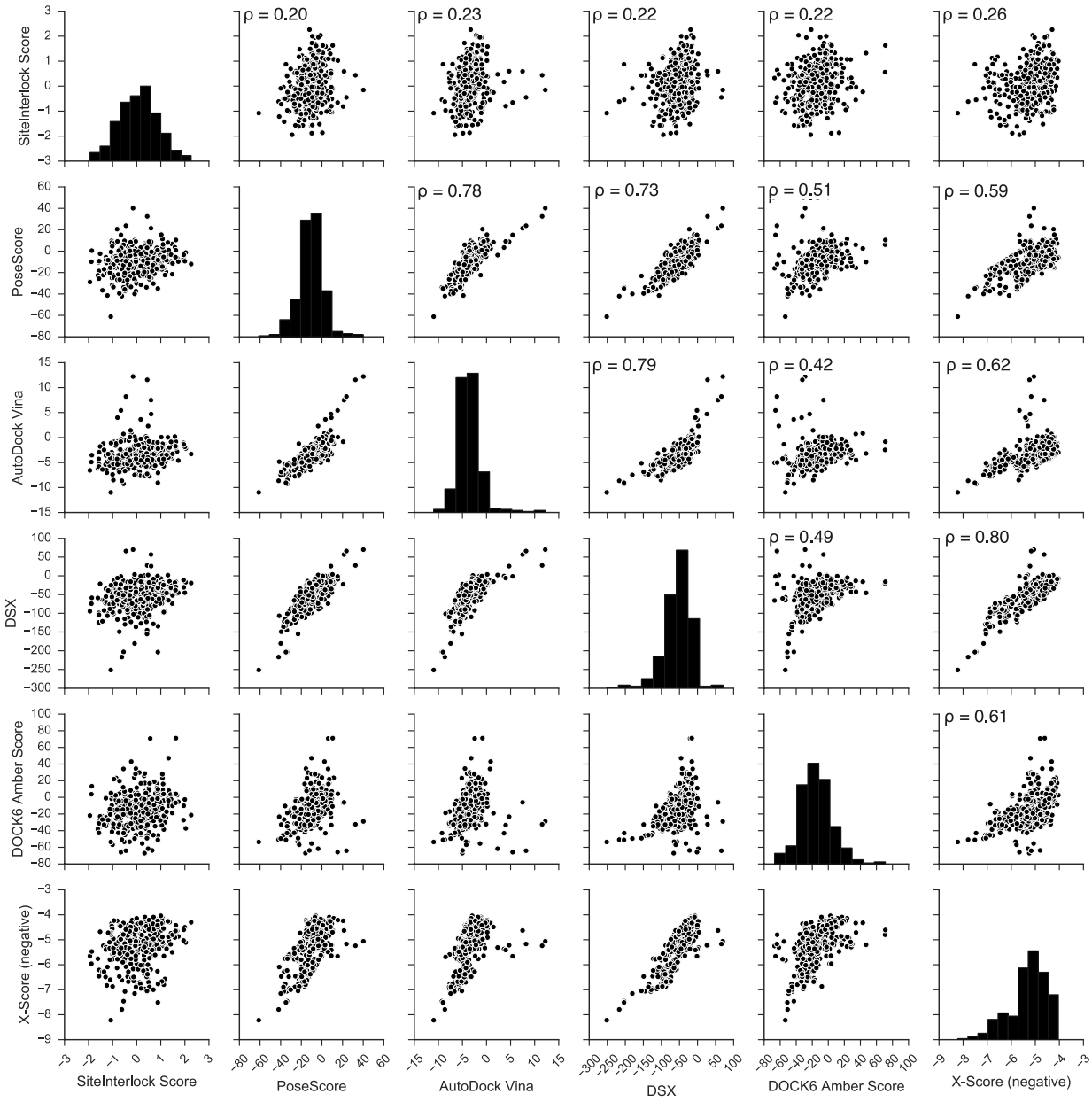


Figure 3.9: Comparison of the values of different scoring functions for all 331 docking poses, as a matrix of pairwise scatter plots. Spearman's rank correlation coefficient, denoted as ρ, is provided for each scoring function pair in the upper triangle, measuring the extent to which the two scoring functions shown in each plot rank the poses in the same order. Along the diagonal appears the histogram of the number of docking poses as a function of score value for each scoring function. The standardization of SiteInterlock score components ProteinAvg and LigandAvg leads to a Gaussian distribution of scores, which helps to distinguish good from average from poor dockings. Some of the other scoring functions exhibit narrow distributions, making the discrimination of good protein-ligand orientations more challenging. To facilitate the comparisons here, X-Score values (last column and row) are presented multiplied by -1, so that more negative values appear as more favorable.

ranging from 0.20-0.26. This indicates that SiteInterlock measures independent feature(s) of the complexes that are not measured by the other methods. SiteInterlock's rigidity measure is novel in that synergy between interactions (their spatial arrangement and coupling) is key to measuring rigidity, rather than just reflecting additive contributions of bonds. Furthermore, this coupling can extend throughout the ligand and binding site rather than being highly localized to the pairs of atoms and functional groups that interact directly. Thus, SiteInterlock can be considered to measure the degree of coupling between interactions in the binding sites, as well as depending on the presence of favorable individual interactions for that coupling to occur.

3.5 Conclusions

SiteInterlock, based on rigidity theory derived from structural mechanics, has been applied here to identify the native complex between a protein and ligand, given the protein structure in either the ligand bound or free conformation and the ligand molecule in a variety of conformations. Several results support the hypothesis that the native complex is characterized by enhanced interfacial rigidity involving both molecules:

- The majority of holo complexes (17 out of 19 diverse proteins) display increased protein rigidity at the interface when the protein is bound, while the remaining two appear equally rigid.
- Including ligand as well as protein interfacial rigidification improves discrimination of the native complex from misdocked complexes.
- SiteInterlock rigidity performs competitively with the best of five commonly used, well-developed docking scoring functions in discriminating near-native poses from a range of decoy poses.
- For the majority (29) of the complexes, SiteInterlock selects ligand poses that are within 2.8 Å RMSD of the native pose, when given a set of sampled (not crystallographic) ligand

conformations. For 25 of the complexes, the best-scoring pose is within 1.5 Å RMSD of the best-sampled pose.

- SiteInterlock has the advantage of avoiding very poor dockings (5 Å or greater RMSD), which are an issue for four of the other scoring functions.

More fundamentally, this work shows that rigidification of the cooperative network of noncovalent bonds upon complex formation is a signature of binding interfaces that is sufficient to detect the native complex. This measure of interaction coupling between the protein and ligand, rather than purely additive interactions, may explain why SiteInterlock rigidity values for complexes have a modest correlation with the values of other scoring functions. Thus, SiteInterlock provides a new feature – interfacial rigidity – and a new way of assessing protein-ligand interfaces that can be used alone or in combination with other methods. We anticipate many useful applications of this interfacial rigidity method for structure-based ligand discovery, with the potential to also aid ligand fitting in crystallography for complexes with moderate resolution.

3.6 Acknowledgements

We would like to thank OpenEye Scientific Software (Santa Fe, NM) for providing academic licenses for the QUACPAC, OMEGA, and OEchem software packages used in this work.

CHAPTER 4

ENABLING THE HYPOTHESIS-DRIVEN PRIORITIZATION OF LIGAND CANDIDATES IN BIG DATABASES: SCREENLAMP AND ITS APPLICATION TO GPCR INHIBITOR DISCOVERY FOR INVASIVE SPECIES CONTROL

Adapted with permission from Raschka, Sebastian, Anne M. Scott, Nan Liu, Santosh Gunturu, Mar Huertas, Weiming Li, and Leslie A. Kuhn. "Enabling the hypothesis-driven prioritization of ligand candidates in big databases: Screenlamp and its application to GPCR inhibitor discovery for invasive species control." *Manuscript in revision.*

4.1 Abstract

While the advantage of screening vast databases of molecules to cover greater molecular diversity is often mentioned, in reality, only a few studies have been published demonstrating inhibitor discovery by screening more than a million compounds for features that mimic a known three-dimensional ligand. Two factors contribute: the general difficulty of discovering potent inhibitors, and the lack of free, user-friendly software to incorporate project-specific knowledge and user hypotheses into 3D ligand-based screening. The Screenlamp modular toolkit presented here was developed with these needs in mind. We show Screenlamp's ability to screen more than 12 million commercially available molecules and identify potent in vivo inhibitors of a G protein-coupled bile acid receptor within the first year of a discovery project. This pheromone receptor governs sea lamprey reproductive behavior, and to our knowledge, this project is the first to establish the efficacy of computational screening in discovering lead compounds for aquatic invasive species control. Significant enhancement in activity came from selecting compounds based on one of the hypotheses: that matching two distal oxygen groups in the three-dimensional structure of the pheromone is crucial for activity. Five of the 15 most active compounds were selected by this hypothesis. A second hypothesis – that presence of an alkyl sulfate side chain results in high activity – identified another 5 compounds in the top 10, demonstrating the significant benefits of hypothesis-driven screening.

4.2 Introduction

4.2.1 Virtual screening for inhibitor discovery

Within the field of virtual screening, structure or receptor-based approaches involve the docking of small molecules into the three-dimensional (3D) structure of an enzyme or receptor binding site to select a set of molecules for experimental testing as activators or inhibitors of the protein. The prioritization of candidates is typically based on ranking the molecules by their predicted binding affinities (Ferrara et al., 2004). However, applications of structure-based screening are limited by the

availability of accurate three-dimensional (3D) structures of the target protein. Moreover, the large number of geometrically feasible solutions when both molecules are considered flexible means that thorough sampling of such docking poses is computationally impractical, even for state-of-the-art computing clusters. As a result, most currently used docking solutions treat the ligand candidate as flexible and the protein as only partly flexible via limited side-chain sampling (Cozzini et al., 2008). Even under these partially-flexible protein assumptions, ligand docking is very computationally expensive. It is not feasible for most academic research groups to dock millions of small, flexible molecules, which requires the use of computing clusters or commercial cloud services (Capuccini et al., 2017). An equally significant problem is that prediction of $\Delta G_{binding}$ of protein-ligand complexes has remained prone to errors typically on the order of several kcal/mol (a substantial percentage of the total $\Delta G_{binding}$), causing the ranking of compounds to be approximate at best (Merz Jr, 2010). This problem is likely to remain difficult and improve incrementally rather than rapidly, due to the difficulty of measuring conformational energies, entropy changes, electrostatics, and solvent contributions to ligand binding (Hou et al., 2011). The most accurate approaches are only feasible for assessing a small set of compounds.

Ligand-based screening, in which database compounds are compared to a known active compound (rather than docked to the protein target) to discover mimics, is frequently employed by pharmaceutical companies due to the success rate and the unavailability of 3D protein structures for many targets of interest. Generally, ligand-based virtual screening is computationally more efficient than structure-based approaches (Drwal & Griffith, 2013). An additional advantage is that errors in modeling protein and solvent flexibility do not come into play in ligand similarity-based scoring, which is based solely on the extent to which a candidate matches the known ligand in 3D volume and charge or atom-type distribution. Ligand-based screening can outperform structure-based approaches in the speed and the enrichment of active molecules (Hawkins et al., 2007; McGaughey et al., 2007; Hu et al., 2012). Furthermore, when performed with a single known active compound for comparison, 3D ligand-based screening is capable of identifying molecular mimics spanning a wide space of structural scaffolds and chemotypes. This desirable feature, known

as lead or scaffold hopping (Rush III et al., 2005; Muegge & Mukherjee, 2016), is important since a significant percentage of inhibitory compounds may undergo attrition during the pharmacological and clinical development process due to not meeting criteria for in vivo absorption, distribution, metabolism, excretion, and toxicity.

The Screenlamp project started when we sought to fill a gap, by developing freely available, effective software to enable typical academic biochemical research groups, rather than just computational chemistry experts, to test their hypotheses about the importance of specific functional groups or pharmacophores (3D spatial relationships between functional groups) that lead to high ligand activity, when performing a broad search for compounds or scaffolds with significant similarity to a known ligand. In a random database, the probability of finding one or more good lead molecules with substantial affinity for the protein target via close mimicry of a known ligand increases with the number of molecules screened. Thus, our second goal was to make the organization and screening of very large databases of millions of commercially available compounds accessible to a typical research lab, rather than being restricted to researchers with parallel computing expertise. Some tools exist to aid users in ligand-based screening, but they are limited by the level of molecular detail they support, the flexibility of use, and cost. The SwissSimilarity webserver was recently launched to support ligand-based virtual screening (Zoete et al., 2016). While this service includes 10.6 million drug-like molecules from ZINC, its screening is based on non-superpositional methods that do not consider the 3D volumes or spatial arrangement of functional groups. Phase is a commercial tool developed by Schroedinger, which allows users to perform 3D ligand-based screening based on abstract hydrogen-bond acceptor and donor, hydrophobic, aromatic, and charged pharmacophore points, which the software derives from known actives (Dixon et al., 2006). Aside from the barrier of substantial licensing costs, its integration as part of the Schroedinger graphical user interface, including assignment of ligand protonation states and conformers and the use of a proprietary scoring function and eMolecules database, limits its flexibility. We have found partial charge and protonation state assignment, quality of 3D conformer sampling, flexible identification of pharmacophores and querying based on functional group relationships, and 3D overlays and similarity scoring to be

variable in quality between existing software packages while being essential to screening success. The ability to choose the modules that work best for a project and provide a freely available, flexible workflow for 3D ligand-based discovery are supported by Screenlamp.

The fact that 3D ligand-based screening of millions of compounds still presents a substantial technical challenge to most users is underscored by only a handful of inhibitor-discovery publications appearing in the literature for this approach over the past 13 years (Nagamine et al., 2009; Koes et al., 2012; Miller & Roitberg, 2013; Murgueitio et al., 2014; Almela et al., 2015; Allen et al., 2015a; Mirza et al., 2016; Johnson & Karanicolas, 2016), in comparison with dozens of publications for screening by docking of similar-sized databases. With Screenlamp, volumetric and partial charge-based alignment of fully flexible molecules and analysis of 3D chemical group matches can be performed on millions of commercially available molecules, such as the ZINC drug-like database (http://zinc.docking.org; Irwin & Shoichet, 2005), within a day on a typical desktop computer. Here we demonstrate its successful application to a challenging problem: discovery of both steroidal and non-steroidal inhibitors with IC50 values under $1\mu M$ for an olfactory GPCR activated by a bile acid pheromone (Li et al., 2002). Because the molecular weight of the pheromone is at the upper-limit for drug-like compounds, the discovery of active compounds benefited from Screenlamp's ability to search expanded sets of molecules from ZINC and the Chemical Abstracts Service Registry (https://www.cas.org/content/chemical-substances).

4.2.2 Pioneering aquatic invasive species control and GPCR inhibitor discovery through virtual screening

This pheromone inhibitor discovery project presents a novel, behaviorally selective approach to aquatic invasive species control, which in the past has involved *in vivo* testing of thousands of pesticides. The sea lamprey is an invasive species that has had greatly deleterious impacts since the 1950s on both the native ecology and commercial fishery of the Great Lakes of North America. Ongoing efforts at reducing sea lamprey populations are labor-intensive and cost millions of dollars per year (Hansen & Jones, 2008). They include the use of in-stream barriers to prevent lamprey

from reaching spawning areas (Lucas et al., 2009) and the application of trifluoromethyl nitrophenol (TFM), a larval lampricide (McDonald & Kolar, 2007). TFM has been successful, leading to a decrease of the sea lamprey population by over 90% between 1960 and 1970 (Scott & Crossman, 1973). However, the discovery of new sea lamprey control approaches remains a high priority for the binational Great Lakes Fishery Commission. Occasionally, TFM has shown off-target toxic effects to amphibians, trout, and most importantly, lake sturgeon, which the U.S. Fish and Wildlife Service lists as threatened or endangered in nineteen of the twenty states of its historic range (https://www.fws.gov/midwest/sturgeon/biology.htm; date accessed: Sept. 2, 2017; Becker, 1983; Boogaard et al., 2003). A recent sea lamprey control approach involves the baiting of traps (Johnson et al., 2009, 2015) with the main component of the male sea lamprey mating pheromone 3kPZS (3-keto petromyzonol sulfate; 7α,12α,24-trihydroxy-5α-cholan-3-one-24-sulfate), which is an agonist for the sea lamprey odorant receptor 1 (SLOR1).

4.2.3 G-protein coupled receptors and olfactory receptors

SLOR1 (UniProtKB ID: S4RTH2) and other pheromone and olfactory receptors in the sea lamprey are categorized in class A of the G protein-coupled receptors (GPCRs) based on sequence homology (Libants et al., 2009). Class A or rhodopsin-like GPCRs form the largest of the five GPCR superfamilies (Katritch et al., 2012). GPCRs play an important role in human medicine, with about half of all human drugs targeting GPCRs and their signaling (Lundstrom, 2009). A well-known agonist of β1-adrenergic receptor, the closest human structural homolog of SLOR1, is epinephrine (also known as adrenaline). Antagonists of this receptor known as beta blockers are commonly used for controlling blood pressure and glaucoma. In humans, olfactory receptors comprise 388 out of our 779 GPCRs (http://gpcr.usc.edu), indicating their importance for responding to chemical cues in the environment, such as oxygen (Chang et al., 2015), smoke (Bessac & Jordt, 2010), scents released by rotten meat (Hussain et al., 2013), scents associated with nutrients (Milligan et al., 2014), and pheromones (Liberles, 2014). Ligands for class A GPCRs are correspondingly diverse, including steroids, peptides, light-responsive chromophores, neurotransmitters, lipids, nucleotides,

and chemokine proteins (Niimura et al., 2009). In insects, non-GPCR olfactory receptors (Benton et al., 2006) also play key roles in sensing and responding to repellants such as DEET, as well as in detecting pheromones that lead to mating and reproduction (Kain et al., 2013).

Sea lamprey mating is governed by sex pheromones released by spermiated mature males. Ovulated females are attracted by the 3kPZS secreted in spawning areas. We hypothesized that blocking the detection of 3kPZS by female sea lamprey will halt the reproductive cycle and reduce the sea lamprey population. The aim of our high-throughput screening was thus to identify sea lamprey-selective inhibitor mimics of 3kPZS that are environmentally benign. Screenlamp was developed and used to screen 12 million commercially available small organic molecules. Those with the most significant volumetric and electrostatic similarity to 3kPZS were further prioritized within Screenlamp by filtering compounds according to a series of hypotheses about the importance of individual chemical groups for activity. In vivo olfactory assays of the selected 299 compounds were then performed, testing their ability to block 3kPZS olfactory responses, and resulting in the discovery of several classes of inhibitors with sub-micromolar IC50 values. Beyond meeting the goals of discovering potent 3kPZS pheromone inhibitors and pioneering the use of computer-aided drug discovery for invasive species control, this project aims to advance other researchers' success in ligand discovery by making the Screenlamp software publicly available.

4.3 Methods

4.3.1 Driving structure-activity hypothesis development by structural modeling of 3kPZS-receptor interactions

Of the available GPCR crystal structures in the Protein Data Bank (Berman et al., 2000), nociceptin, adenosine, and β1-adrenergic receptors are all structural homologs (Sander & Schneider, 1991) based on pairwise identity of 24-27% covering most of the 330-residue SLOR1 sequence (Altschul et al., 1990). SLOR1 is most similar to the β1-adrenergic receptor, based on evaluation of sequence similarity in the extracellular loops and the inter-helical cleft comprising the orthosteric (activating ligand) binding site, the absence of non-helical insertions within transmembrane helices, and the

conservation of motifs, including the E/DRY ionic lock motif in helix 3, which interacts with acidic residues in helix 6 in the inactive state of class A GPCRs (Rosenbaum et al., 2009).

A homology-based structural model of SLOR1 was constructed using the crystal structure of avian β1-adrenergic receptor as a template (PDB code: 2vt4; Warne et al., 2008), by using ModWeb Modeller version SVN.r972 (Eswar et al., 2008). The protein backbone structures of other related class A GPCRs with their bound ligands, such as rhodopsin, adenosine (A2A), and β1-adrenergic receptors were overlaid with the SLOR1 model to define the orthosteric binding region in SLOR1. All favorable-energy conformations of 3kPZS, generated via OpenEye Omega (version 2.4.1), were docked into the SLOR1 ligand binding cavity to predict their mode of interaction by using the SLIDE software with default settings (Zavodsky et al., 2002).

4.3.2 Development of Screenlamp, a hypothesis-based screening toolkit

To facilitate the virtual screening of millions of flexible, three-dimensional structures for ligand discovery, including 3kPZS antagonists, the Screenlamp toolkit was developed in Python. It leverages high-performance memory-buffered multi-dimensional arrays (Van Der Walt et al., 2011) and data frames (McKinney, 2010; Raschka, 2017a). Screenlamp first allows selection of those molecules meeting specific physicochemical or spatial properties, such as the presence of two functional groups within a certain distance. Screenlamp then interfaces with robust tools that are freely available to academic researchers to assign partial charges to ligand atoms, sample energetically favorable 3D conformers, and generate 3D overlays): the OpenEye molcharge utility in QUACPAC for assigning partial charges (Halgren, 1996; Jakalian et al., 2002), OMEGA (Hawkins et al., 2010; Hawkins & Nicholls, 2012) for conformer generation, and ROCS (Hawkins et al., 2007; Sheridan et al., 2008) for 3D molecular overlays with the reference molecule (for example, 3kPZS). The modules and tasks that can be performed within Screenlamp are summarized in Figure 4.1. While an early internal version relied on an SQL database (Chamberlin & Boyce, 1974) for recordkeeping and an HDF5 database (Folk et al., 2011) for storing 3D coordinates of molecules, the application program interface has been simplified and computationally accelerated. Screenlamp

works efficiently without SQL or HDF5 and can be applied to any molecular database organized as multi-MOL2 (3D) formatted files.

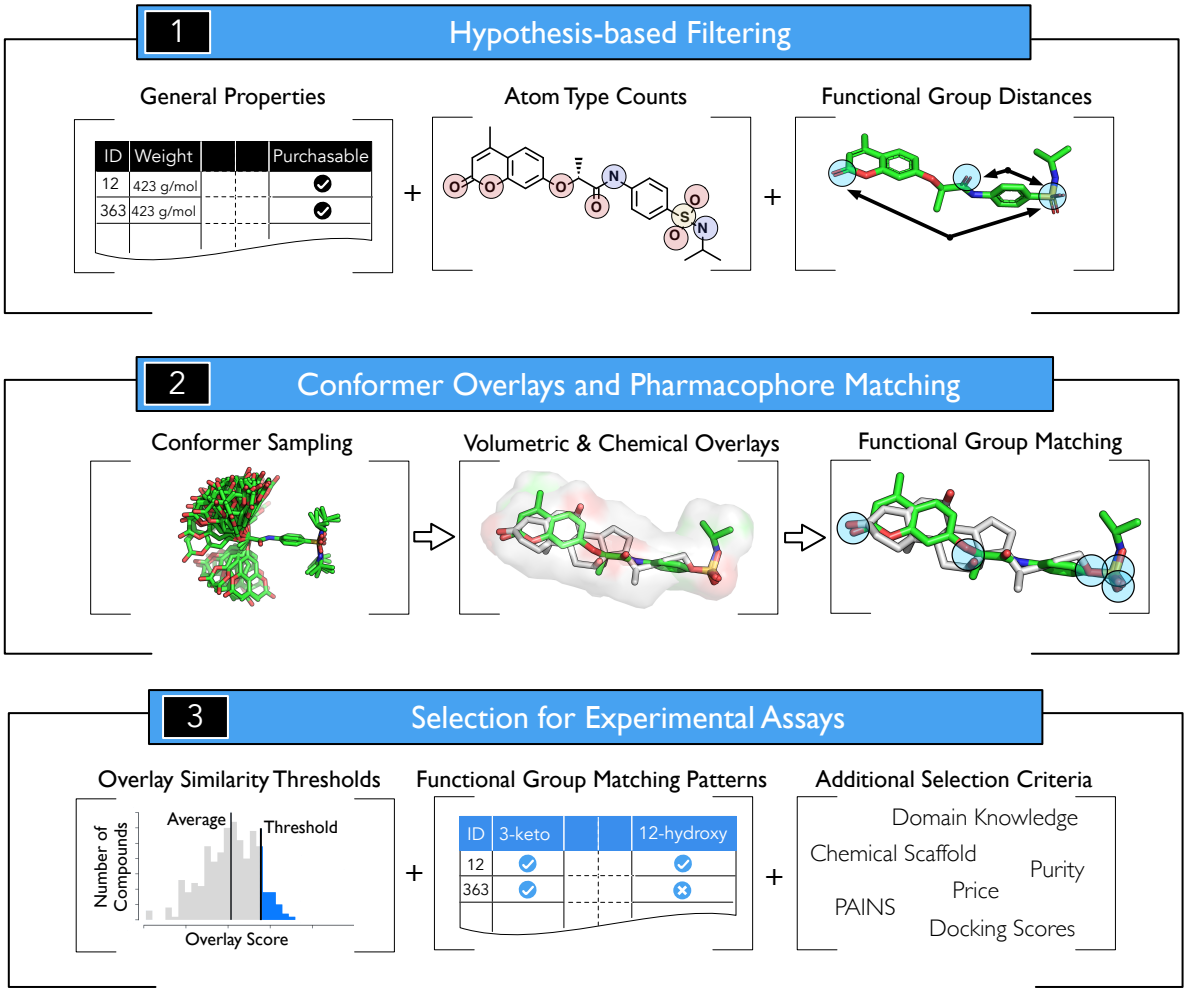


Figure 4.1: Summary of the tools provided or augmented by Screenlamp. (1) Filtering tasks that can be performed within Screenlamp to meet hypothesis-driven criteria and retrieve the structures of a subset of candidate molecules. (2) Once flexible conformers of the candidate database molecules have been sampled and overlaid with the reference molecule (for example, by using Omega and ROCS from OpenEye), Screenlamp can identify functional group matches in those pairwise overlays to discover functional group mimics of a reference molecule. (3) Based on the information that is available from the 3D overlays and functional group matching, as well as user-specified selection criterion, molecules are ranked for experimental testing.

The modules in the Screenlamp toolkit (Figure 4.1) allow researchers to rearrange and recombine subsets of filtering, alignment, and scoring steps in a pipeline that meets their own hypothesis-driven selection criteria. For instance, a module within Screenlamp allows users to select subsets

of molecules based on properties such as molecular weight, number of hydrogen bond acceptors and donors, number of rotatable bonds, or any other property data given in column format in a text file such as the property files (for example, http://zinc.docking.org/db/bysubset/23/23_prop.xls) of the ZINC commercially available compound database (Irwin & Shoichet, 2005). The use of molecular property data – molecular weight, number of freely rotatable bonds, etc. – can also be obtained by using open-source chemoinformatics tools such as RDKit (http://www.rdkit.org), while being optional for use in Screenlamp. Additional Screenlamp modules are available for filtering, such as selecting only those molecules that contain functional groups of interest, or optionally, functional groups in a particular spatial arrangement. Based on the 3D alignments, Screenlamp provides a module that can generate fingerprints representing the presence or absence of spatial matching between the database entries and a series of 3D functional group matches in the reference molecule. These molecular fingerprints and functional group matching patterns can further be used for exploratory data analysis or machine learning-based predictive modeling of structure-activity relationships (Mitchell, 2014). Along with volumetric and electrostatic scores provided by the overlay tool, and filtering based on molecular properties, the user can then test hypotheses about the biological importance of these user-specified features by identifying a matching set of compounds and procuring these for biological assays.

Once the user has selected a subset of molecules according to the current hypothesis to be tested, expressed as a set of criteria on presence or absence of certain atoms or properties, the corresponding structures are sent for conformer generation and 3D alignment with the known ligand reference molecule (typically, a known inhibitor, agonist, or substrate). The following sections provide details on how a typical workflow was implemented, in this case for the discovery of potent mimics of 3kPZS as pheromone antagonists. The Screenlamp software and full documentation are available to download from GitHub (https://github.com/psa-lab/screenlamp).

4.3.3 Preparation of millions of drug-like molecules for ligand-based screening

The 3D coordinate files, in Tripos MOL2 format, of 12.3 million molecules were downloaded from ZINC12 (Irwin & Shoichet, 2005) using the "drugs now" criteria (compounds with drug-like properties, available off-the-shelf). They were processed as illustrated in Figure 4.1 according to the hypothesis criteria, which are summarized in the paragraph *Hypothesis-based candidate selection* at the end of this section. Additional screening data sets of antagonist candidates were prepared, as described below, to enable the testing of close analogs of 3kPZS and known ligands of GPCRs.

Combinatorial analog dataset. Isomeric SMILES string (simplified molecular-input line-entry system) structural representations (Weininger, 1988) of 332 close variants of 3kPZS were created by sampling different combinations of alternative functional groups at the 3, 7, and 12 positions in 3kPZS (Figure 4.2) and different configurations (5-α planar or 5-β bent relationship between the A and B rings) of the steroid ring system. These SMILES representations were used as search queries in SciFinder (http://www.cas.org/products/scifinder) to identify purchasable compounds that exactly (or nearly exactly, showing $\geq$ 99 percent similarity) match the 332 analogs. Chemical Abstract Service Registry (CAS Registry; http://www.cas.org/content/chemical-substances/faqs) identifiers were found for 84 commercially available molecules. The corresponding SMILES strings were translated into 3D structures for virtual screening by using OpenEye QUACPAC/molcharge (version 1.6.3.1; OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com) with the AM1BCC (Jakalian et al., 2002) force field for partial charge assignment.

CAS Registry steroids. The ZINC database covers many, but not all, vendors of small organic molecules; thus, the CAS Registry of 91 million compounds (https://www.cas.org/content/chemical-substances) was searched with SciFinder Scholar (Chemical Abstracts Service, Columbus, OH) for all commercially available steroid molecules that were not already present in the ZINC database. Using SciFinder (which limits the number of molecules that can be processed at a time to 100), batches of CAS Registry steroid structures were exported and processed into SMILES

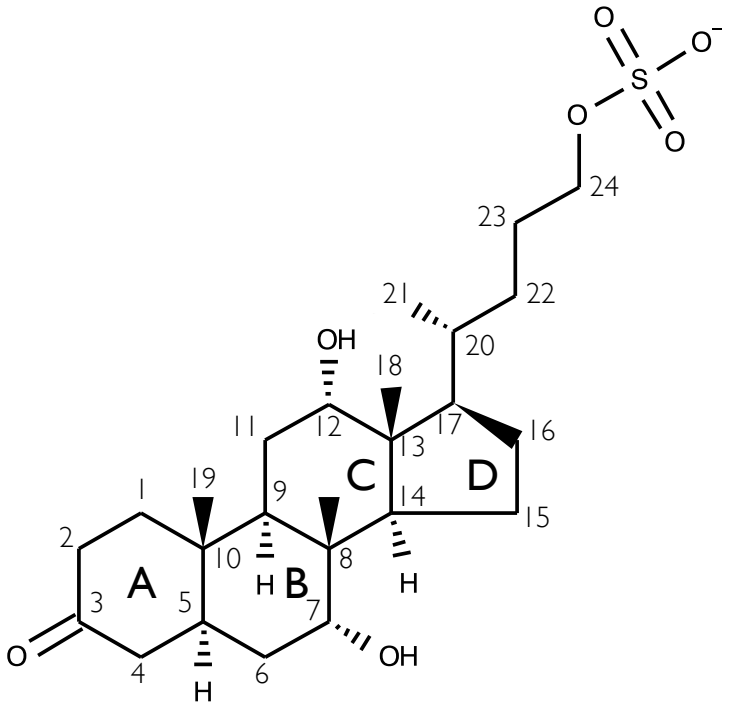


Figure 4.2: The molecular structure of 3kPZS.

strings using CACTUS (http://cactus.nci.nih.gov). Three-dimensional structures were created from the SMILES strings as described for the combinatorial analog dataset above, resulting in 2,995 additional steroids for screening.

GPCR Ligand Library (GLL). The GLL database consists of approximately 24,000 known ligands for 147 GPCRs (http://cavasotto-lab.net/Databases/GDD/; Gatica & Cavasotto, 2012). To prepare this database for our virtual screening pipeline, partial charges were added to the existing 3D structures of these molecules using OpenEye QUACPAC/molcharge with the AM1BCC force field (Jakalian et al., 2002).

4.3.4 Identification of incorrect steroid substructures in molecular database

In version 12 of ZINC (http://zinc.docking.org), if a vendor did not provide complete stereochemistry information for chiral centers in a steroid molecule, up to four different stereoisomeric structures were automatically provided by ZINC, each with a separate ID. However, at most one of those four structures had a valid steroid configuration (with a 5-α planar or 5-β ring structure and 18- and 19-methyl group orientations as shown in Figure 4.2). Thus, we developed a custom steroid checking tool using the OpenEye OEChem toolkit by comparing each molecule with an isomeric SMILES representation of the canonical steroid core atom connectivity and chirality, to filter out invalid steroid configurations. This steroid checker is included in Screenlamp and has recently been implemented in ZINC, by coordination with the developers at UCSF.

4.3.5 Step 1: Hypothesis-based molecular filtering

The Screenlamp toolkit provides a user-friendly interface to efficiently select those molecular structures that are relevant for a given screening hypothesis or objective. For instance, the first step in the 3kPZS inhibitor screening (Figure 4.3) selected those drug-like molecules listed as commercially available by either ZINC or CAS. Drug-like properties were defined as satisfying Lipinski's rule of 5 (Lipinski et al., 1997), plus a rotatable bond criterion to filter out highly flexible molecules because their significant loss of entropy upon protein binding detracts from the $\Delta G_{binding}$ between receptor and ligand. The drug-like criteria used were: (1) molecular weight between 150 and 500 g/mol; (2) octanol-water partition coefficient less than or equal to 5; (3) 5 or fewer hydrogen bond donors and 10 or fewer hydrogen bond acceptors; (4) polar surface area less than 150 $Å^2$; and (5) fewer than 8 rotatable bonds. In addition, the filtering query excluded all molecules that were flagged as invalid steroids.

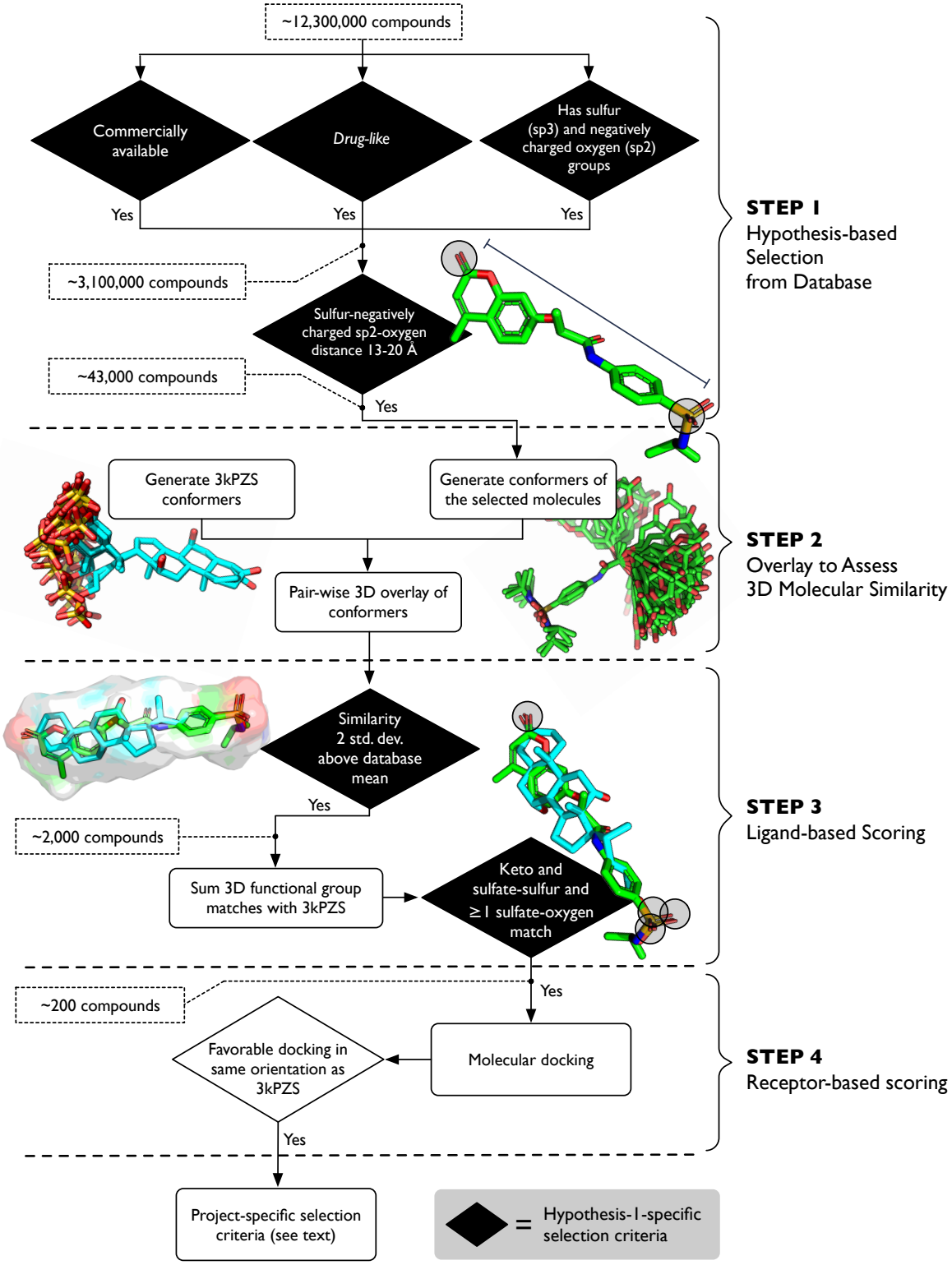


Figure 4.3: Using Screenlamp to identify compounds to test the hypothesis that compounds with negatively charged sulfate and sp^2-hybridized oxygen groups matching the 24-sulfate-3-keto oxygen distance in 3kPZS will mimic 3kPZS and block its binding.

4.3.6 Step 2: Sampling favorable molecular conformations

Low-energy conformations of the reference molecule 3kPZS were generated by sampling rotatable bond torsions with OpenEye OMEGA (version 2.4.1; Hawkins et al., 2010), using its default settings. Forty-eight favorable 3D conformers in a somewhat extended rather than folded conformation, required to fit within the ligand-binding site of the SLOR1 structure (Figure 4.1), were kept (Figure 4.3); specifically, the distance between the 3-keto group and the sulfur atom in the sulfate group in these conformers was 13-20 Å. Up to 200 favorable-energy conformations, following conformational clustering by OMEGA, were retained for each of the database molecules selected by the filtering steps.

4.3.7 Generation of overlays to compare molecular shape and charge distribution with a known ligand

In addition to property and pharmacophore-based filtering, Screenlamp invokes the ROCS (Hawkins et al., 2007) software to generate 3D molecular overlays to evaluate similarity in volumetric and partial charge distributions. Thus, 48 low-energy conformers of the 3kPZS reference molecule were overlaid with up to 200 conformers for each of the selected database molecules using OpenEye ROCS (version 2.4.6). The 3D overlays were ranked by the TanimotoCombo metric, which consists of equally contributing components that assess the degree of volumetric (shape) and partial charge ("color") overlay. The TanimotoCombo metric requires perfect match between all parts of two molecules (rather than exact substructure matches) to achieve a perfect score, which ranges between 0 (no overlap/similarity) and 2 (perfect overlap). For each database entry, only the configuration of the best-overlaid pair of conformers between the 3kPZS reference and the database molecule was saved.

4.3.8 Step 3: Ligand-based scoring

Molecules with a similarity score two standard deviations above the mean, showing a high degree of similarity to the 3kPZS reference molecule, were considered as potential 3kPZS mimics and

evaluated for functional group matches with 3kPZS. Functional group matches were identified based on the atom type, atomic charge, and hybridization of the following 3kPZS groups: 3-keto, 3-hydroxyl, 7-hydroxyl, and 12-hydroxyl oxygens; 18- and 19-methyl groups; sulfate ester oxygen and the three sulfate terminal oxygens (Figure 4.2). In each case, an atom (functional group component) in the database molecule was considered to match if it was within 1.3 Å of the same functional group in 3kPZS (matching the atom type, hybridization, and charge), given the highest-scoring ROCS overlay of the database molecule with 3kPZS.

4.3.9 Docking the highest-ranking compounds with the SLOR1 structural model to assess goodness of fit

For the selected set of inhibitor candidates, prioritized by multiple criteria as described in the Results, flexible docking was performed by using SLIDE (version 3.4) with default settings (Zavodsky et al., 2002) to compare the mode of interaction of a given ligand candidate with 3kPZS docked into SLOR1. A ligand docking was considered to mimic 3kPZS if a salt bridge was formed with His110 in SLOR1, similar to that observed for the 3kPZS sulfate tail. In addition, significant hydrophobic, rather than amphipathic, interaction of a ligand candidate with the hydrophobic wall in the ligand binding site of SLOR1 was evaluated as a characteristic of 3kPZS interaction.

4.3.10 Hypothesis-driven selection of ligand candidates

Upon screening the compounds from the above databases to identify those with significant ROCS TanimotoCombo scores and functional group matches with 3kPZS as tabulated by Screenlamp, top-scoring compounds were selected for electro-olfactogram (EOG) assay to directly assess their ability to reduce the sea lamprey olfactory response of 3kPZS, according to the following series of hypotheses. Many of these hypotheses involve the presence of oxygen in a spatial position equivalent to the 3-keto oxygen in 3kPZS and negatively charged oxygen in positions equivalent to the terminal sulfate oxygens in 3kPZS. The focus on these functional groups is based on prior results, which indicated that both 3-O and sulfated tail groups were associated with high olfactory potency

(Li et al., 1995) and were present in a few other steroid compounds known to elicit 3kPZS-like activity (Burns et al., 2011). Specific hypotheses that were tested:

1. *Compounds matching the 3-keto and one or more sulfate oxygens in extended conformations of 3kPZS*, with TanimotoCombo score greater than 2 standard deviations above the mean. This tests the hypothesis that high overall shape and electrostatic similarity and matching the 3-keto and sulfate groups are sufficient to mimic and block 3kPZS activity.

2. *Similar to the above, compounds with 3-hydroxyl groups spatially matching the 3-keto moiety in 3kPZS* were selected to test the hypothesis that 3-hydroxyl containing compounds can block 3kPZS olfaction. Additional criteria for this set: ROCS similarity score to 3kPZS of 0.8 or above, high ROCS electrostatics complementarity score (0.25 or above), matching a sulfate terminal oxygen and at least one of the other functional groups in 3kPZS (sulfate oxygen, hydroxyl or steroid methyl substituents), and docking with the sulfate group proximal to His110 in SLOR1 with a favorable predicted $\Delta G_{binding}$ (< -7kcal/mol).

3. *All compounds with a planar steroid ring system and alpha configuration of hydroxyl groups matching 3kPZS* (rather than equatorial configuration), 3-keto and sulfate oxygen matches, and ROCS TanimotoCombo scores greater than 0.65. This set tests the hypothesis that close steroidal analogs matching the oxygen-containing groups in 3kPZS will mimic 3kPZS activity. The emphasis on planar (5-α) steroids derives from the fact that sea lamprey are the only fish that synthesize planar steroids with sulfated tails (Hagey et al., 2010), and these features are expected to be species-selective olfactory cues. In fact, 5-β steroid relatives of 3kPZS are far less potent (Burns et al., 2011).

4. *Phosphate or sulfate tail analogs. Aliphatic chains with at least 3 methyl(ene) groups terminating in a phosphate or sulfate group*were identified by the ZINC search tool, to test whether mimicking the sulfate tail moiety of 3kPZS alone is sufficient to block 3kPZS olfaction, and whether a phosphate group can mimic the sulfate group.

5. *Compounds with a high degree of shape/electrostatic match with the C and D steroid rings and sulfated aliphatic tail structure in PAMS-24*, another sea lamprey pheromone identified by the Li lab (Brant, 2015). This region corresponds to atoms 8, 9, and 11-24 plus the sulfate group (Figure 4.2), with the addition of an isopropyl group at C-24. This set tests whether compounds matching a tail fragment inhibit the 3kPZS response.

6. *5-β steroid structures with at least 2 sulfate oxygen matches with 3kPZS or at least 5 functional group (oxygen and methyl) matches* were chosen to test whether bent rather than planar steroids can block 3kPZS olfaction. (None of the compounds could match both the 3-keto and sulfate tail, due to the bent geometry of the 5-β steroid ring system when overlaid with ROCS to match all atoms.) Prior work on 5-β steroids tested their ability to act as 3kPZS agonists rather than inhibitors (Burns et al., 2011).

7. *Compounds with highly negative sulfate oxygen-matching atoms* (with charges at least 0.3 units more negative than the sulfate oxygen charge in 3kPZS) were selected, testing the hypothesis that strongly negatively charged groups can form stronger interactions with SLOR1 (e.g., salt bridge with His110) and outcompete 3kPZS for binding.

8. *Compounds with highly negative sulfate oxygen-matching atoms* (with charges at least 0.3 units more negative than the sulfate oxygen charge in 3kPZS) were selected, testing the hypothesis that strongly negatively charged groups can form stronger interactions with SLOR1 (e.g., salt bridge with His110) and outcompete 3kPZS for binding.

9. *Epoxide-containing steroids.* Epoxide functional groups are labile, tending to spring open due to bond strain and react with nearby protein groups. Previous research (Davis et al., 2007) indicated that epoxide cross-links can be site-specific, preferring histidine side chains. 3kPZS-like, epoxide-containing steroids were tested because cross-linking with the active-site His110 in SLOR1 could result in very strong inhibition of the 3kPZS receptor. Epoxide opening or cross-linking from the equivalent of the 3-oxygen position in 3kPZS could also

create an antenna-like group, potentially making favorable interactions with the collar of the binding site as has been found in other GPCRs (Kruse et al., 2013).

10. *Taurine tail-containing steroids with significant overall chemical similarity to 3kPZS.* Taurolithocholic acid was observed to significantly block the olfactory response of sea lamprey to 3kPZS. Other compounds with taurine tails and overall good matches to 3kPZS that could also block its binding to SLOR1 were identified as candidates for testing.

11. *CAS steroids with high volumetric and electrostatic similarity to 3kPZS.* The top 25 steroids from the CAS Registry, ranked by volumetric and electrostatic similarity to 3kPZS in ROCS overlays, were selected for experimental assays. These compounds are highly similar to 3kPZS while not being biased by prior knowledge of activity determinants.

12. *Compounds known to be bioactive in complex with the β1-adrenergic receptor*, the GPCR of known structure with highest binding site sequence similarity to SLOR1. Three molecules known to be active versus β1-adrenergic receptor were selected for assaying: carvedilol (agonist; ZINC01530579), atenolol (selective antagonist; ZINC00014007), and dobutamine (partial agonist; ZINC00003911).

4.3.11 Assays to measure inhibition of olfactory response of 3kPZS

Electro-olfactogram assays (EOGs) are commonly used to measure *in vivo* olfactory responses to environmental stimuli in vertebrates (Scott & Scott-Johnson, 2002). EOGs record the sum of action potentials (the field potential) generated upon the activation of olfactory receptors (predominantly GPCRs) in the olfactory epithelia after exposure to an odorant. The sea lamprey EOG assays were conducted following a standard protocol described in Brant et al., 2016. Adult sea lamprey were anesthetized with 100 mg/L of 3-aminobenzoic acid ethyl ester (MS222, Sigma-Aldrich Chemical Co.) and injected with 3 mg/kg gallamine triethiodide (Sigma-Aldrich Chemical Co.). Then, the gills were exposed to a continuous flow of aerated water with 50 mg/L MS-222 throughout the experiments. All tested compounds were delivered directly to the olfactory rosette using a small

capillary tube. Water used in the EOGs was charcoal filtered fresh water. At the beginning of each experiment, and after each candidate compound test, the olfactory rosette was flushed with charcoal filtered fresh water for 2 minutes before the responses to 3kPZS (10^{-6} M) and L-arginine (10^{-5} M) were recorded. To test the effect of the candidate inhibitor compounds on the olfactory detection of 3kPZS, the olfactory rosette was continuously exposed to a 10^{-6} M solution of the candidate inhibitor for 2 minutes. The responses were measured for a mixture of 10^{-6} M 3kPZS and 10^{-6} M of the candidate inhibitor, and also for a mixture of 10^{-5} M L-arginine and 10^{-6} M of the candidate inhibitor. The two mixtures were recorded for 4 seconds each.

Let R_{mixture} be the response of the 3kPZS and inhibitor mixture, R_{3kpzs} the 3kPZS response before candidate inhibitor, and B the response to blank charcoal filtered fresh water. Then, the percent reduction of the 3kPZS olfactory response was calculated as

$$1 - \frac{R_{\text{mixture}} - B}{R_{\text{3kpzs}} - B} \times 100.$$

If the candidate compound either inhibited or acted as an agonist of the 3kPZS receptor (competing for the binding site with 3kPZS), a reduction of the 3kPZS signal would be observed. The response of a mixture of L-arginine and the candidate inhibitor were recorded, and the percent reduction of the L-arginine response was calculated to ensure that the inhibitor candidate had a specific effect in the 3kPZS receptor. L-Arginine is a common stimulant of the olfactory epithelium of sea lamprey that does not compete with the 3kPZS signal transduction pathway or receptor (Burns et al., 2011). Recordings for each candidate compound were repeated two to five times, and the reported signal reduction was computed as the average signal reduction among the replicates.

4.3.12 Graphics

Molecular graphics and renderings were produced using MacPyMOL v1.8.2.2 (DeLano, 2002). Data plots were generated using matplotlib (version 2.0.0; Hunter, 2007), and diagrams were drawn using vector graphics software: OmniGraffle (version 7.3.1), Affinity Designer v1.5.5, and

Autodesk Graphic (version 3.0.1). All images were exported into bitmap format using macOS Preview (version 9.0).

4.4 Results and Discussion

4.4.1 Structural model for interactions between 3kPZS and SLOR1

A 3D atomic structure of SLOR1 was built by MODELLER and energy-minimized with CHARMM, as described in the Methods. MODELLER computed 25.5% sequence identity between SLOR1 and the β1-adrenergic receptor template (PDB entry 2vt4; Warne et al., 2008), with a highly significant expectation value of $6.4e^{-11}$. Structural evaluation by PROCHECK showed 95% of the SLOR1 residues to have main-chain dihedral values in the most-favored region, comparable to high-resolution crystal structures. SwissModel Workspace evaluation tools indicated that all-atom contacts (Benkert et al., 2011) were similar in favorability to those found in the 2vt4 template structure, and the model has a favorable overall energy (Zhou & Zhou, 2002). The structural model of SLOR1 and generation of flexible conformers starting with the 3D structure of 3kPZS determined by NMR and mass spectrometry (Li et al., 2002) enabled prediction of their interaction by docking. The most favorable docking mode predicted by SLIDE, with a predicted $\Delta G_{binding}$ of -9.0 kcal/mol, showed the sulfate tail binding deeply in the orthosteric cleft, surrounded by the transmembrane helices and open to the extracellular space. The planar steroid binds in this cleft almost parallel to the transmembrane helical axes, with the specificity-determining 3-keto group pointing towards the solvent-exposed extracellular loops (Figure 4.4).

The main sulfate-binding residue, His110, is 10 Å above a regulatory sodium site elucidated in the high-resolution adenosine receptor structure and thought to occur in many class A GPCRs (PDB entry 4eiy; Liu et al., 2012). Most of the sodium-ligating side chains are identical or similar in side-chain chemistry in SLOR1 (Figure 4.5). The strongly attractive salt bridge between the 3kPZS sulfate group with both side chain nitrogen atoms on neighboring His110, reinforced by a hydrogen bond with Tyr203 and through-space electrostatic attraction with the postulated buried sodium ion (Venkatakrishnan et al., 2013), help explain the sensitivity of SLOR1 to 3kPZS in low

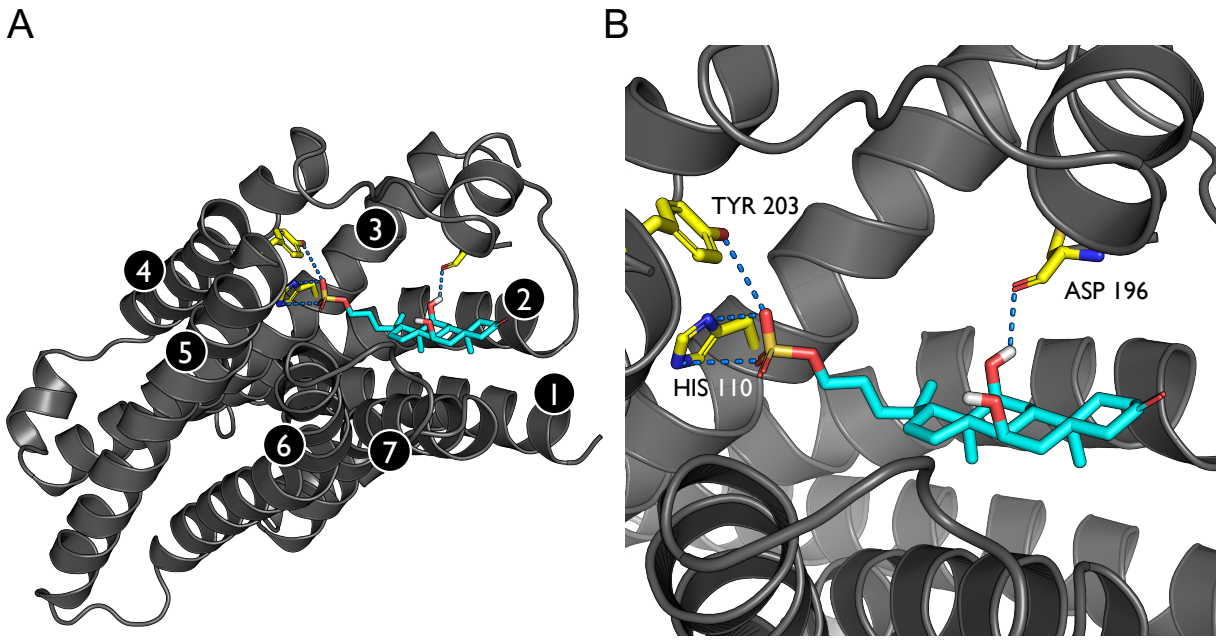


Figure 4.4: SLOR1 homology model. The backbone of the SLOR1 homology model is shown in gray ribbons with side-chain carbon atoms in yellow. 3kPZS carbon atoms are shown in cyan, with oxygen atoms in red, nitrogen atoms in blue, the sulfate sulfur in yellow, and hydrogen atoms in white. Hydrogen bonds between SLOR1 and 3kPZS are drawn as blue dashed lines. (a) Full view of the SLOR1-3kPZS complex, with GPCR transmembrane helices enumerated 1-7 from N- to C-terminus. (b) Close-up of key polar side chains forming intermolecular hydrogen bonds and salt bridges between SLOR1 and 3kPZS, as determined by SLIDE docking.

nanomolar concentration. The di-methylated face of the steroid system in 3kPZS (Figure 4.2 and Figure 4.4) is predicted to bind to a highly hydrophobic wall in the SLOR1 cleft, comprised of hydrocarbon side chain groups from Phe87, Met106, Leu109, His110, Asp196, Pro277, Tyr280, and Thr284. The 12-hydroxyl group on the opposite face of the steroid ring hydrogen-bonds with the Cys194 main-chain oxygen. This mode of interaction is supported by a very similar cholate binding prediction for SLOR1 from CholMine (http://cholmine.bmb.msu.edu; Liu et al., 2015). The position of the 3-keto group at the solvent interface, not directly contacting SLOR1, suggests it interacts with the thirty N-terminal residues of SLOR1 that are absent from the model due to lack of homology with any PDB structure. Structures of ligand-interacting lid peptides in class A GPCRs (which includes the CDYLVVLFL sequence in SLOR1), are highly individualized according to receptor type (Venkatakrishnan et al., 2013).

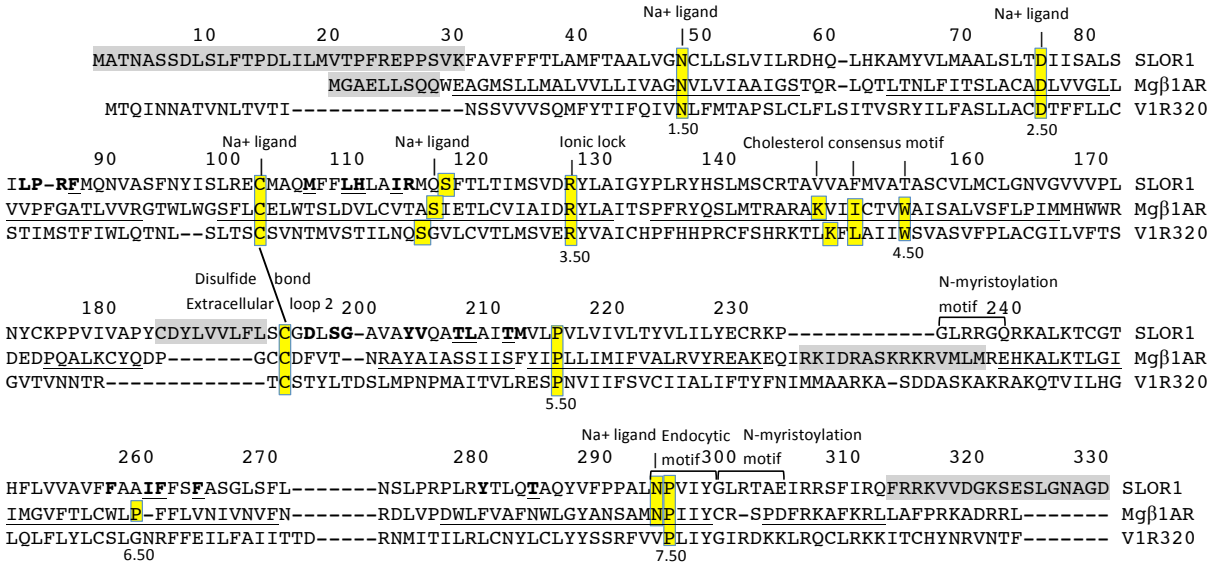


Figure 4.5: Alignment of the sea lamprey SLOR1 sequence with the closest GPCR of known 3D structure, β1-adrenergic receptor. The V1R320 receptor (Libants et al., 2009), another class A GPCR lamprey activated by 3kPZS, is shown for comparison. Yellow highlighting indicates residues associated with a cholesterol, sodium ion, N-myristoylation, or endocytic binding motif. The most highly conserved residues in GPCRs are labeled by their Ballesteros-Weinstein numbers X.50, where X is the transmembrane helix number and 50 is the position number assigned to the most conserved residue in that helix across all GPCRs (Ballesteros & Weinstein, 1995). Residue numbers for SLOR1 appear above the alignment. Boldface indicates predicted ligand binding site residues in SLOR1, based on occurring within 5 Å of the retinal ligand in rhodopsin (PDB entry 2z73; Murakami & Kouyama, 2008) or cyanopindolol in the β1-adrenergic receptor structure (PDB entry 2vt4; Warne et al., 2008), following their structural superposition on SLOR1 by DaliLite (http://ekhidna.biocenter.helsinki.fi/dali_lite). Gray highlighting indicates residues with no structural model in SLOR1 due to low homology with the β1-adrenergic receptor (Mgβ1AR) or absence of crystallographic coordinates in this region of 2vt4. Underlined residues form the transmembrane helices in PDB 2vt4, based on DSSP main-chain hydrogen-bonding analysis provided by the PDB (Kabsch & Sander, 1983).

4.4.2 Screenlamp discovery of potent 3kPZS antagonists

A typical screening run for a single hypothesis, for instance, obtaining 3kPZS volumetric and pharmacophore mimics with 3-oxygen and 24-sulfate matches (Figure 4.3) starting from 12 million commercially available, drug-like molecules in ZINC, was completed within a day on a standard desktop computer (2 Intel^{TM} Xeon^{TM} CPU E5-2620 v2 at 2.10GHz, 16 GB DDR3 SDRAM, and 7200 RPM hard drive). Candidates from the hypothesis-based screens described in the Methods provided a set of 307 commercially available compounds, including 8 samples from different vendors for some of the 299 unique compounds. The entire set was procured and tested by EOG for

the ability to reduce the 3kPZS olfactory response. Following the EOGs, the most and least active compounds were analyzed structurally to identify features that correlate with activity (Figure 4.6).

4.4.3 Structure-activity relationships of Screenlamp compounds

Seven of the 15 most active compounds, which reduced the response to 3kPZS by 41-92%, were steroidal. Interestingly, most of the top 15 inhibitors other than petromyzonol sulfate (PZS; ZINC72400307; the 3-hydroxyl analog of 3kPZS), lacked 3kPZS-like hydroxyl groups in the 7- and 12-positions of the steroid ring system. The three most active compounds had 3-hydroxyl groups in place of the 3-keto group in 3kPZS (Figure 4.7), suggesting that this group acts as a switch between agonist and inhibitor functions. In the two most active compounds, PZS and ZINC35044325, the 3-hydroxyl groups overlapped with the 3-keto group of 3kPZS following pairwise overlay (Figure 4.6).

Other interesting structure-activity relationships were revealed by the five sulfate tail analogs among the 10 most active compounds (Figure 4.7), which matched the sulfate tail moiety of 3kPZS. For instance, the three compounds ZINC01845398 (n-butylsulfate), ZINC01532179 (lauryl sulfate), and ZINC02040987 (tetradecyl sulfate) consist entirely of aliphatic hydrocarbon chains terminating in a sulfate group (Figure 4.7) and were found to reduce the olfactory response of 3kPZS by 43-45%. This is a useful insight, as it indicates that matching the 3-keto oxygen and steroid ring system in 3kPZS is not absolutely essential. Sulfated alkanes like these are inexpensive compounds, though they vary in vertebrate toxicity and are likely to be less target-selective than molecules capable of making additional 3kPZS-like interactions. The trisulfated variant of PZS (ZINC72400309) is another molecule that would not have been predicted by a typical drug discovery approach, due to its high polarity and bulk relative to the reference compound, 3kPZS. However, this turns out to be one of the most effective antagonists of 3kPZS according to in vivo EOG results, and trisulfated PZS shows even greater promise in our ongoing behavioral tests with sea lamprey in natural stream water. Another hypothesis-based structure-activity result was that matching the 3-keto and one or more sulfate oxygens in extended conformations of the 3kPZS pheromone led to high activity; five

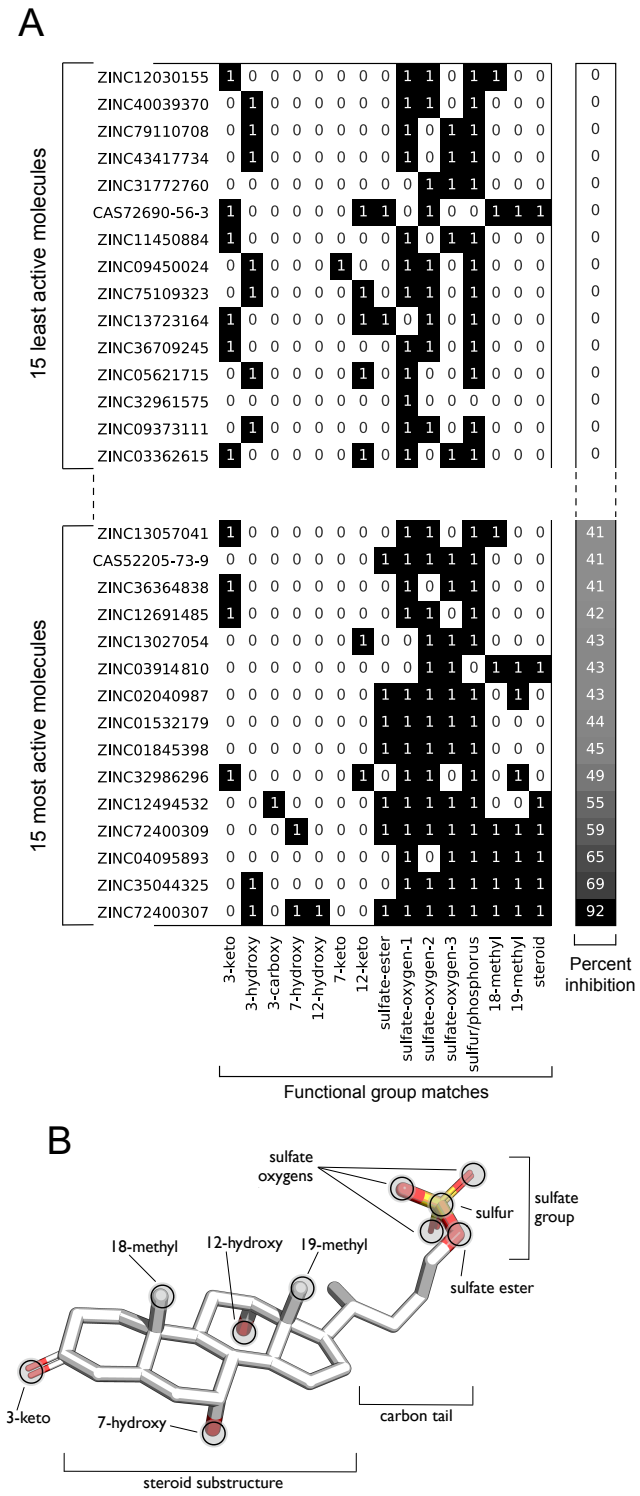


Figure 4.6: Functional group matches of the 15 most active and 15 least active molecules. (a) Heat map showing the functional group matches of the 15 most active and 15 least active molecules when overlaid with 3kPZS. The percent inhibition was computed as the average inhibition over two or more independent electro-olfactogram assays. Heat map cells containing 1's indicate the presence of a match and 0's indicate the absence of a match. (b) 3D representation of an energetically favorable 3kPZS conformer with functional group labels, to aid in interpreting the heat map x-axis labels.

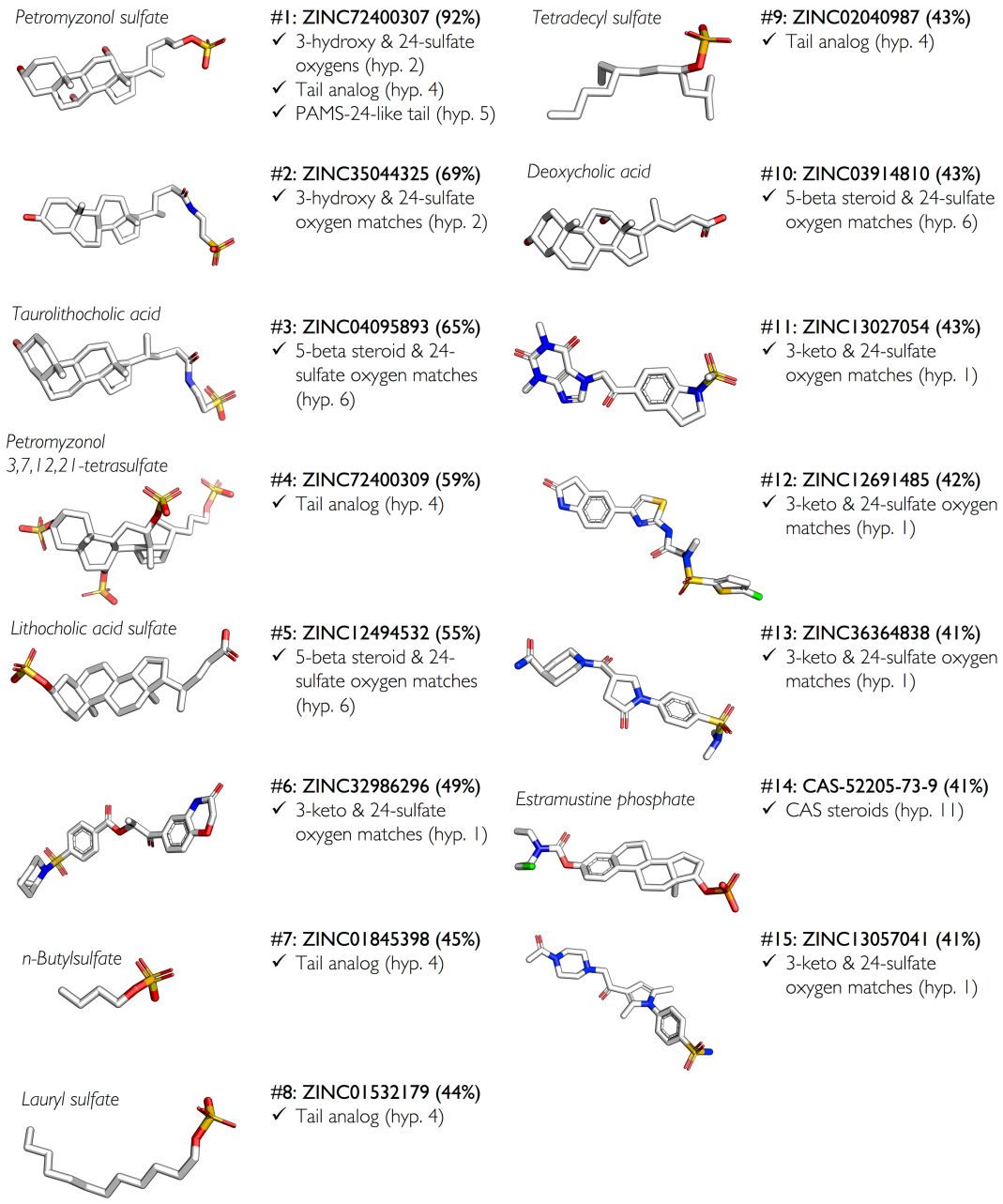


Figure 4.7: 3D structures of the 15 most active molecules. The 3D structures of the 15 most active molecules from screening the ZINC drug-like dataset, the combinatorial analog dataset, CAS registry steroids, and the GPCR Ligand Library, as described in the Methods section Preparation of millions of drug-like molecules for ligand-based screening, are shown. The molecule structures are numbered from highest percent inhibition (#1) to lowest (#15). Accession codes in the CAS registry and ZINC databases are provided along with average percent inhibition over two or more replicates. Hypothesis-based selection criteria are listed below the compound IDs, referencing the hypothesis descriptions given in the Methods. The ZINC13057041 compound has been flagged as a potential PAIN (pan-assay interference compound) containing a functional group that leads to false-positive assay results via the server available at http://cbligand.org/PAINS/ (Baell & Holloway, 2010).

of the 15 most active compounds were identified by this criteria.

4.4.4 Enrichment of active molecules through hypothesis-based filtering criteria

While shaped-based ranking methods for ligand-based virtual screening such as ROCS yield favorable results when tested for the ability to identify active molecules among a large set of decoys (Hawkins et al., 2007), the development of Screenlamp was driven by the need to include domain-based knowledge, such as spatial relationships between a subset of functional groups observed in pheromones, to discover compounds highly enriched in activity. The benefits of incorporating system-specific criteria when screening is underscored by considering the results of using only shape and charge scoring. Upon comparing the ROCS TanimotoCombo scores alone, based on shape and partial charge similarity to 3kPZS, with EOG assay values representing percent inhibition of 3kPZS response for the 299 compounds, no apparent correlation between the molecular similarity scores and percent inhibition of 3kPZS response values was observed (Figure 4.8).

This is typical in ligand-based (or protein structure-based) scoring: high similarity (or complementarity with the protein) is necessary for binding but is typically not sufficient. "It's in the details" applies to the determinants of biological activity. Nature requires molecules to form an exquisitely selective set of interactions in order to exclude the possibility of potentially lethal binding by the wrong ligands. The key is to identify which groups are making those interactions. Based on the experimental EOG data of 299 tested compounds, it is apparent that using hypothesis-driven functional group matching criteria in addition to ROCS-based similarity scoring yields greater enrichment of activity (Figure 4.9), and it also allows identification of those critical groups. For instance, while retrieving compounds matching 3kPZS with a ROCS TanimotoCombo similarity score of 1.03 or more recovered 4 of the 5 most active molecules (with at least 50% inhibition), this set of retrieved molecules also included many (161) non-active molecules (Figure 4.9b). Including additional selection criteria, such as the presence of a steroidal substructure, a sulfur or phosphorus overlay with the 24-sulfur atom in 3kPZS, and three sulfate oxygen matches, also yields 4 active molecules but only 13 inactives (Figure 4.9a). These results support that hypothesis-based chemical

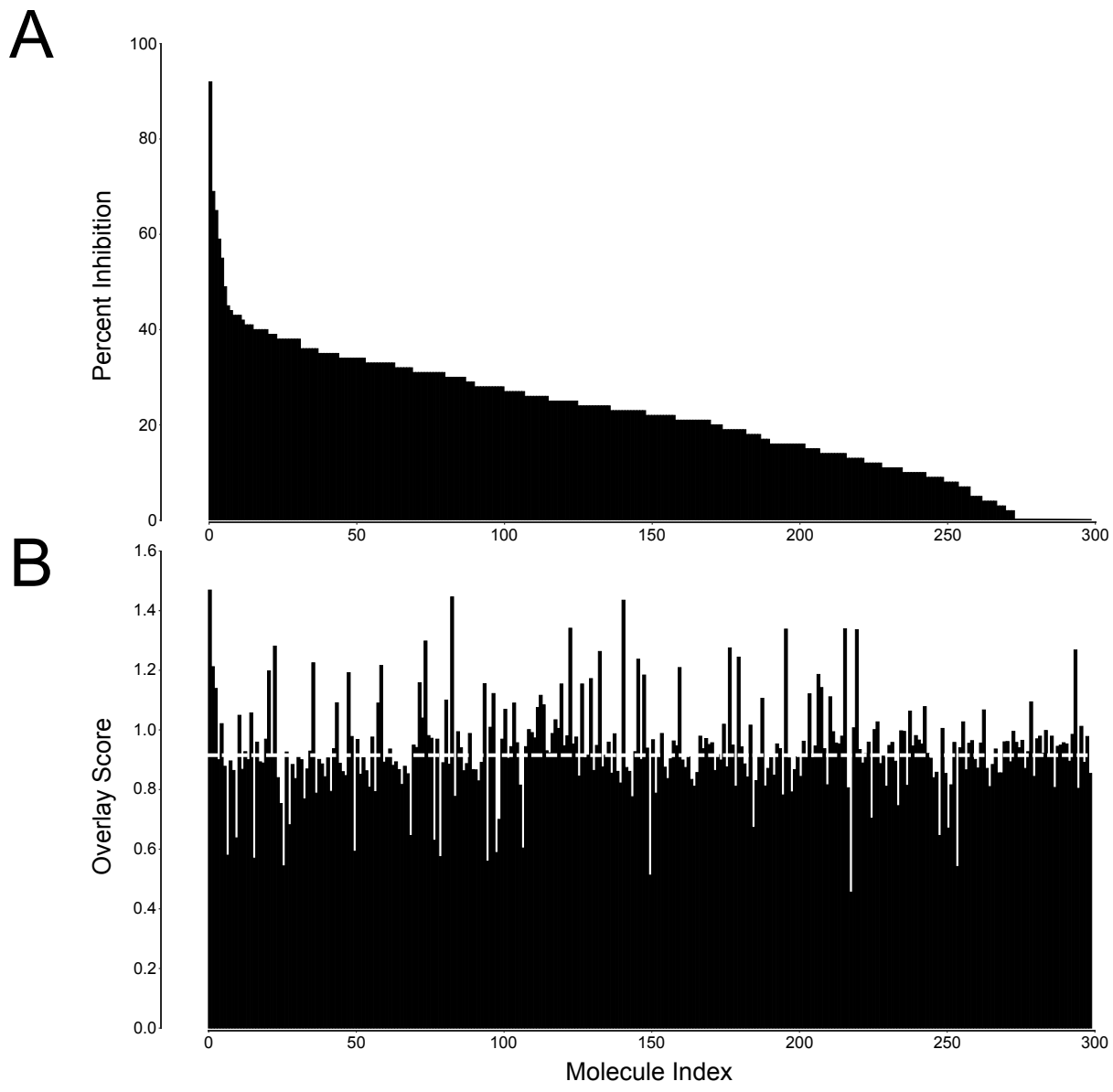


Figure 4.8: Quantitative comparison of EOG percent inhibition values for 3kPZS inhibitor candidates with their molecular similarity scores upon 3D overlay with 3kPZS. (a) The 299 assayed compounds were sorted by EOG activity values, from highest percent inhibition of 3kPZS response (left end of x-axis) to lowest (right end). (b) For these compounds shown in the same x-axis order as in (a), the ROCS TanimotoCombo molecular similarity scores following 3D flexible overlay with 3kPZS were plotted, equally weighting the electrostatic and volumetric components, with a maximum possible sum of 2.0. If the overlay similarity scores alone were highly predictive of activity, we would expect to see a pattern of high overlay scores corresponding to high percent inhibition values (that is, a similar profile of high scores decreasing to low scores, left to right, in (b) as well as (a)). However, the pattern of overlay scores in (b) is highly variable across the compounds, even for those with the highest percent inhibition values. While for most hypotheses, only compounds with reasonably high overlay scores were assayed (meaning we pre-selected for overall molecular similarity), the data in (b) shows that overlay scores alone are not enough to predict the ability of a compound to inhibit 3kPZS activity. This drove the development of the tools in Screenlamp for identifying functional group patterns associated with biological activity (hypothesis-driven screening).

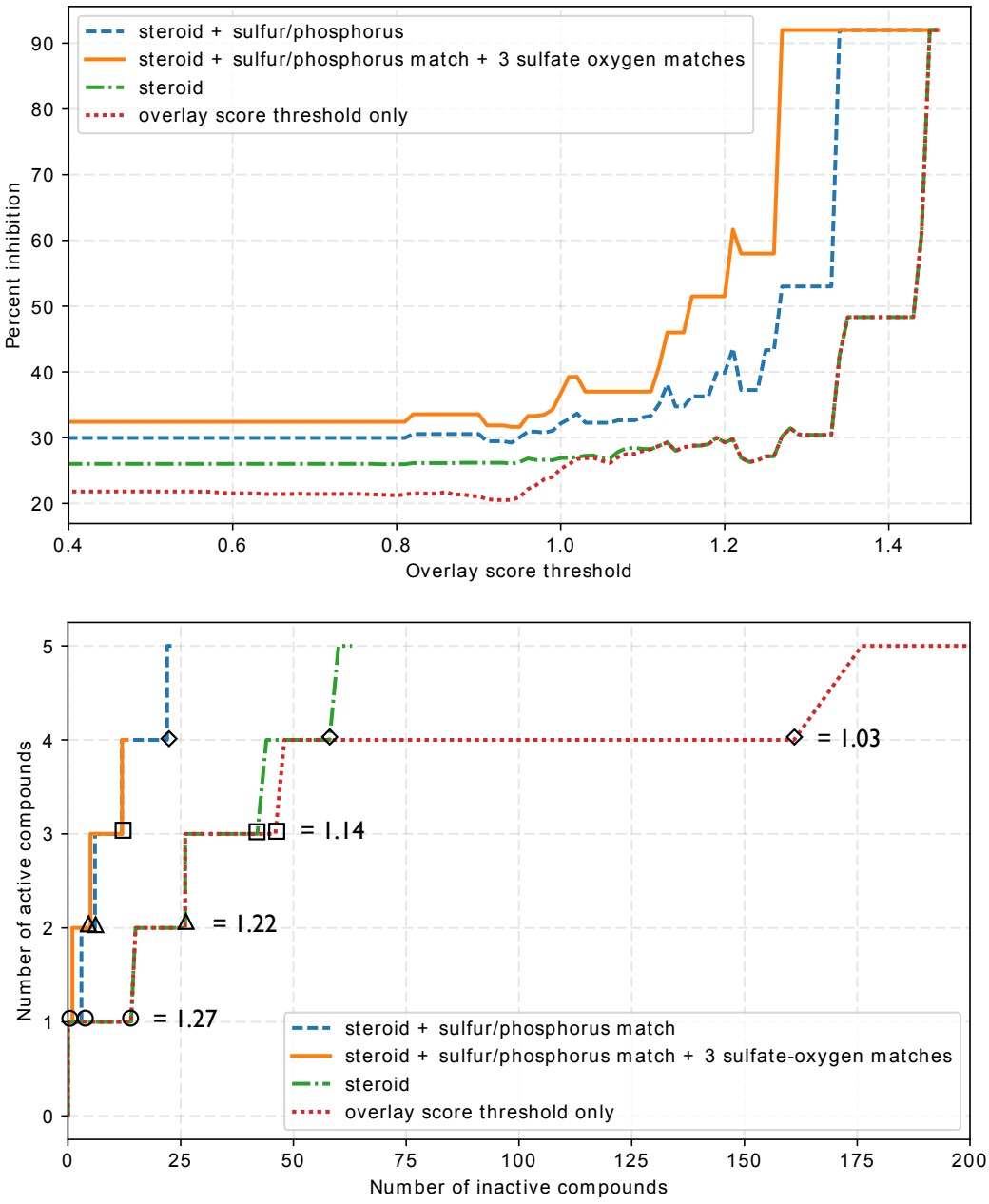


Figure 4.9: Enrichment curves and receiver operating characteristic comparing the performance via hypothesis selections criteria to ligand-based overlay score selections. (a) Enrichment graphs showing the percent inhibition as a function of chemical and volumetric similarity of candidate molecules to 3kPZS, based on inhibition assays for 299 molecules selected by structure-activity hypotheses (as described in the text). The overlay score threshold refers to the ROCS TanimotoCombo score, equally weighting volumetric and electrostatic similarity.

Figure 4.9 (cont'd). The different traces on the graphs compare the enrichment for several hypotheses relative to using 3D similarity (overlay score) alone. (b) Receiver operating characteristic curve (rate of retrieval of true positives vs. false positives), with triangles, square and circle symbols used to show the point on each curve corresponding to a given score threshold, for cases in which the overlay score alone was used to select candidate compounds, versus when the overlay score was augmented by increasingly selective steroid-based hypotheses, resulting in fewer molecules being tested but a greater enrichment in active compounds. The curves show the number of active molecules (at least 50 percent inhibition of 3kPZS in experimental assays) for different overlay thresholds.

group filtering criteria, as facilitated by Screenlamp, not only decreases computational costs by appropriately reducing the chemical search space, but also is invaluable for increasing the rate of retrieval of active compounds.

4.5 Conclusions

Incorporating a hypothesis-driven strategy in computational and organismal biology has identified an inhibitor which in low concentration virtually nullifies the olfactory response of sea lamprey to the major mating pheromone, 3kPZS. Other highly active compounds were identified and are showing great promise as mating behavioral deterrents in ongoing stream trials. The ligand-based screening with Screenlamp also provided a series of simpler, nonsteroidal compounds that are significantly active and provide useful structure-activity information. To our knowledge, this presents the first successful application of structure-based drug discovery techniques to identify potent lead compounds for aquatic invasive species control. To enable other projects to benefit from this scalable, hypothesis-driven strategy which works easily with very large datasets, we have documented and are distributing the Screenlamp toolkit free of charge (https://github.com/psa-lab/screenlamp; see *Results* section on *Development of Screenlamp* for details).

CHAPTER 5

AUTOMATED INFERENCE OF CHEMICAL DISCRIMINANTS OF BIOLOGICAL ACTIVITY

Adapted with permission from Raschka, Sebastian, Anne M. Scott, Mar Huertas, Weiming Li & Leslie A. Kuhn. 2017. "Automated Inference of Chemical Discriminants of Biological Activity." *Methods in Molecular Biology: Computational Drug Discovery and Design (M. Gore, ed.), Springer Protocols. In Press*.

5.1 Abstract

Ligand-based virtual screening has become a standard technique for the efficient discovery of bioactive small molecules. Following assays to determine the activity of compounds selected by virtual screening, or other approaches in which dozens to thousands of molecules have been tested, machine learning techniques make it straightforward to discover the patterns of chemical groups that correlate with the desired biological activity. Defining the chemical features that generate activity can be used to guide the selection of molecules for subsequent rounds of screening and assaying, as well as help design new, more active molecules for organic synthesis.

The quantitative structure-activity relationship machine learning protocols we describe here, using decision trees, random forests, and sequential feature selection, take as input the chemical structure of a single, known active small molecule (for example, an inhibitor, agonist, or substrate) for comparison with the structure of each tested molecule. Knowledge of the atomic structure of the protein target and its interactions with the active compound are not required. These protocols can be modified and applied to any data set that consists of a series of measured structural, chemical, or other features for each tested molecule, along with the experimentally measured value of the response variable you would like to predict or optimize for your project, for instance, inhibitory activity in a biological assay or $\Delta G_{binding}$. To illustrate the use of different machine learning algorithms, we step through the analysis of a dataset of inhibitor candidates from virtual screening that were tested recently for their ability to inhibit GPCR-mediated signaling in a vertebrate.

5.2 Introduction

In this chapter, we will apply machine learning to analyze the results from a virtual screening (VS) project for discovering inhibitors of GPCR signaling in a vertebrate, to infer the importance of functional groups for their biological activity. Computer-based ligand screening, also known as ligand-based screening, is frequently used in pharmaceutical discovery because it performs robustly in identifying active molecules from the top-scoring set and does not require the availability of an atomic structure of the protein target (Ripphausen et al., 2011; Geppert et al., 2010). Further, it has

been shown that ligand-based virtual screening is capable of exploring different active scaffolds, making it a valuable alternative to structure-based methods such as molecular docking, even when atomic structures of the target are known (Pérez-Nueno et al., 2008; Hawkins et al., 2007).

However, scientists typically focus on the most active handful of compounds and test their closest analogs while not making use of the activity data available from all the tested compounds to identify correlations between their chemical groups and activity values. Part of this may be due to the need to establish spatial correspondences between chemical groups in compounds containing different molecular scaffolds (for example, comparing substituents on a steroid ring system versus a purine nucleotide). This problem has been circumvented in the protocols presented here by considering all molecules as fully flexible 3D structures and determining their optimal overlay based on the volumes and partial charges of the atoms, followed by comparing the chemical identities of neighboring atoms and small organic groups such as $–NH_2$. We will use the term "functional groups" to refer to single or small groups of atoms that are being compared between molecules. This flexible overlay procedure provides a rational and quantitative way of comparing chemical groups between compounds.

5.2.1 Discovering biologically active molecules through virtual screening

The most prominent approaches in the computer-aided discovery of biologically active molecules are *structure-based* screening (Sukuru et al., 2006; Lyne, 2002; Ghosh et al., 2006; Li & Shah, 2017) and *ligand-based* screening (Ripphausen et al., 2011; Geppert et al., 2010; Yan et al., 2016; Raschka et al., 2017) as well as hybrids thereof (Zavodszky et al., 2009; Buhrow et al., 2013). Traditionally, structure-based screening is restricted to applications where an experimentally determined, high-resolution three-dimensional (3D) structure of the ligand's binding partner (usually a protein or nucleic acid) is available from X-ray crystallography or nuclear magnetic resonance experiments.

While ligand-based screening does not require knowledge of the binding target, it assumes that active molecules are likely to share shape and chemical similarities with a known, biologically active ligand. In short, ligand-based screening can be described as a similarity search between a

known ligand and the molecules in a database (Figure 5.1).

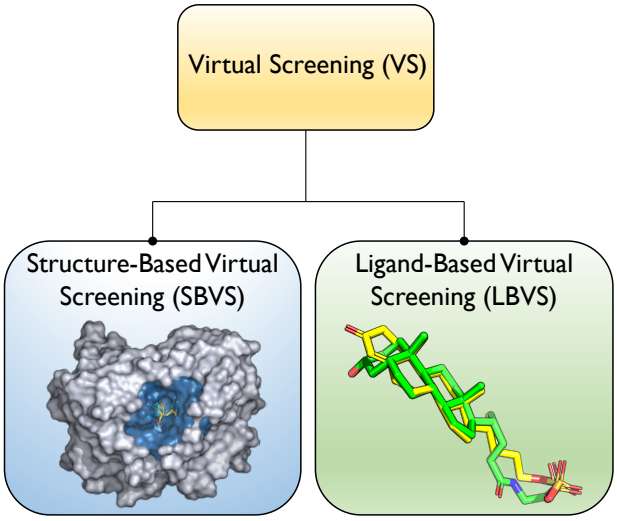


Figure 5.1: An illustration of the two broad categories of virtual screening. Structure-based virtual screening involves docking into a binding site to maximize protein-ligand surface complementarity, and ligand-based virtual screening involves evaluating small-molecule similarity with a known ligand.

5.2.2 Using machine learning to identify functional groups associated with biological activity

To guide virtual screening, understand biological mechanisms, and aid the design of more potent inhibitors or activators of molecular processes, several different techniques have been developed to analyze datasets of molecular descriptors and measured activity. A common goal in quantitative structure-activity relationship (QSAR) modeling includes the prediction of the *in vitro* or *in vivo* activity of molecules given their features. Another common goal is to gain insights into the importance of individual functional groups for binding or chemical activity; such insights are invaluable for the discovery and optimization of potent agonists or inhibitors. More detailed discussions of QSAR can be found in Kubinyi et al., 2006 and Verma et al., 2010.

To infer which functional groups are most important for biological activity, this chapter focuses on the use of supervised machine learning algorithms to discover functional group matching patterns that explain the relative activity of the tested inhibitor candidates. Primarily, the analysis of the discriminants of biological activity presented here employs tree-based machine learning algorithms. A decision tree (Breiman et al., 1984) that separates active from non-active molecules provides a

model that is readily interpretable, resulting in a set of decision rules that if chained together, can explain the hierarchy of features in a molecule that are most important for distinguishing actives from non-actives. Secondly, multiple decision trees will be combined via the random forest method (Breiman, 2001). Each decision tree in a random forest is fit to a random sample of the training data and feature set. This produces an ensemble of different decision trees, which together provide a robust predictive model that is less prone to overfitting the training data than any individual decision tree (Breiman, 2001). Furthermore, a random forest facilitates the computation of feature importance as the average information gain over the individual trees, as it will be explained in more detail in the Methods section. Lastly, we will utilize an implementation of sequential backward selection, a sequential feature selection algorithm that identifies subsets of features to maximize the performance of a given model in a greedy (fastest improvement, rather than exhaustive) fashion (Ferri et al., 1994; Raschka, 2017b). Sequential feature selection algorithms can be combined with any machine learning algorithm, and hence, they provide a flexible, model-agnostic solution for the analysis of combinations of functional groups that explain biological activity.

5.2.3 Predicting the essential features of GPCR inhibitors: a real-world case study

This chapter presents an automated, machine learning-based approach to infer the discriminants of activity in molecules from assays performed on compounds prioritized by ligand-based screening. To explain the methodology behind this approach, we will consider a novel dataset of 56 molecules that have been prioritized as candidates for inhibiting GPCR-mediated pheromone signaling in an invasive species control project. Readers can access the same data and software and then perform the same analyses and compare their results with ours.

The goal of this invasive species control project is to inhibit a pheromone-induced GPCR olfactory signaling pathway. We hypothesized that the inhibition of pheromone detection by the olfactory system will prevent mature female sea lamprey from reaching mature males at spawning grounds in tributaries of the Laurentian Great Lakes and thus reduce the invasive sea lamprey population. Controlling the sea lamprey with pesticide applications currently costs millions of

dollars per year, with native fish populations and commercial fishing continuing to be impacted by sea lamprey parasitism (Hansen & Jones, 2008). The rationale behind the screening side of this project is based on a recently completed project (Raschka et al., 2017), focusing on inhibiting the GPCR signaling pathway induced by a male sea lamprey mating pheromone, 7α,12α,24-trihydroxy-5α-cholan-3-one-24-sulfate (3kPZS).

The dataset analyzed here consists of the chemical structures and assay data for another male sea lamprey mating pheromone, the sulfate-conjugated bile alcohol, 3,12-diketo-4,6-petromyzonene-24-sulfate (DKPES, Figure 5.2). The 56 molecules prioritized by ligand-based screening according to their degree of 3D DKPES similarity were assayed for their ability to block the *in vivo* sea lamprey olfactory response to DKPES, as measured by an electro-olfactogram assay (EOG). The activity data was then analyzed using machine learning algorithms to uncover structure-function patterns. A brief summary of the virtual screening approach that we used to identify inhibitory mimics of

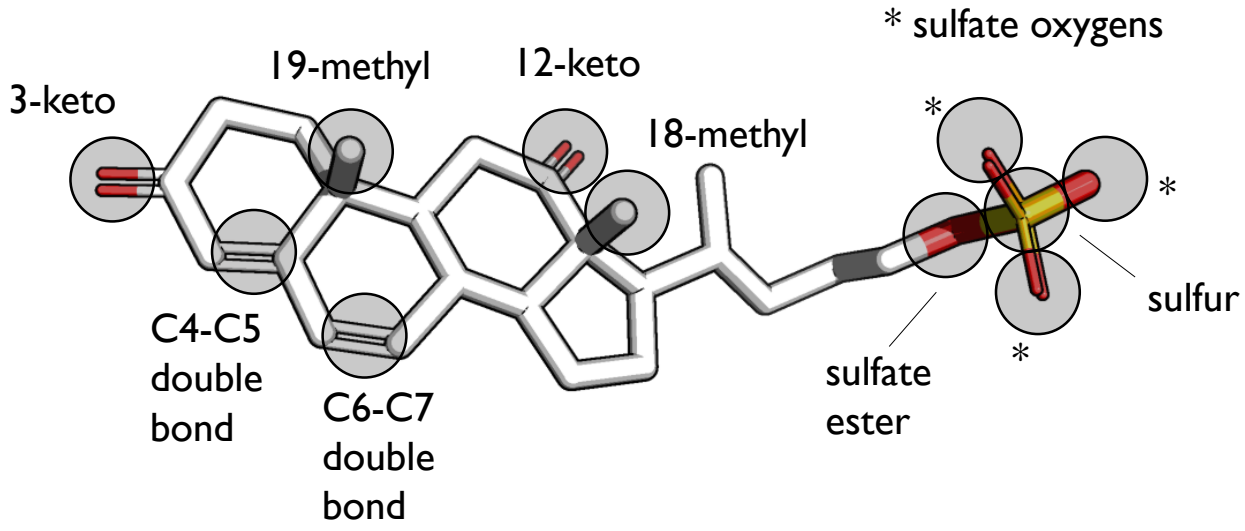


Figure 5.2: 3D structure of a favorable (low-energy) DKPES conformer. The functional group features corresponding to the columns in the olfactory response dataset are highlighted with gray circles. White indicates carbon, red indicates oxygen, and yellow indicates sulfur.

DKPES is provided in Figure 5.3.

As a result, 56 candidate molecules were prioritized for biological assays based on the following criteria:

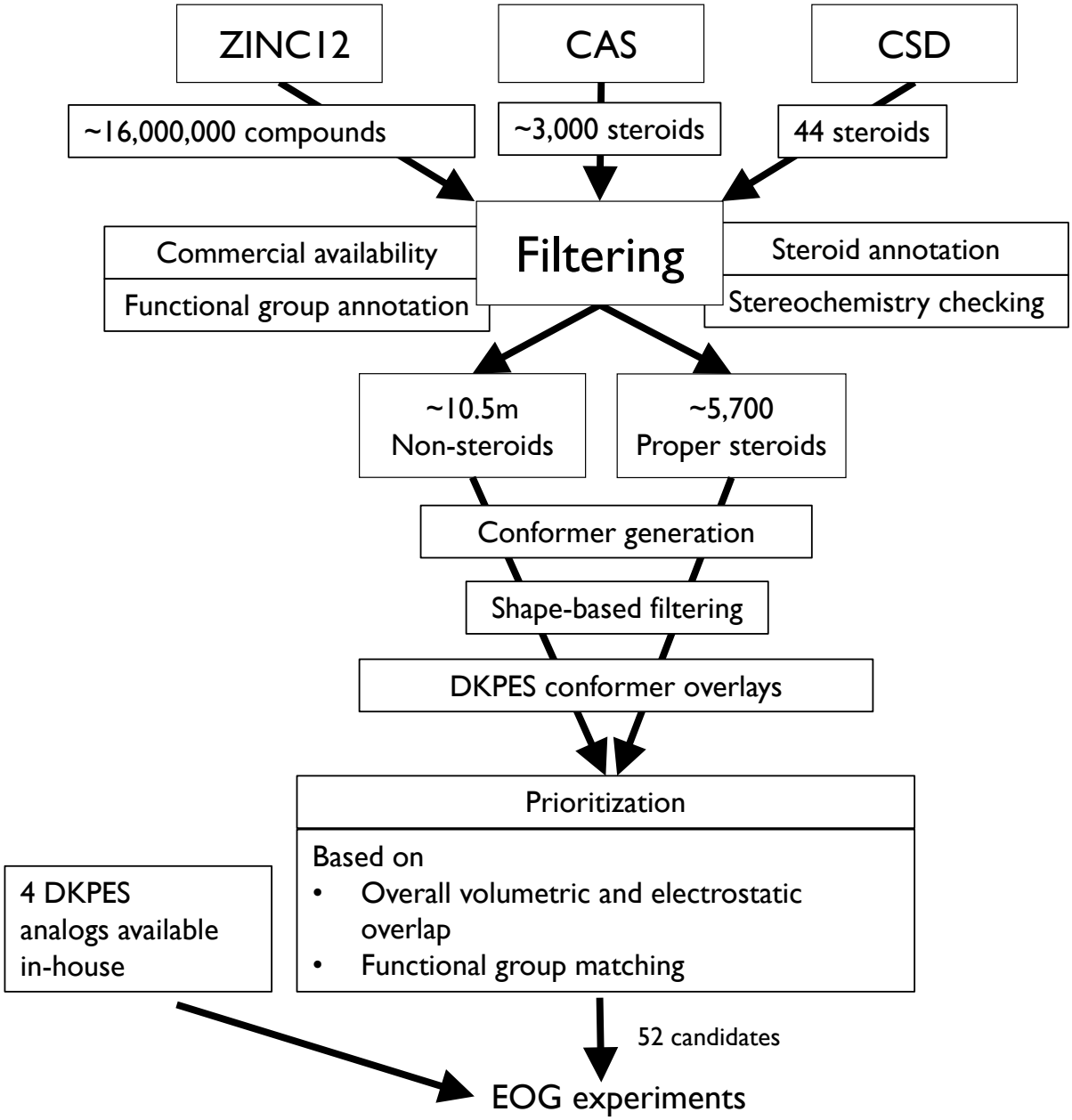


Figure 5.3: Summary of the virtual screening workflow to prioritize molecules for electro-olfactogram (EOG) assays. The Screenlamp toolkit (https://github.com/psa-lab/screenlamp) was used to prepare the virtual screening pipeline, including OpenEye OMEGA and ROCS (https://www.eyesopen.com). The screening databases of small molecules, mostly commercially available, were the drug-like molecules in ZINC12 (http://zinc.docking.org; Irwin & Shoichet, 2005), steroid structures from Chemical Abstracts Service Registry (CAS; https://www.cas.org), and steroid structures from the Cambridge Structural Database (CSD; https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/; Allen, 2002).

- Steroidal substructure containing molecules with a high degree of shape and charge match to the reference pheromone, DKPES (Figure 5.2).
- Four DKPES analogs with oxidized (double bond-containing) rings that are naturally produced by mature male sea lamprey (molecule IDs: ENE 1-4, Figure 5.4; Johnson et al., 2014).
- Diverse compounds to test the hypothesis that compounds with the best charge and shape match will mimic DKPES (without requiring a steroid core or sulfate tail match).
- Non-steroid compounds having 3-keto or 3-hydroxy and 12-keto or 12-hydroxy matches and at least one sulfate oxygen match that overlay on the corresponding oxygen atoms in DKPES (Figure 5.2).

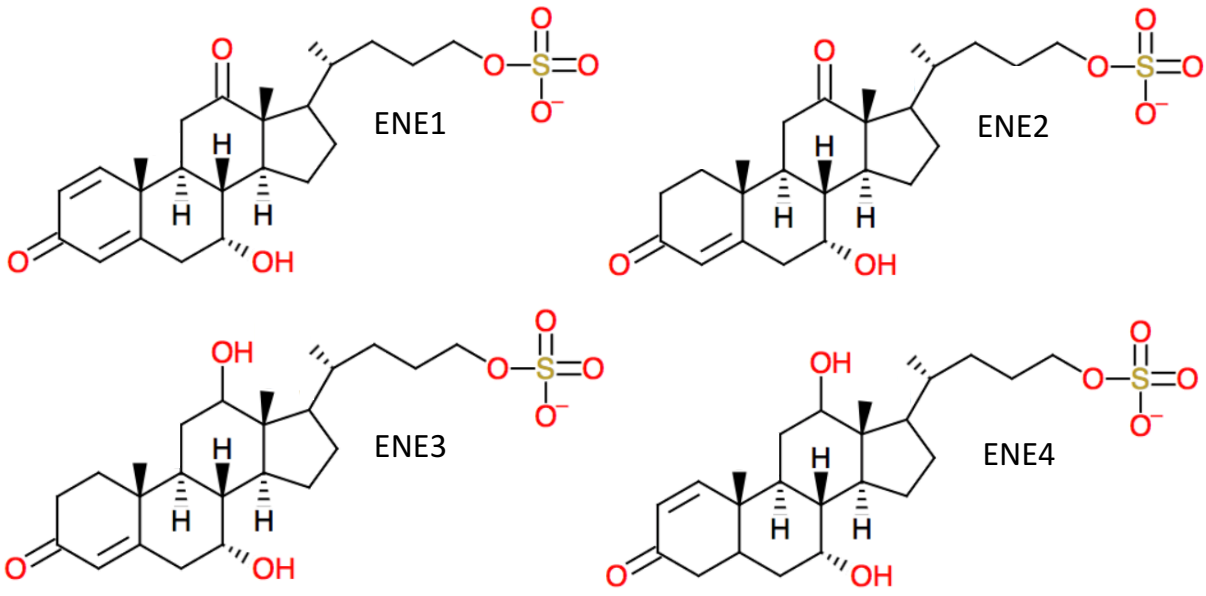


Figure 5.4: 2D structures of the 4 combinatorial DKPES analogs ("ENE" compounds). ENE1: 7,24-dihydroxy-3,12-diketo-1,4-choladiene-24-sulfate; ENE2: 7,24-dihydroxy-3,12-diketo-4-cholene-24-sulfate; ENE3: 7,12,24-trihydroxy-3-keto-4-cholene-24-sulfate; ENE4: 7,12,24-trihydroxy-3-keto-1-cholene-24-sulfate.

To measure the biological activity of the 56 DKPES inhibitor candidates selected with the above screening and prioritization criteria, we used an electro-olfactogram (EOG) as described in

Raschka et al., 2017. The measured EOG response, acting as the target variable in this dataset was the percentage reduction of the standard DKPES signal when the sea lamprey nose was perfused with a known concentration (10^{-6} M) of inhibitor candidate (computed as the average of 2 to 5 experimental replicates). Figure 5.5 shows four of the 56 molecules for illustrative purposes, two actives and two non-actives, with the percent DKPES olfactory inhibition for each. In the context of this project, *non-actives* were defined as molecules that block the olfactory response by less than 40% in EOG assays, and molecules that block the signaling response by at least 60% were defined as *actives*. It shall be noted that a similar workflow can be used to model continuous response data. However, in our experience, working with continuous target data can often lead to noisier, less interpretable results.

The DKPES dataset for analysis by machine learning contains the ROCS overlay scores from ligand-based screening (Figure 5.3) as well as the functional group matching information provided by Screenlamp in tabular form (https://github.com/psa-lab/predicting-activity-by-machine-learning; Raschka et al., 2017).

Using the DKPES dataset as a case study, the Methods section will explain how to work with such tabular datasets consisting of samples and molecular features using open source libraries for data parsing, visualization, and machine learning. The code and data used in the following section is freely available at https://github.com/psa-lab/predicting-activity-by-machine-learning.

5.3 Materials

5.3.1 Python interpreter

To perform the analyses described in the Methods section, a recent Python (Van Rossum, 2007) version (3.5 or newer) is required (Python 3.6 is recommended). A Python installer for all major operating systems (macOS, Windows, and Linux) can be downloaded from https://www.python.org/downloads/.

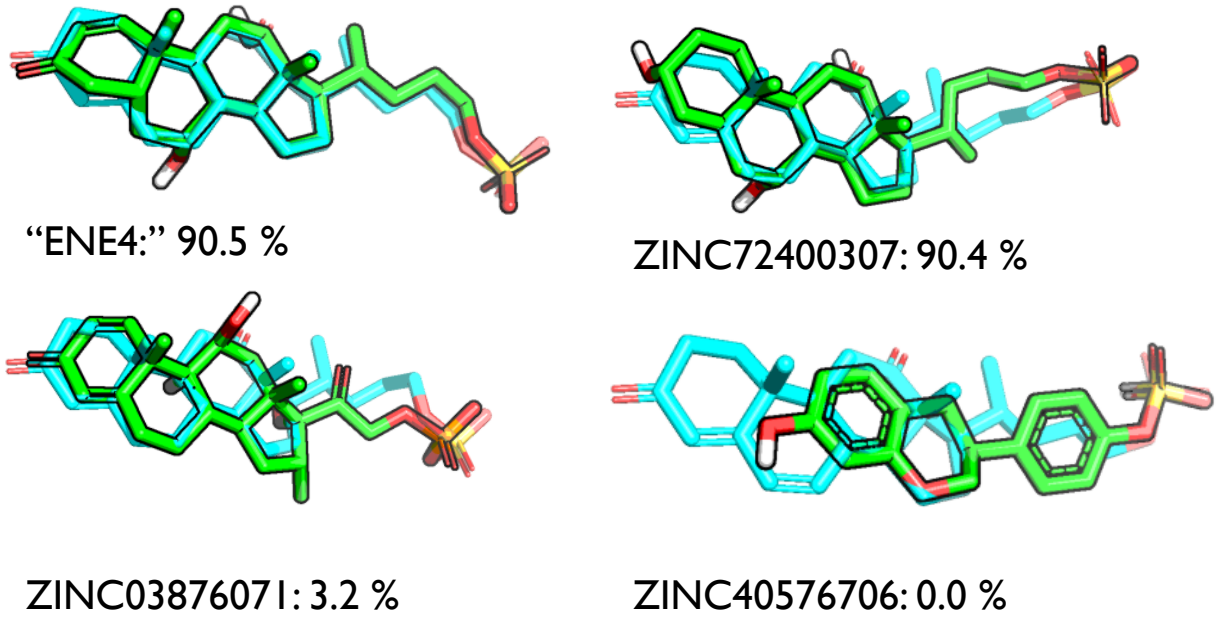


Figure 5.5: 3D structures and percent DKPES olfactory inhibition of the two most active and inactive molecules. The two most active molecules (actives, top row) and two low-activity molecules (non-actives, bottom row) from the screening set are shown in green as overlayed with the best-matching DKPES 3D conformer (cyan).

5.3.2 Python libraries for scientific computing

The following list specifies the Python libraries used in this chapter, the recommended version number, and a short description of their use:

- NumPy version 1.13.0 or newer (http://www.numpy.org); numerical array library (Van Der Walt et al., 2011)
- SciPy version 0.19.0 or newer (https://www.scipy.org); advanced functions for scientific computing (Jones et al., 2001)
- Pandas version 0.20.1 or newer (http://pandas.pydata.org); handling of CSV files and working with data frames (McKinney, 2010)
- Matplotlib version 2.0.2 or newer (https://matplotlib.org); 2D plotting (Hunter, 2007)

- Scikit-learn version 0.18.1 or newer (http://scikit-learn.org/stable/); algorithms for machine learning (Pedregosa et al., 2011)
- MLxtend version 0.7.0 or newer (http://rasbt.github.io/mlxtend/); sequential feature selection algorithms (Raschka, 2017b)

The scientific computing libraries listed above can be installed using Python's in-built Pip module (https://pypi.python.org/pypi/pip) by executing the following line of code directly from a macOS/Unix, Linux, or Windows MS-DOS terminal command line:

```
pip install numpy scipy pandas \
  matplotlib scikit-learn pydotplus mlxtend
```

If you encounter problems with version incompatibilities, you can specify the package versions explicitly, as shown in the following terminal command example:

```
pip install numpy==1.13.0 scipy==0.19.0 pandas==0.20.1 \
  matplotlib==2.0.2 scikit-learn==0.18.1 \
  pydotplus==2.0.2 mlxtend==0.7.0
```

5.3.3 Graph visualization software

To visualize the decision trees later in this chapter, an installation of GraphViz is needed. The GraphViz downloader is freely available at http://www.graphviz.org with the installation and setup instructions.

5.3.4 Dataset

The datasets being used in this chapter, as well as the source files of all the accompanying code, are available online under a permissive open source license at https://github.com/psa-lab/predicting-activity-by-machine-learning.

5.3.5 Additional resources

If you are unfamiliar with Python and the Python libraries that you installed in section *2.2. Python Libraries for Scientific Computing*, it is highly recommended to familiarize yourself with their basic functionality by reading these freely available resources:

- *Python Beginner Guide*: https://wiki.python.org/moin/BeginnersGuide
- *NumPy Quickstart Tutorial*: https://docs.scipy.org/doc/numpy-dev/user/quickstart.html
- *Introduction to NumPy*: https://sebastianraschka.com/pdf/books/dlb/appendix_f_numpy-intro.pdf
- *10 Minutes to pandas*: http://pandas.pydata.org/pandas-docs/stable/10min.html
- *Matplotlib Tutorials*: https://matplotlib.org/users/index.html
- *An introduction to machine learning using scikit-learn*: http://scikit-learn.org/stable/tutorial/basic/tutorial.html

5.4 Methods

This section walks through the individual steps involved in a typical analysis pipeline for identifying which functional groups and atoms (or other molecular properties or *features*) are predictive of the measured biological activity of the molecules. The first section explains how the tabular DKPES dataset can be loaded into a Python session for analysis.

5.4.1 Loading and inspecting the biological activity dataset

This section explains how to load a CSV-formatted dataset table into a current Python session. A convenient way to parse a dataset from a tabular plaintext format, such as CSV, is to use the `read_csv` function from the Pandas library as shown in the code example in Figure 5.6, which loads the DKPES dataset into a Pandas DataFrame object (`df`) for further processing.

For this section, we used a CSV file where the features and target variable (signal inhibition) were stored as columns separated by commas. Note that the `read_csv` function does not strictly require this input format. For instance, pandas's `read_csv` function supports any possible column delimiter (for example, tabs, whitespaces, and so forth), which can be specified via the `delimiter` function parameter. For more information about the `read_csv` function, please refer to the official documentation at https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html. Furthermore, if you are planning to work with datasets where the features are stored as rows as opposed to columns, you can use the transpose method (`df = df.transpose()`) after loading a dataset to transpose the data frame index and columns.

```
>>> import pandas as pd
>>> df = pd.read_csv('../data/csvs/dkpes.csv')
>>> df.head(10)
```

	index	Signal-inhibition	3-Keto	3-Hydroxy	12-Keto	12-Hydroxy	19-Methyl	18-Methyl	Sulfate-Ester	Sulfate-Oxygens	...
0	ENE4	0.905	1	0	0	1	1	1	1	3	...
1	ZINC72400307	0.904	0	0	0	1	1	1	1	3	...
2	ENE3	0.897	1	0	0	1	1	1	1	3	...
3	ENE1	0.893	1	0	1	0	1	1	1	3	...
4	ENE2	0.845	1	0	1	0	1	1	1	3	...
5	ZINC12494532	0.741	0	0	0	0	0	0	1	3	...
6	ZINC35044325	0.739	0	1	0	0	1	1	0	3	...
7	ZINC04095893	0.722	0	0	0	0	1	1	0	3	...
8	ZINC01532179	0.686	0	0	0	0	0	0	1	3	...
9	ZINC70666191	0.627	0	0	0	1	1	1	0	0	...

Figure 5.6: Code for reading in the DKPES dataset into a data frame. The characters >>> denote a Python interpreter prompting for a command to enter and execute. The table resulting from the execution of this code example (`df.head(10)`) shows an excerpt from the DKPES data table sorted by signal inhibition: the 10 most active molecules from the EOG experiments and their functional group matching patterns.

As a result from executing the code shown in Figure 5.6, the `df.head(10)` call will display the first ten rows in the dataset, to confirm that the data file has been parsed correctly. The DKPES dataset consists of 56 rows, where each row stores the functional group matching information for

an assayed molecule with the reference molecule DKPES.

Please note that this work assumes that a tabular dataset containing information of the molecules as well as the assay response have already been collected in tabular form. However, the analysis approach outlined in this chapter is a general one, and it is not restricted to the specific functional group matching patterns shown in Figure 5.6. For more information on how this functional group matching data can be generated from a ligand-based screening, see (https://github.com/psa-lab/screenlamp; Raschka et al., 2017).

Further, throughout the Methods section, we assume that the data frame of activity data was already sorted by signal inhibition in decreasing order. While sorting the data frame is not essential for fitting the machine learning models in the later section, you may consider sorting your datasets for the heat map visualization, to show the 10 molecules with the highest inhibition activity, for example. To sort the data frame df, you can use `sort_values` method of a given pandas data frame object. For example, the following code sorts the molecules stored as a data frame `df` from most active to least active: `df = df.sort_values('Signal-Inhibition', ascending=False)`. More information about this `sort_values` method can be found in the official pandas documentation at https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.sort_values.html.

The first column of the DKPES data table (Figure 5.6), "index," numbers each molecule. The "Signal Inhibition" column contains the response variable measured by the biological assay, in this case ranging from 0 (non-active) to 1 (highly active, with 100% DKPES signal inhibition). For instance, we can see from the table (Figure 5.6) that ENE4 and ZINC72400307 (petromyzonal sulfate) were the most promising candidate inhibitors, as they reduced the olfactory response to DKPES by 90.5% and 90.4%, respectively, when each inhibitor candidate was used at the same equimolar concentration (10^{-6} M) as DKPES. The consequent columns, labeled as 3-Keto, 3-Hydroxy, and so forth, contain information about whether an atom or functional group in the candidate molecule overlayed (within 1.3 Å) with the same group in DKPES. This functional group matching data is stored as a binary variable, where 0 indicates "no overlay" and 1 indicates "overlay." In addition, the ROCS shape and charge ("color") overlay scores were appended to

the dataset. For information on how the overlay scores are computed, the reader is referred to https://docs.eyesopen.com/rocs/shape_theory.html and Hawkins et al. (Hawkins et al., 2007); however, it shall be noted that the ROCS scoring data is not essential for this analysis.

While we recommend working with 3D structures because they provide spatial relationships between chemical groups, molecular features can also be derived from 1D string representations of molecules or 2D structural representations. For example, the presence of certain substructures or atom types, using so-called molecular fingerprints, can be computed using the open-source toolkit OpenBabel (https://openbabel.org/docs/dev/Fingerprints/intro.html). To convert a 1D or 2D representation of a molecule into a 3D structure as input for the spatial functional group matching in the DKPES dataset that was done via Screenlamp (Raschka et al., 2017) using ROCS overlays (OpenEye Scientific Software, Santa Fe, NM; https://www.eyesopen.com/rocs), you may find the following tools helpful:

- The CACTUS online SMILES translator and structure file generator (https://cactus.nci.nih.gov/translate/).
- OMEGA (OpenEye Scientific Software, Santa Fe, NM; https://www.eyesopen.com/omega), which creates multiple favorable 3D conformers of a given structure from 1D, 2D, or 3D representations (Hawkins et al., 2010; Hawkins & Nicholls, 2012). This software is available free for academic researchers upon completion of a license agreement with OpenEye.

Further, you may find the BioPandas toolkit helpful (http://rasbt.github.io/biopandas/; Raschka, 2017a), which reads 3D structures from the common MOL2 file format into the pandas data frame format. This allows users can be useful if you are working with large MOL2 databases that contain thousands or millions of structures that you want to filter for certain properties prior to generating overlays via ROCS or compute the functional group matching patterns via Screenlamp (https://github.com/psa-lab/screenlamp).

It is always helpful to perform exploratory analyses when working with a new dataset. The following code snippet shown in Figure 5.7 will generate the histogram of the signal inhibition

values shown, plus the 2D scatter plot comparing the signal inhibition values with the molecular similarity measured in the overlays. First, the signal inhibition data from the data frame (`df`) is assigned to a variable y, and the functional group columns of interest to a variable X. Next, the code in Figure 5.7 demonstrates how to use matplotlib to create two sublots, `ax[0]` and `ax[1]`, showing a histogram of the signal inhibition and a scatter plot of the signal inhibition versus molecular similarity side by side. (If you are new or unfamiliar with the matplotlib syntax, it is recommended to consult the tutorials and resources listed in Section 5.3.5 Additional Resources.)

```
>>> y = df['Signal-inhibition'].values
>>> fgroup_cols = ['3-Keto', '3-Hydroxy', '12-Keto',
...                '12-Hydroxy', '19-Methyl', '18-Methyl', 'Sulfate-Ester',
...                'Sulfate-Oxygens', 'C4-C5-DB', 'C6-C7-DB', 'Sulfur']
>>> X = df[fgroup_cols].values
>>> fig, ax = plt.subplots(1, 2, figsize=(10, 3))
>>> ax[0].hist(y, bins=np.arange(0, 1.1, 0.1))
>>> ax[0].set_ylabel('Molecule count')
>>> ax[0].set_xlabel('Signal inhibition')
>>> ax[1].scatter(y, df['TanimotoCombo'].values)
>>> ax[1].set_ylabel('Tanimoto similarity')
>>> ax[1].set_xlabel('Signal inhibition')
>>> plt.show()
```

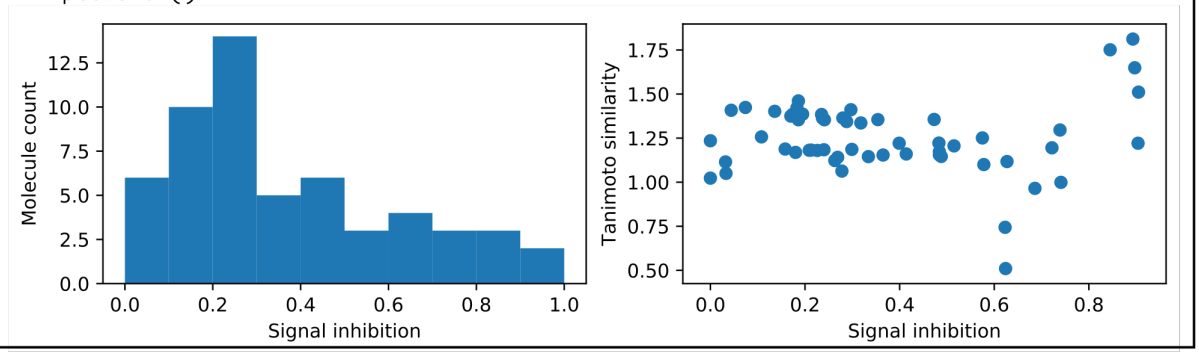


Figure 5.7: Code for performing exploratory analysis in Python using the matplotlib library to plot a histogram of the "Signal Inhibition" data and a scatter plot to inspect the relationship between the signal inhibition and overlay scores. In the corresponding programming code, the "Signal Inhibition" column is first assigned to a variable `y`, and the functional groups of interest are assigned to the variable `fgroup_cols`, which is then used to create the matrix X that stores the functional groups matching patterns of those functional groups of interest. Next, a figure with two subplots is initialized by calling `plt.subplots` from matplotlib. The `plt.hist` function adds the histogram to the first subplot (`ax[0]`), and the `plt.scatter` function draws the scatter plot in the subplot to the right (`ax[1]`). The resulting plots show the DKPES inhibitor activity distribution for the 56 compounds that were assayed (left) and the relationship between activity and overlay similarity from ROCS (right), given as the TanimotoCombo score in the range 0 to 2, where 2 means that two 3D structures have an identical volume and partial charge distribution.

From the histogram (left panel of Figure 5.7), we can see that most molecules inhibit the DKPES

signal by less than 50% in in vivo EOG assays. The scatter plot in Figure 5.7 shows that 4 out of the 5 most active molecules have a high overlay similarity value of 1.5 or greater. The TanimotoCombo value is the sum of the volumetric and chemical similarity components, where an exact match (two identical molecules in the same conformation) will result in a maximum score of 1 for each, summing to a maximum score of 2. While the top 4 most active molecules have the highest overlay similarity, no correlation between overlay similarity and signal inhibition can be observed across the remaining 52 molecules. This indicates that more specific determinants of activity are at play, motivating the pattern analysis of functional groups matching DKPES. Interestingly, the outlier with a very low Tanimoto similarity score and a moderately high signal inhibition value of 0.62 is a sulfate tail-containing natural product produced by sea squirts (ZINC14591952; Aiello et al., 2000; shown in Figure 5.8).

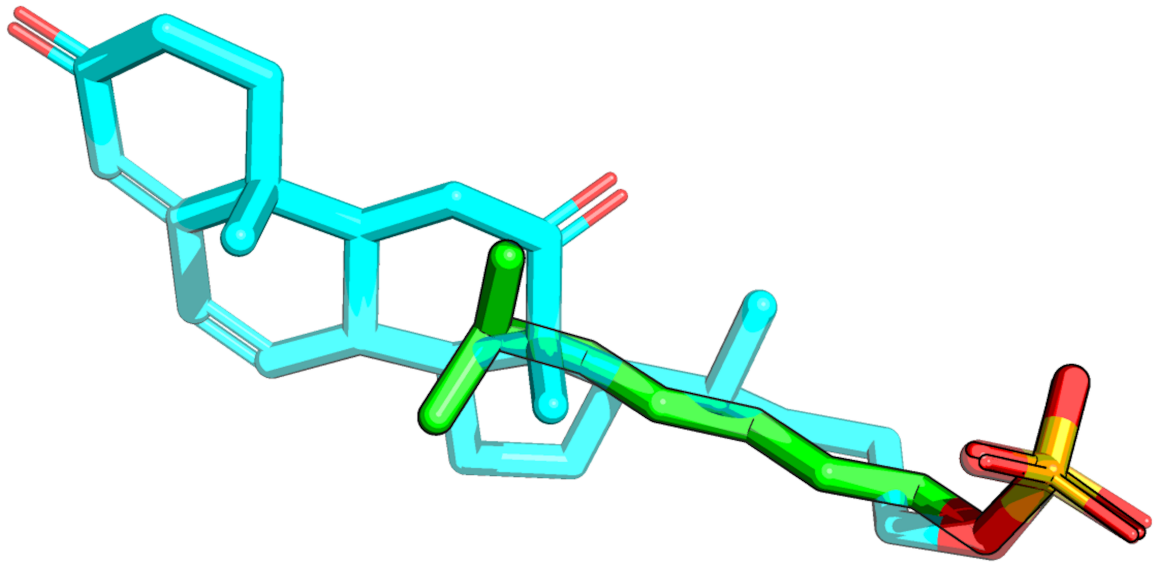


Figure 5.8: Sulfate tail compound sodium 6-methylheptyl sulfate. The Sulfate tail compound sodium 6-methylheptyl sulfate (carbon atoms shown in green; ZINC14591952, average of 62% signal inhibition in EOG experiments) is shown overlaid with the most similar DKPES conformer (carbon atoms shown in cyan).

From this molecule we can conclude that mimicking the sulfate group in DKPES alone can block the olfactory response of DKPES by approximately 60%, likely by competing with interactions of the similar tail in DKPES with the GPCR ligand binding site.

5.4.2 Chemical and functional groups

This section explains how to visualize the other features in the dataset: the functional group matches with the DKPES reference molecule (Figure 5.2). Using the code in Figure 5.9, we will plot the functional group matching pattern of the top 10 most active and 10 least active molecules via two heat maps shown side by side for a visual comparison.

We chose a 1.3 Å cut-off between overlayed atoms to identify functional group matches in 3D. If two molecules share the same atom type at a distance greater than 1.3 Å, this was not considered a functional group match. This relatively generous distance cut-off (nearly a covalent bond- length) was chosen to account for minor deviations in the crystal structures and overlays when comparing functional groups between pairs of molecules. Note that changing the distance threshold generally will affect the resulting functional group matching patterns. For instance, the 3-hydroxy group in ZINC72400307 (Figure 5.5) does not overlay with the 3-keto group of the DKPES query (Figure 5.2) in our analysis since the distance between those two atoms is 1.7 Å. We recommend choosing distance thresholds up to 1.3 Å.

Looking at the heat maps in Figure 5.9, the following conclusions can be drawn:

- The top 9 most active molecules have a sulfur match and match three of the oxygen atom in the DKPES sulfate group.
- Sulfur and sulfate oxygen atom matches alone are not sufficient for activity. From the previous scatter plot analysis (Figure 5.7), we know that the sulfate tail analog alone (Figure 5.8; ZINC14591952) shows a signal inhibition of 60%. It matches the 3 terminal sulfate-oxygens and sulfur atom. However, a compound with the same matching pattern (ZINC22058386 in Figure 5.9) has no biological activity in the same assay, likely due to its greater bulk (Figure 5.5).

However, casual inspection of the data does not always lead to insights that apply to all of the compounds, and it can miss interesting trends, especially for large datasets. The next section will

```
>>> fig, ax = plt.subplots(1, 2, figsize=(10, 3))
>>> ax[0].imshow(X[:10], cmap='binary')
>>> ax[0].set_title('10 most active molecules')
>>> plt.sca(ax[0])
>>> plt.xticks(range(len(fgroup_cols)),
...            fgroup_cols, rotation='vertical')
>>> plt.yticks(range(10), df['index'][:10])
>>> ax[1].imshow(X[-10:], cmap='binary')
>>> ax[1].set_title('10 least active molecules')
>>> plt.sca(ax[1])
>>> plt.xticks(range(len(fgroup_cols)),
...            fgroup_cols, rotation='vertical')
>>> plt.yticks(range(10), df['index'][-10:])
>>> plt.show()
```

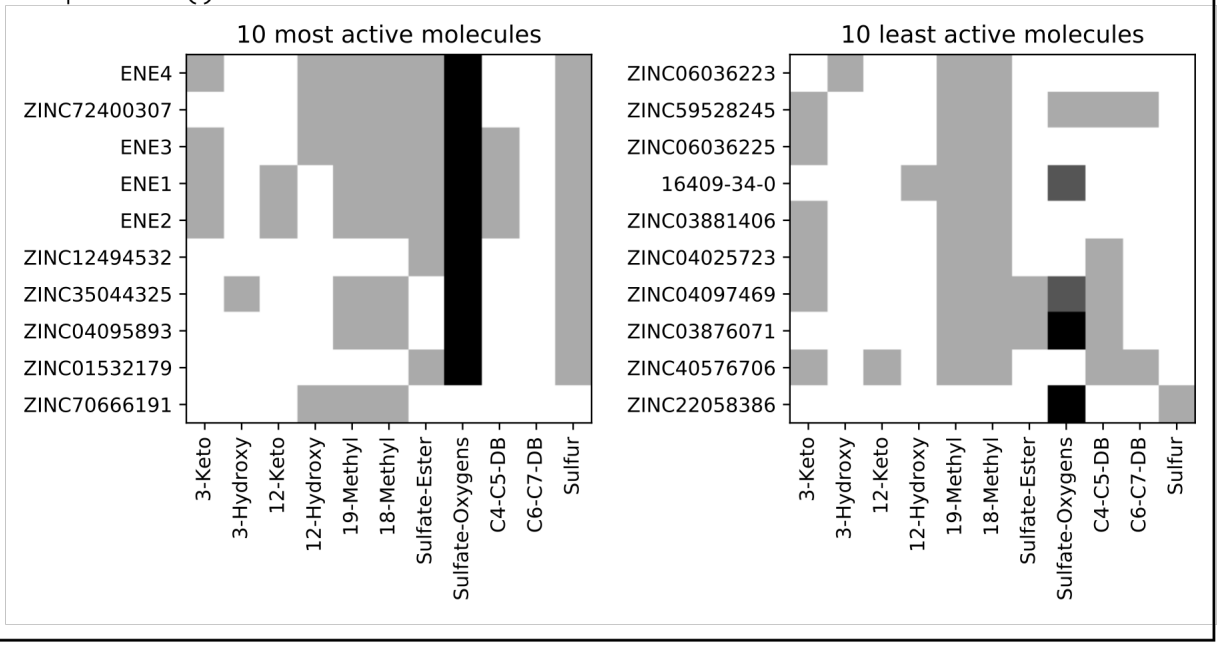


Figure 5.9: Code to generate heat maps showing matches of functional groups in DKPES by the 10 most active and 10 least active molecules tested in EOG assays. Using the matplotlib.pyplot module that was important as `plt` earlier (Figure 5.7), we create two subplots stored in the array `ax`. Using matplotlib.pyplot's `imshow`, we plot the functional group patterns of the ten most active molecules (`X[:10]`, the first ten elements in the sorted data array) as a heat map in left subplot (`ax[0]`). Similarly, we plot the ten least active molecules (the last ten molecules in the array, `X[-10:]`) as a heat map in the right subplot (`ax[1]`). As the heat maps show, all features except sulfate-oxygens are encoded as binary variables (0: white cell background, no match; 1: light gray, match). Sulfate-oxygens refers to the three terminal oxygens, excluding the sulfate ester oxygen. This variable has values from 0-3 (up to all 3 terminal oxygens being matched), where black cell backgrounds correspond to 3 matches, dark gray corresponds to 2 matches, and light gray to 1 match, respectively.

introduce several machine learning approaches for deducing the importance of functional groups for biological activity.

5.4.3 Tracing preferential chemical group patterns using decision trees

Decision tree classifiers are a good choice if we are concerned about the interpretability of the combinations of features used to predict activity. While decision trees can be trained to predict outcomes on a continuous scale (regression analysis), we fill focus on decision trees for classification in this chapter, that is, predicting whether a molecule is active or non-active. While the discretization of the continuous target variable (here: *signal inhibition in percent*) is to some extent arbitrary, it helps with generating less noisy predictions and with improving the interpretability of the selected features as they can be directly interpreted as discriminants of active and non-active molecules. For the following analysis, we considered molecules with a signal inhibition of 60% or greater as active molecules.

As you will see, within a tree it is easy to trace the path of decisions comprising the model that best separates different classes of molecules (here: active vs non-active). In other words, based on the functional group matching information in the DKPES dataset, the decision tree model poses a series of questions to infer the discriminative properties between active and non-active molecules. While there is technically no minimum number of molecules required for using the techniques outlined in this chapter, we recommend collecting datasets of at least 30 structures for the automatic inference of functional groups that discriminate between active and non-active molecules.

Although this is difficult to achieve in practice, an ideal dataset for supervised machine learning would be balanced, that is, with an equal number of positive (active) and negative (non-active) training examples. While there is no indication that class imbalance was in issue for the DKPES dataset, as the results of the decision tree analysis were unambiguous, imbalance may be an issue in other datasets. There are many different techniques for dealing with imbalanced datasets, including several resampling techniques (oversampling of the minor-

ity class or under-sampling of the majority class), the generation of synthetic training samples, and reweighting the influence of different class labels during the model fitting. A review of techniques for working with imbalanced datasets can be found in (Raschka & Mirjalili, 2017). For machine learning with scikit-learn, a compatible Python library that has been developed to deal with imbalanced datasets (http://contrib.scikit-learn.org/imbalanced-learn/). Also note that classifiers in scikit-learn, including the `DecisionTreeClassifier`, accept a `class_weight argument`, which can be used to put more emphasis on a particular class (e.g., active or non-active) during model fitting, thereby preventing that the decision tree algorithm becomes biased towards the most frequent class in the dataset. For more information on how to use the `class_weight` parameter of the `DecisionTreeClassifier`, refer to the documentation at http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

The learning algorithm that is constructing a nonparametric decision tree model from the dataset works as follows. Starting at the tree root and splits the dataset (the active and non-active molecules) on the feature (for example, presence of a sulfur match) that results in the largest information gain. In other words, the objective function of a decision tree is to learn, at each step, the splitting criterion (or decision rule) that maximizes the information gain upon splitting a parent node into two child nodes. The information gain is computed as the difference between the impurity of a parent node and the sum of its child node impurities. Intuitively, we can say the lower the impurity of the child nodes, the larger the information gain. The impurity itself is a measure of how diverse the subset of samples is, in terms of the class label proportion, after splitting. For example, after asking the question "does a molecule have a positive sulfur match?" a pure node would only contain either active or non-active molecules when answering this question with a "yes." A node that consisted of 50% non-active and 50% active samples after applying a splitting criterion would be most impure – such a result would indicate that it was not a useful criterion for distinguishing between active and non-active molecules. In the decision tree implementation that we are going to use in this chapter, the metric for computing the impurity of a given node is measured as Gini impurity as used in the CART (classification and regression tree) algorithm (Breiman et al., 1984). Gini impurity is

defined as

$$\text{Impurity}(t) = \sum_{i=1}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t))^2, \tag{5.1}$$

where t stands for a given node, i is a class label in $c = \{\text{active}, \text{non-active}\}$, and $p(i|t)$ is the proportion of the samples that belongs to class i for a particular node t. Looking at the previous equation, it is easy to see that the impurity of a given node is minimal if the node is pure and only contains samples from one class (for examples, actives), since $1 - (12 + 02) = 0$. Vice versa, if samples at one node are perfectly mixed, the Gini impurity of a node is maximal: $1 - (0.52 + 0.52) = 0.5$. In an iterative process, the splitting procedure is then repeated at each child node until the leaves of the tree are pure, which means that the samples at each node all belong to the same class (either active or non-active), or cannot be separated further due to the lack of discriminatory information in the dataset. For more information about decision tree learning, see Raschka, 2015 and (Louppe, 2014).

To build a decision tree classifier (as opposed to a decision tree regressor), we discretize the signal inhibition variable, creating a binary target variable `y_binary`. Using the code in Figure 5.10, active molecules are specified as molecules with signal inhibition of 60% or greater (class 1), and molecules with less than 60% signal inhibition are labeled as non-active (class 0).

```
>>> y_binary = np.where(y >= 0.6, 1, 0)
>>> np.sum(y_binary)
12
```

Figure 5.10: Code for discretizing the continuous signal inhibition variable. The `numpy.where` function creates a new array, `y_binary`, where all molecules with more than 60% signal inhibition will be labeled as 1 (active), and all other molecules will be labeled with a 0 (non-active).

As can be seen from computing the sum of values in the `y_binary` array (`np.sum(y_binary)`; Figure 5.10), discretization of the continuous signal inhibition variable resulted in 12 molecules labeled as active; consequently, the remaining 44 molecules in the dataset are now labeled as non-active. In the next step, we will initialize a decision tree classifier from scikit-learn with default

values, let it learn the decision rules that discriminate between actives and non-actives from the dataset, and export the model and display it as a decision tree (Figure 5.11). Deep, unpruned decision trees with many decision points are notoriously prone to overfitting. This is analogous to the overfitting problem in parametric regression, where including more terms with adjustable weights allows better fit to a set of training data, while resulting in complex decision rules that are hard to interpret and do not perform well on held-out or new data. This is why we preferred classification trees over decision trees for regression analysis for the single decision tree and random forest analyses in this chapter.

We conclude from the binary classification tree (Figure 5.11) that a majority of the active inhibitors (8 of 12) share a sulfur atom and a sulfate ester group that overlay with the respective functional groups in DKPES; none of the non-active compounds have these characteristics. With decision trees, the resulting models can offer intuitive insights into the hypothesis space. Specifically, the tree in Figure 5.11 indicates that, given a set of molecules initially selected as having high volumetric and chemical similarity with DKPES, the presence of a sulfur atom and sulfate ester group matching those two groups in DKPES predicts the subset of molecules that are active as DKPES inhibitors.

Using machine learning to derive decision rules objectively and automatically is convenient and less error-prone in providing insights compared with visual analysis of functional group patterns in a heat map. Note that the problem analyzed here as a case study is not a classical example of machine learning, in which a classifier is fit to a training dataset, and then its accuracy of prediction (and generalizability to new data) is estimated on held-out data by using a test set or cross-validation techniques. In this chapter, we are describing general approaches for analyzing the importance of various functional groups for the activity of molecules. Our primary goal is not to build a predictor to classify new molecules as active or non-active, although the models developed in this chapter could indeed be used in such a way.

```
>>> from sklearn import tree
>>> clf = tree.DecisionTreeClassifier()
>>> clf = clf.fit(X, y_binary)
>>> dot_data = tree.export_graphviz(clf, out_file=None,
...                                 feature_names=fgroup_cols,
...                                 class_names=['non-active', 'active'],
...                                 filled=True, rounded=True)
>>> graph = pydotplus.graph_from_dot_data(dot_data)
>>> graph.write_pdf("tree.pdf")
```

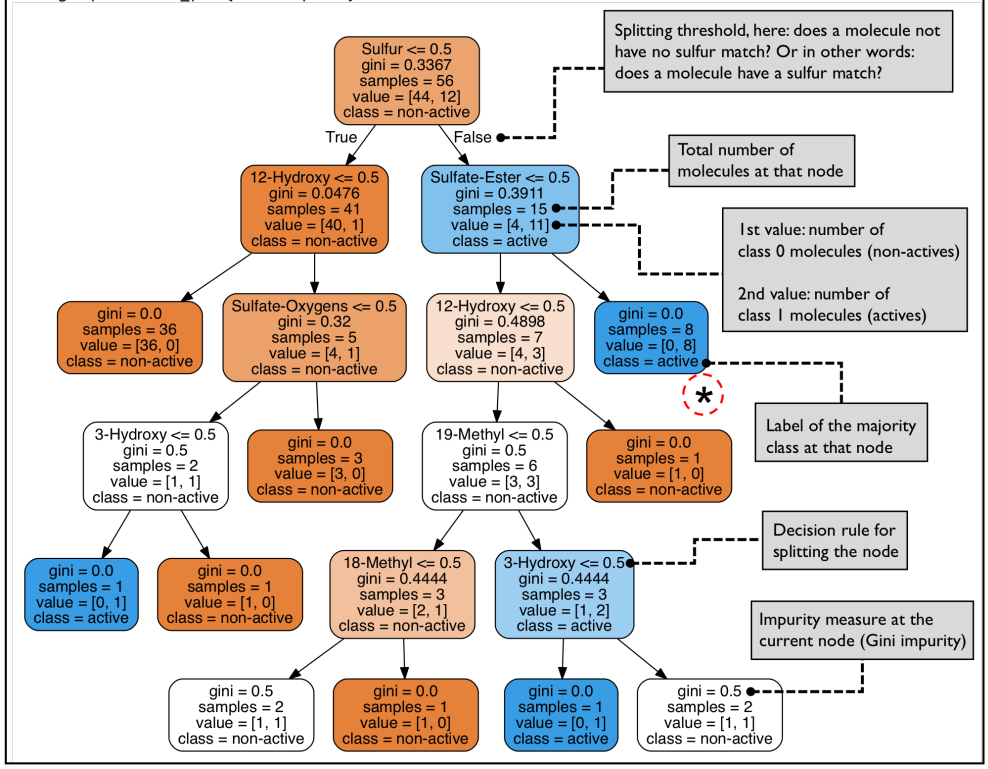


Figure 5.11: Binary classification tree separating active from non-active compounds. After importing the tree submodule from the scikit-learn machine learning library, the first line of code initializes a new `DecisionTreeClassifier` object that is then learning the decision rules from the functional group matching pattern array (`X`) and the discretized response variable (binary labels of the active and non-active molecules, `y_binary`) by calling the `fit` method. The last three lines of code then export the fitted decision tree as a PDF image, which is shown here. The first node at the top of the tree, for example, uses a decision rule asking which molecules in the 56-molecule dataset (44 actives and 12 non-actives) match a sulfur group in DKPES. Note that this question is posed as a conditional (true/false) statement "Molecules do not contain a sulfur group match", due to the implementation of the decision tree in scikit-learn. The molecules for which the condition is "False" – that is, molecules that do match the sulfur group in DKPES – are then passed to the child node on the right (here: 4 non-actives and 11 actives), where the next conditional statement is "Molecules do not contain a 'Sulfate-Ester' match." Each node in the tree contains the impurity measure after the split (Gini impurity), reflecting the degree of separation between active and non-active compounds; a Gini impurity value of 0 reflects a set containing purely active or non-active compounds. The number of samples refers to the compounds at each node that pass the filtering criteria. The first value within brackets in the bottom row in each terminal node denotes the number of non-active compounds at that node, and the second number denotes the number of active compounds. Highlighted with an asterisk is the terminal node (to the center-right of the plot), which contains eight active compounds and no non-active compounds. For visual clarity, containing more non-active molecules than actives are labeled in orange, and nodes that contain more actives than actives are colored in blue. The higher the color intensity, the higher the ratio of active molecules or non-active molecules, respectively.

5.4.4 Deducing the importance of chemical groups via random forests

To estimate the relative importance of the different functional groups based on active and non-active labels, we will now construct a random forest model (Breiman, 2001), which is an ensemble method based on multiple decision trees. In the random forest models, the feature importance is measured as the averaged impurity decrease computed from multiple decision trees. In the following code example (Figure 5.12), we will use the random forest algorithm implemented in scikit-learn to create and ensemble of 1,000 decision trees, which are grown from different bootstrap samples of the compound dataset and randomly selected subsets of functional group feature variables. (A bootstrap sample is generated by randomly drawing samples from the original dataset with replacement to generate a resampled dataset of the same size as the original one.)

```
>>> from sklearn.ensemble import RandomForestClassifier
>>> forest = RandomForestClassifier(n_estimators=1000,
...                                 random_state=0,
...                                 n_jobs=-1)
>>> forest.fit(X, y_binary)
```

Figure 5.12: Code for fitting a random forest. Similar to fitting a `DecisionTreeClassifier` (Figure 5.11), we first initialize a new `RandomForestClassifier` object from scikit-learn and fit it to the functional group matching pattern array (`X`) and labels of the active and non-active molecules (`y_binary`). By setting `n_estimators=1000`, we will use 1000 decision trees for the forest. Here, `n_jobs=-1` means that we are utilizing all processors on our machine to fit those decision trees in parallel to speed up the computation. The `random_state` parameter accepts an arbitrary integer for the bootstrap sampling and feature selection in the decision tree to make the experiment deterministic and reproducible.

Based on the random forest model, we can infer feature importance by averaging the impurity decrease for each feature split from all 1000 trees in the forest. Conveniently, the random forest implementation in scikit-learn already computes the feature importance upon model fitting, so that we can access this information from the forest, after calling the `fit` method, via its `feature_importances_ attribute`. The code in Figure 5.13 will create a bar plot of the feature importance values, which are normalized to sum up to 1 for easier interpretation.

As shown by the bar plot in Figure 5.13, the feature importance values computed from the 1,000 regression trees agree with the conclusions drawn previously in the Methods sections 3.3 and 3.4:

```
>>> importances = forest.feature_importances_
>>> indices = np.argsort(importances)[::-1]
>>> feature_labels = np.array(fgroup_cols)
>>> plt.bar(range(X.shape[1]),
>>>         importances[indices],
>>>         align='center')
>>> plt.xticks(range(X.shape[1]),
...            feature_labels[indices], rotation=90)
>>> plt.xlim([-1, X.shape[1]])
>>> plt.ylabel('Relative feature importance')
>>> plt.tight_layout()
>>> plt.show()
```

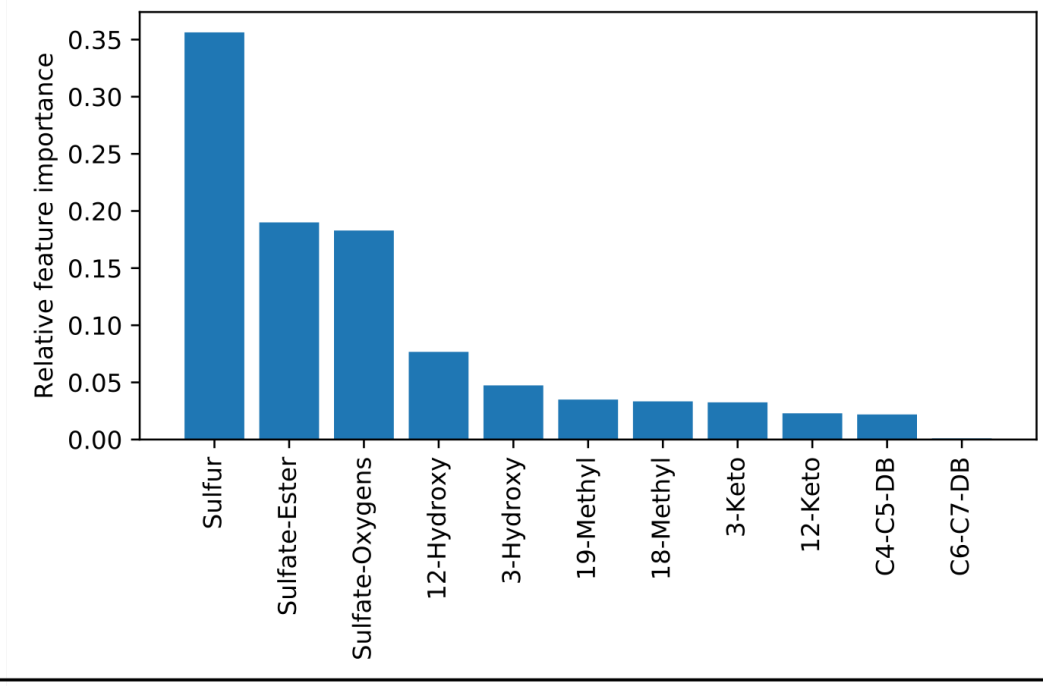


Figure 5.13: Relative feature importance of the functional group matches. The relative feature importance of the functional group matches is inferred from the random forest model that was trained to discriminate between active and non-active molecules. First, the importances values are sorted from highest to lowest using NumPy's `argsort` function. Next, we summarize the computed feature importance in a bar plot using matplotlib's pyplot submodule, which was imported as `plt` earlier.

sulfur, sulfate ester, and sulfate oxygen groups are the most important functional group features for DKPES inhibitor activity (Figure 5.11).

While the feature importance values provide us with a numeric value to quantify the importance of features, these quantities do not provide information about whether the presence or absence of the particular functional group matches are characteristic of the active molecules. However, we can easily determine whether active molecules match a certain functional group by inspecting the heat map visualizations of active and non-active molecules (Figure 5.9).

Concerning the interpretation of feature importance values from random forests, note that if two or more features are highly correlated, one feature may be ranked much higher than the other feature, or both features may be equally ranked. In other words, the importance or information in the second feature may not be fully captured. The potential bias in interpreting the feature importance from random forest models has been discussed in more detail by (Strobl et al., 2008). In general, this issue can be pre-assessed by measuring the degree to which series of values for two features across a set of compounds are correlated by calculating the Pearson linear correlation coefficient or by calculating the Spearman rank correlation coefficient to assess similar ranking of values between the features across a set of compounds (which does not assume a linear relationship between variables). The Spearman and Pearson correlation coefficients can be computed using the `peasonr` and `spearmanr` functions from the scipy.stats package (please refer to the official SciPy documentation at https://docs.scipy.org/ for more information). While the predictive performance of a random forest is generally not negatively affected by high correlation among feature variables (multicollinearity), it is recommended to exclude highly correlated features from the dataset for feature importance analysis, for instance, via recursive feature importance pruning (Strobl et al., 2009).

5.4.5 Sequential feature selection with logistic regression

As an alternative approach and to probe the robustness of our conclusions, we will apply a sequential backward selection (SBS) algorithm combined with logistic regression (Walker & Duncan, 1967) for

the classification of active versus non-active compounds. SBS is a model-agnostic feature selection algorithm that evaluates different combinations of features, shrinking the subset of features to be considered one by one. Here, model-agnostic refers to the fact that SBS can be combined with any machine learning algorithm for classification or regression.

In general, sequential feature selection algorithms are greedy search algorithms that reduce the d-dimensional feature space to a smaller k-dimensional subspace, where $k < d$. The sequential feature selection approach selects the best-performing feature subsets automatically and can help optimizing two objectives: improving the computational efficiency and reducing the generalization error of a model by getting rid of features that are irrelevant.

The SBS algorithm removes features initial feature subset sequentially until the new, reduced feature subspace contains a specified number of features. To determine a feature that is to be removed at each iteration of the SBS algorithm, we need to define a criterion function J, which is to be minimized. For instance, this criterion function is defined as the difference between the performance of the model before and after the feature removal. In other words, at each iteration of the algorithm, the feature that results in the least performance loss (or most performance gain) after removal is eliminated. This removal of features is repeated in each iteration of the algorithm until the desired, pre-specified size of the feature subset is reached. More formally, we can express the SBS algorithm in the following pseudo-code notation adapted from Raschka & Mirjalili, 2017:

1. Initialize the algorithm with $k = d$, where d is the dimensionality of the full feature space $\mathbf{X}_d$.

2. Determine the feature x^- that maximizes the criterion: $x^- = \arg\max J(\mathbf{X}_d - x)$, where $x \in \mathbf{X}_k$.

3. Remove the feature x^- from the feature set: $\mathbf{X}_{k-1} = \mathbf{X}_k - x^-; k = k - 1$.

4. Terminate if k equals the number of desired features; otherwise, go to step 2.

The reason why we chose sequential feature selection to deduce functional group matching patterns that are predictive of active and non-active molecules is that it presents an intuitive method that has shown to produce accurate and robust results. For more information on sequential feature selection, please refer to Ferri et al., 1994.

Sequential feature selection constitutes just one of many approaches to select feature subsets. Univariate feature selection methods that consider one variable at a time and select features based on univariate statistical tests, for example, percentile thresholds or p-values. A good review of feature selection algorithms can be found in Saeys et al., 2007. However, the main advantage of sequential feature selection over univariate feature selection techniques is that sequential feature selection analyzes the effect of features on the performance of a predictive model considering the features as a synergistic group. Other techniques, related to sequential feature selection, are genetic algorithms, which have been successfully used in biological applications to find optimal feature subsets in high-dimensional datasets as discussed in Raymer et al., 2000 and Raymer et al., 1997.

Logistic regression is one of the most widely used classification algorithms in academia and industry. One of the reasons why logistic regression is a popular choice for predictive modeling is that it is easy to interpret as a generalized linear model: The output always depends on the sum of the inputs and model parameters. However, note that sequential feature selection can be used with many different machine learning algorithms for supervised learning (classification and regression).

To introduce the main concept behind logistic regression, which is a probabilistic model, we need to introduce the so-called *odds ratio* first. The odds ratio computes the *odds* in favor of a particular event E, which is defined as follows, based on the probability p of a positive outcome (for instance, the probability that a molecule is active),

$$\text{odds} = \frac{p}{(1-p)}. \tag{5.2}$$

Next, we define the *logit* function, which is the logarithm of the odds ratio,

$$\text{logit(p)} = \frac{p}{(1-p)}. \tag{5.3}$$

The logit function takes values over the range 0 to 1 (the probability p) and transforms them to real numbers that describe the relationship between the functional group matching patterns, multiplied with weight coefficients (that need to be learned) and the odds that a given molecule is active,

$$\text{logit}(p(y = 1|x)) = w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + b = \sum_{i=1}^{m} w_i x_i + b. \tag{5.4}$$

Here, m is an index over the input features (functional group matches, x), w refers to the weight parameters of the parametric logistic regression model, and b refers to the y-axis intercept (typically referred to as *bias* or *bias unit* in literature). The input to the logit function, $p(y = 1|x)$, is the conditional probability that a particular molecule is active, given that its functional group matches x.

However, since we are interested in modeling the probability that a given molecule is active, we need to compute the function inverse ϕ of the logit function, which we can compute as

$$\phi(z) = \frac{1}{1 + e^{(} - z)}. \tag{5.5}$$

Here, z is a placeholder variable defined as

$$z = \sum_{i=1}^{m} w_1 x_1 + b. \tag{5.6}$$

The logistic regression implementation used in this section learns the weights for the parameters (matched chemical features) of the logistic regression model that minimizes the logistic cost function, which is the probability of making a wrong prediction given the number of n active and non-active molecule labels in the set of compounds, where the binary vector y stores the class labels (1=active, 0=non-active),

$$l(w) = \sum_{i=1}^{n} \left[y^{(i)} \log \left(\phi\left(z^{(i)}\right) \right) + (1 - y^{(i)} \log \left(1 - \phi\left(z^{(i)}\right) \right) \right]. \tag{5.7}$$

For more information about logistic regression, see reference (Walker & Duncan, 1967).

Now, by combining a logistic regression classifier with a sequential feature selection algorithm, we can identify a fixed-size subset of functional groups that maximizes the probability of correct prediction of which compounds are active.

Since we are interested in comparing feature subsets of different sizes to identify the smallest feature set with the best performance, we can run the SBS algorithm stepwise down to a set with only one feature, allowing it to evaluate feature subsets of all sizes, by using the code shown in Figure 5.14. Furthermore, the SBS implementation uses k-fold cross-validation for internal performance validation and selection. In particular, we are going to use 5-fold cross-validation.

In 5-fold cross-validation, the dataset is randomly split into k non-overlapping subsets or folds (a molecule cannot be in multiple subsets). From the five splits, four folds are used to fit the logistic regression model, and one fold is used to compute the predictive performance of the model on held-out (test) data. 5-fold cross-validation repeats this splitting procedure five times so that we obtain five models and performance estimates. The model performance is then computed as the arithmetic average of the five performance estimates. For more details about k-fold cross-validation, please see the online article, "Model evaluation, model selection, and algorithm selection in machine learning – Cross-validation and hyperparameter tuning" at https://sebastianraschka.com/blog/2016/model-evaluation-selection-part3.html.

We chose 5-fold cross-validation to evaluate the logistic regression models in the sequential backward selection since $k = 5$ it is a commonly used default value in k-fold cross-validation. Generally, small values for k are computationally less expensive than larger values of k (due to the smaller training set sizes and fewer iterations). However, choosing a small value for k increases the pessimistic bias, which means the performance estimate underestimates the true generalization performance of a model. On the other hand, increasing the size of k increases the variance of the estimate. Unfortunately, the No Free Lunch Theorem (Wolpert, 1996) – stating that there is no algorithm or choice of parameters that is optimal for solving all problems (as shown by Bengio & Grandvalet, 2004) – also applies here. For an empirical study of bias, variance, and bias-variance trade-offs in cross-validation, also see Kohavi, 1995.

```
>>> from mlxtend.feature_selection import SequentialFeatureSelector as SFS
>>> from mlxtend.plotting import plot_sequential_feature_selection as plot_sfs
>>> from sklearn.linear_model import LogisticRegression
>>> classifier = LogisticRegression()
>>> sfs = SFS(classifier,
...           k_features=1,
...           forward=False,
...           floating=False,
...           scoring='accuracy',
...           verbose=0,
...           cv=5)
>>> sfs = sfs.fit(X, y_binary)
>>> fig1 = plot_sfs(sfs.get_metric_dict(), kind='std_err')
>>> plt.ylim([0.5, 1])
>>> plt.ylabel('Accuracy')
>>> plt.xlabel('Number of features in the selected subset')
>>> plt.grid()
>>> plt.show()
```

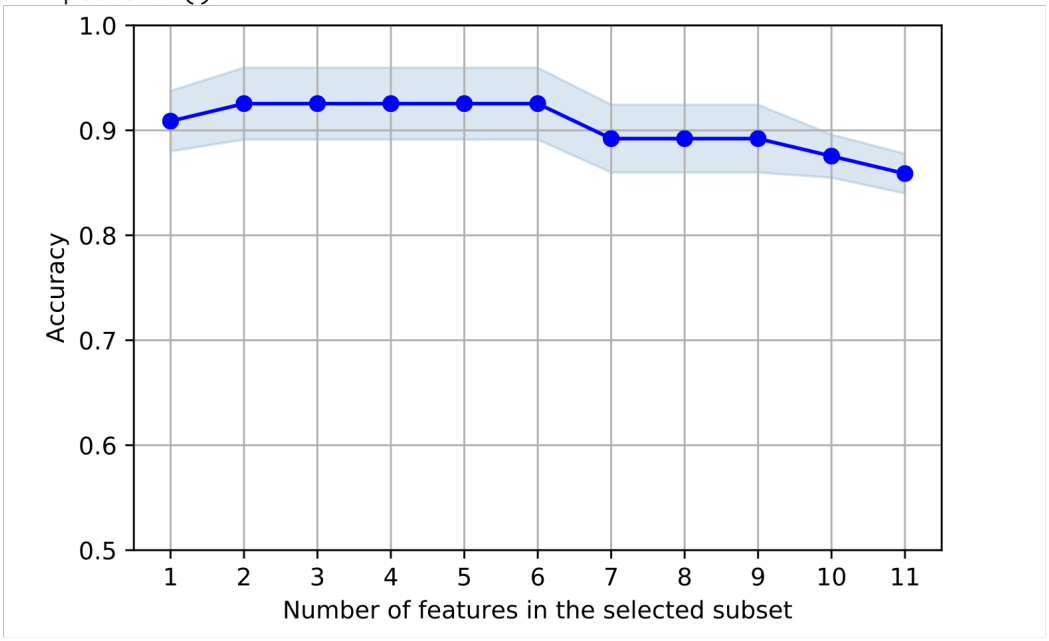


Figure 5.14: Performing sequential feature selection using logistic regression to identify features that discriminate between active and non-active molecules. After importing the Python classes for fitting the `LogisticRegression` classifier within the `SequentialFeatureSelector`, by setting `forward=False` and `floating=False`, we specify that the sequential feature selector should perform regular backward selection. Then we use the `plot_sfs` function to visualize the results with matplotlib's pyplot submodule. The resulting plot in this figure shows the classification accuracy of the logistic regression models trained on different feature subsets (functional group matching patterns) via sequential backward selection. The prediction accuracy (0=worst, 1=best), where 1 corresponds to 100% accuracy in predicting active versus non-active compounds across the input set, was then computed via 5-fold cross validation. The plot presents the average prediction accuracy (whether the model can predict held-out active and non-active molecules given their functional group matching patterns) across the 5 different test sets. The error margin (pale blue region above and below the dark blue average points) shows the standard error of the mean for the 5-fold cross validation.

As can be seen in Figure 5.14, the performance of the classification algorithm does not change significantly across the different feature subset sizes. The feature subsets with size 2-6 have the highest accuracy, indicating that adding more features to the 2-feature subset does not provide additional discrimination between active and non-active molecules. The decline in accuracy after adding a seventh feature to the set is likely due to the curse of dimensionality (Hughes, 1968). In brief, the curse of dimensionality describes the phenomenon that a feature space becomes increasingly sparse if we increase the number of dimensions (for instance, by adding additional functional group matching features) given a fixed number of samples in the training set, which will more likely result in overfitting and less accurate results. While the execution of the code in Figure 5.14 provided us with insights regarding the best-performing feature subset sizes via SBS in predicting active or non-active molecules, we haven't determined what those features are. Since there is no information gain by going beyond two-feature set (Figure 5.14), we will use the following code (Figure 5.15) to extract the feature names:

```
>>> sfs.subsets_[2]
{'avg_score': 0.92545454545454553,
 'cv_scores': array([ 1., 1., 0.81818182, 0.90909091, 0.9 ]),
 'feature_idx': (10, 6)}
```

Figure 5.15: Code to obtain the feature names of the best-performing feature subset from sequential backward selection. The `subsets_` attribute of the sequential feature selector (`sfs`) refers to a Python dictionary that stores the feature (functional group matches) indices and cross-validation information. By looking up the dictionary entry at index position 2, we can access the feature indices of the 2-feature subset, 10 and 6, and by using `sfs.subsets_[2]` as an index to the `feature_labels` array that we defined earlier (Figure 5.13) and reporting the feature labels, we can see that "Sulfur" and "Sulfate-Ester" matches are the most discriminatory features of active and non-active molecules.

The output from the code executed in Figure 5.14 shows that the 2-feature subset consisting of "Sulfur" and "Sulfate-Ester" matches has the most discriminatory information for separating active and non-active molecules as DKPES mimics. This information is consistent with the conclusions drawn from the previous random forest and decision tree analyses.

Now we have shown how to use decision trees, random forest models, and logistic regression to analyze which features can best discriminate between active and inactive compounds, and to assess

the relative importance of the different features for discrimination. Such methods provide human interpretable information on chemical features important for activity, and concurrence between the methods strengthens the conclusions.

In a related pheromone inhibitor project, we used the results of feature importance analysis to drive the selection of compounds in a subsequent round of virtual screening that required few compounds to be assayed and resulted in significant enhancement of activity and new knowledge about functional group importance. Those compounds are now being tested by members of our research team for invasive species behavioral modification in the tributaries of the Great Lakes of North America under an EPA permit (Raschka et al., 2017). Analysis of whether the set of features and their relative importance hold equally well for different subsets of assayed compounds (e.g., steroids versus non-steroids) is another valuable direction of inquiry.

The chemical features identified as most important by machine learning will depend on the chemical diversity within the set of molecules for which assay results and chemical structures are analyzed. For instance, if only steroid compounds are tested versus only non-steroids, likely the chemical features found to be most important will differ. In our case, for the steroid set, the side groups providing specific interactions were most important (since the steroid scaffold is in common to all of them), whereas for the non-steroids, compounds that mimic and shape and hydrophobic interactions of the steroidal pheromone may also be important. Thus, considering the set of compounds to be analyzed, and testing the generalizability of the features derived is worth some thought. If you have different chemical classes of compounds to analyze, and a significant number of compounds in each, you can carry out the machine learning analysis of the most important features for each of the groups of compounds separately, as well as all of the compounds together, to discern the extent to which highly ranked features that discriminate between actives and inactives are shared among compounds based on different chemical scaffolds. For instance, training models for molecules from different classes of molecules can help against averaging-effects where a certain pattern only occurs in once class of molecules but not the other. While predictions obtained from different models trained on different subsets of molecules cannot be trivially combined, separate

analyses on different classes of molecules can provide useful information about the relationship between physicochemical properties and activity (for example, while having similar functional group matching patterns as active molecules, individual molecules could be inactive as they are too large or bulky to bind to a given receptor binding pocket.

5.4.6 Conclusions

From the decision tree analysis (Section 5.4.3), random forest feature importance estimation (Section 5.4.4), and sequential feature selection results (Section 5.4.5), we can conclude that the sulfate groups (Sulfur, Sulfate-Ester, and Sulfur-Oxygen features) are the most discriminatory features for distinguishing active from non-active compounds in DKPES-mediated olfactory responses. From the inspection of heat maps showing the top 10 active and 10 least active molecules (Section 5.4.2), we also observed that presence of sulfate tails are a consistent determinant of activity. One compound consisting only of a sulfate tail (ZINC14591952; Figure 5.8) resulted in 62% signal inhibition, which supports the hypothesis that sulfate groups are a key feature of active molecules. Figure 5.16 summarizes the results from the random forest feature importance estimation by comparing the importance values to the proportion of functional group matches in active and non-active molecules.

The data in Figure 5.16 shows that matching "Sulfur" and "Sulfur Oxygens" are the most discriminatory features for a random forest to distinguish actives from non-actives, and both features also have a high rate of occurrence in active versus non-active molecules. Features that do not appear substantially more frequently in active molecules than in non-actives (are not discriminatory of activity), for example, 18-methyl, 19-methyl, 3- keto, or the presence of either the C4-C5 or C6-C7 double bond ("DB") also have a low random forest feature importance. Interestingly, the feature importance of Sulfate-Ester is much less than the feature importance of Sulfur or Sulfur-Oxygens, which may be because it is highly correlated with the sulfur and sulfur oxygen matches in the sulfate group and thereby, to some extent, redundant. An alternative explanation is that the ester oxygen is less highly charged than the terminal sulfate oxygens (causing it to make weaker

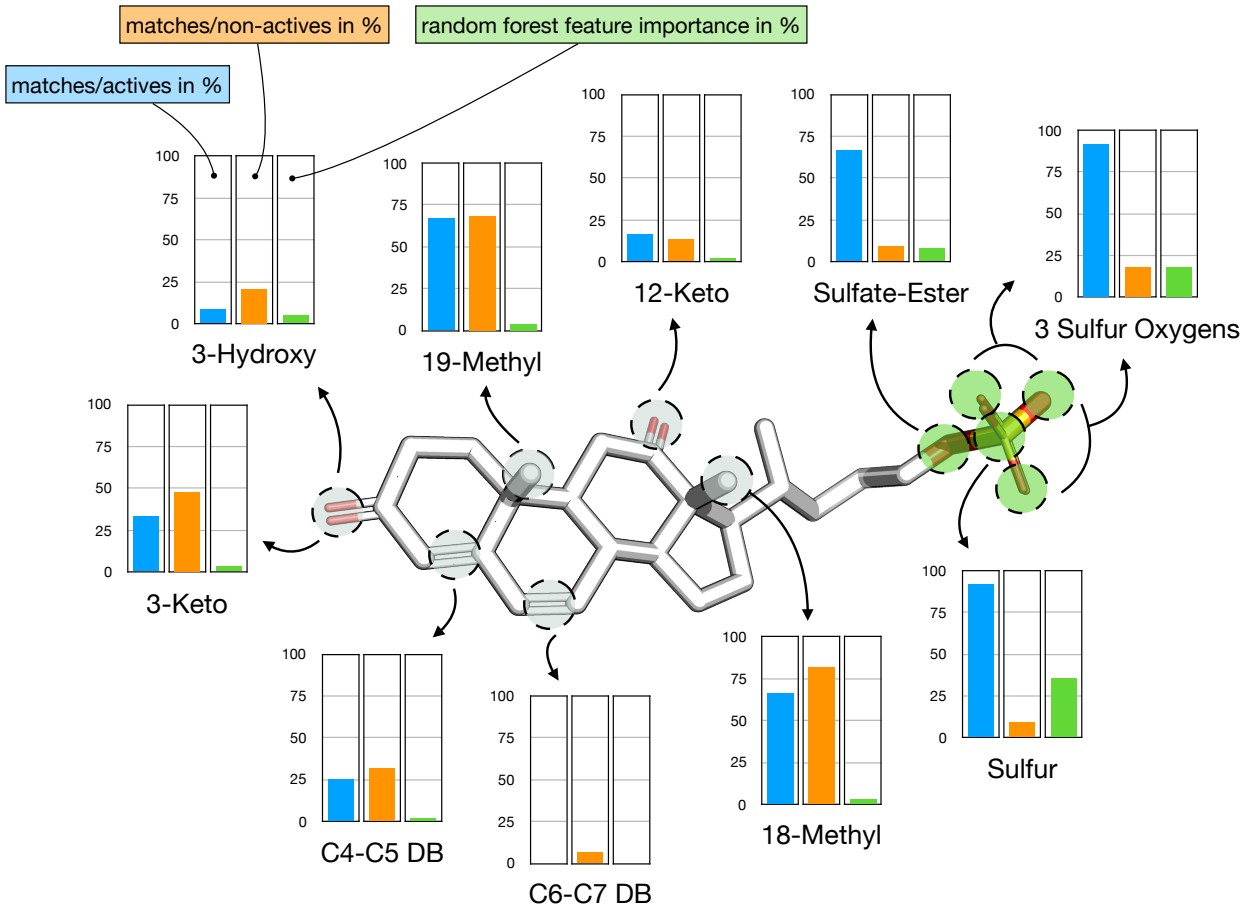


Figure 5.16: Proportion and relative importance of functional group matches. This figure summarizes the proportion of functional group matches across the 12 active and 44 inactive molecules and relative functional group feature importance (from the random forest analysis, Figure 5.13) mapped onto the DKPES reference molecule. DB refers to "double bond."

hydrogen bonds) and is also less accessible for interaction with the receptor.

The machine learning techniques presented in this chapter can be used for *any* kind of data for which a set of feature values across a set of objects is used to predict activity (or any observable value determined by an experimental technique, for example, solubility, selectivity, reactivity, and so forth). We hope this chapter has whetted your appetite for machine learning, which provides robust models that relate features of interest to molecular activity and other observables. The code provided here and on the corresponding website (https://github.com/psa-lab/predicting-activity-by-machine-learning) makes it possible for you to learn and then use these techniques in your own research. For further information about machine learning, and to carry out further explorations

with prepared datasets or your own data, we recommend the following tutorials and references: Raschka & Mirjalili, 2017, Raschka et al., 2016b, Friedman et al., 2001, Mueller & Guido, 2017, and the scikit-learn online tutorials (http://scikit-learn.org/stable/tutorial/index.html).

CHAPTER 6

3D EPITOPE-BASED VIRTUAL SCREENING: A NEW METHOD FOR DISCOVERING SMALL MOLECULE INHIBITORS OF PROTEIN-PROTEIN INTERACTIONS

Computational methods and work described in this chapter have been developed by Sebastian Raschka and Leslie Kuhn. Experimental assays (results are not included in this work) are currently carried out by our collaborators in Dr. Marc Basson's lab (formerly at Michigan State University, now at University of North Dakota), by Bixi Zeng, Dr. Basson, and other members of his lab.

6.1 Introduction

The goal of this project is twofold: developing a novel protocol for blocking protein-protein interactions using 3D ligand-based virtual screening to mimic a protein epitope, and testing this approach to inhibit the interaction between two protein kinases involved in cancer metastasis, focal adhesion kinase (FAK) and protein kinase B (AKT1). In particular, this work presents a virtual screening protocol developed to identify candidate molecules with the potential to block the FAK-AKT1 interaction by performing 3D ligand-based virtual screening of more than 10 million molecules with drug-like properties to mimic regions within a seven-residue, primarily helical epitope of FAK known to bind AKT1. After an extensive literature search, and to the best of our knowledge, this is the first instance of using template-directed 3D screening to discover small organic molecules with a very similar volume and hydrophobic/polar atom distribution to a known protein epitope (part of the FAK protein) as the basis for outcompeting the intact protein (FAK) interaction with a protein partner (AKT1). Prior work on protein-protein interaction has focused either on docking small molecules to the protein surface to identifying potential inhibitors, creating constrained peptidyl mimics of the protein epitope, or performing high-throughput screening without structural information to identify molecules that interfere with the protein-protein interaction. Based on this epitope-mimic screening, a set of 13 small molecules is presented which are currently being experimentally tested for the potential to block the interaction between FAK and AKT1 proteins with the goal of further development as cancer therapeutics.

6.1.1 Blocking FAK-AKT1 activated cancer cell adhesion

Metastasis, the spreading of tumor cells from the primary site to different parts of the body, reduces the survival chance of cancer patients drastically (Sandru et al., 2014). Surgical procedures may further promote the metastasis process as viable tumor cells may be dislodged and enter the blood stream and circulate via the bloodstream or other fluids such as cerebrospinal fluids and lymph more readily (Yamaguchi et al., 2000; Hayashi et al., 1999). Thus, preventing cell adhesion of

dislodged tumor cells may decrease the risk of tumor metastasis in surgical procedures.

It has previously been shown that the direct interaction between these two kinases is facilitates the adhesion of tumor cells in high-pressure milieus that occur during surgical procedures in colon cancer resections, focal adhesion kinase (FAK) and the serine/threonine kinase AKT1 (Wang & Basson, 2011). While it has been shown that FAK-AKT1 interaction promotes pressure-induced tumor cell adhesion, the exact mechanism is not fully understood. It is speculated that the FAK-AKT1-mediated cell adhesion process is regulated by the affinity of β-integrin binding cell matrix proteins (Wang & Basson, 2011). However, it is known that the signaling pathway stimulated by FAK-AKT1 interaction in cancer cell adhesion is different from the signaling that occurs between non-cancerous, suspended cells. Hence, blocking the direct FAK-AKT1 interaction is likely not affecting regular cellular processes in non-cancerous cells (Zeng et al., 2017).

The FAK protein is composed of three distinct domains: the FERM domain, FAT domain, and kinase domain. The FAT domain is a required component in the intergrin-mediated signaling pathway and responsible for localizing FAK to focal adhesion sites (Thomas et al., 1999). The kinase domain is responsible for phosphorylating cytoskeletal proteins (α-actinin) to cross-link stress fibers (actomyosin) and to connect them to focal contacts (Thomas et al., 1999). The FERM domain is involved in the auto-inhibition of FAK (Thomas et al., 1999) as well as its direct interaction with activation by AKT1 (Basson et al., 2015).

Previous work identified a seven-residue peptide, LAHPPEE, that inhibits AKT1-FAK-mediated signaling by blocking pressure-induced FAK phosphorylation and the AKT1-FAK interaction (Zeng et al., 2017). The LAHPPEE peptide is corresponds to the wild type sequence of a short helix within the FAK FERM domain (residues 113-119). The FERM domain (residues 35-362 in FAK) has previously been shown to be important for FAK activity (Shiratsuchi & Basson, 2004). Experimental results showed that this short peptide was sufficient to pull down AKT1 in co-immunoprecipitation assays while a scrambled peptide containing the same residues (negative control) was not. In addition, the same LAHPPEE peptide interfered in pull-down of intact FAK by AKT1 (Zeng et al., 2017). The experiments by Zeng et al. (Zeng et al., 2017) showed that

the seven-residue peptide blocks the direct interaction between FAK and AKT1, and additionally the pressure-induced phosphorylation of the FAK Tyr397 (autophosphorylation) was also blocked by the peptide, while the kinase activity of AKT1 was preserved. Hence, it is hypothesized that LAHPPEE blocks the direct interaction between AKT1 and FAK while not affecting the catalytic mechanism of AKT1. In addition to probing the direct AKT1-FAK interaction, the researchers showed that the LAHPPEE peptide was also able to block cell adhesion of suspended cancer cells after surgical wound incision under pressure activation of FAK (Zeng et al., 2017).

6.1.2 Inhibiting protein-protein interactions with small molecules

Blocking protein-protein interactions (PPIs) with small molecules poses one of the biggest challenges in structural biology and drug discovery. What makes PPIs challenging targets of small molecule binders is the large surface areas involved in the protein-protein interaction. Also, the surface of protein-protein interfaces is relatively flat compared to typical small molecule binding pockets, which are rather deep and more concave. This results in fewer favorable interactions and enhanced mobility (higher entropy) of the interacting molecules. Further, protein-protein binding sites are largely hydrophobic, and the contributions of residues in protein-protein interfaces to the overall binding affinity is unevenly distributed: only a few individual residues (hot spots), sometimes organized within small patches (hot spot regions), contribute a relatively large fraction of the binding affinity (Keskin et al., 2005). Thus, defining hot spots of binding and designing ligands that specifically interact with them is important for outcompeting native protein-protein binding.

While a seven-residue peptide has been shown to effectively block the direct AKT1-FAK interaction involved in pressure-induced tumor metastasis (Zeng et al., 2017), the development and use of a peptides as therapeutics (drugs other than vaccines) has historically been largely unsuccessful. Compared to small molecules, peptides are generally cheaper to produce, better tolerated as drugs, and have metabolism that is more predictable. However, properties such as low membrane permeability, the tendency for aggregation, short half-lives, and chemical and physical instabilities (especially acid hydrolysis, leading to decomposition of the peptide upon oral

administration) pose challenges for using peptides as therapeutics. For instance, it is very rare that peptides do not violate Lipinski's Rule of 5 for drug-like molecules (Lipinski et al., 1997; Balgir & Sharma, 2017). While Rule-of-5 drug-like molecules do not necessarily satisfy all desired ADMET (absorption/administration, distribution, metabolism, excretion, and toxicity) properties for successful drugs, peptide drugs are known to have poor ADME properties (Falchi et al., 2014) and generally must be administered intravenously or parenterally rather than orally.

As outlined previously, blocking protein-protein interactions still remains a challenge, and this work proposes a new method of identifying small molecule inhibitors of PPIs using a 3D ligand-based approach. One advantage of the method pioneered here is that it does not require knowledge about the exact set of interactions between two protein binding partners, just knowledge of the epitope or surface on one of the molecules involved in binding. Secondly, by seeking to directly mimic part of a known structure rather than finding a molecule that will bind *to* the structure, this approach avoids the inaccuracy of docking scoring functions and evaluation of detailed electrostatic and solvent contributions to binding. For instance, the exact binding interactions between AKT1 and FAK are unknown. However, we know that a seven-residue peptidyl epitope in the helical region of the FAK FERM domain is sufficient for direct AKT1 interaction and that it competes with intact FAK for AKT1 binding. According to the proposed method, this seven-residue peptidyl region in FAK can then be used as a query molecule to identify small molecule mimics as candidate inhibitors of the FAK-AKT1 interaction.

Most experimental methods of probing PPIs involve the relatively cost intensive mutation of protein residues to identify hot spot residues and binding epitopes. Such knowledge can then be used to dock small molecules to hot spots as potential PPI inhibitors. However, this approach has been shown to be particularly challenging due to the high hydrophobicity of PPIs; docking for inhibitor discovery for PPIs tends to have high false positive and false negative rates (Zahiri et al., 2013). This work introduces an alternative method, which can be summarized in four steps: 1) identifying a small peptide region on one of the proteins involved in a PPI; 2) using the crystal structure (or modeling the structure) of the peptide as a template for identifying drug-like molecules

that mimic the 3D structure and chemistry of part or all of this peptide; 3) use 3D-ligand based virtual screening to prioritize a number of small molecule mimics of the peptide epitope; 4) assay those molecules for the ability to pull down the partner protein, block the PPI between the two intact proteins, etc.

Based on the strategy just described, this work presents a selection of small molecule candidates that mimic a peptide sequence within the FAK FERM domain, which has been found to be essential for AKT1 binding (Zeng et al., 2017). It is hypothesized that mimics of the peptide sequence of interest bind to AKT1 and block the direct AKT1-FAK interaction by competing with FAK for binding. The candidates prioritized from the screening are currently being tested by collaborators in Dr. Marc Basson's Lab (University of North Dakota School of Medicine, Grand Forks) for their binding affinity towards AKT1, the inhibition of FAK binding, and the effect on the AKT1-FAK signaling in pressure-induced tumor cell adhesion.

6.2 Methods

This work aimed to discover small molecules with drug-like properties (Lipinski et al., 1997) that block the FAK-AKT1 interaction involved in pressure-induced tumor growth. More specifically, we employ a ligand-based virtual screening protocol to identify small molecule mimics of a peptide sequence within the FAK FERM domain that is essential for AKT1 binding (Zeng et al., 2017). To enable the virtual screening of small molecule mimics, two query molecules were designed based on peptide sequences in the FAK FERM domain, as described in the following sections.

6.2.1 Modeling a peptide sequence from the FAK FERM domain for ligand-based screening

As a template for modeling, the crystal structure of the FAK FERM domain (PDB entry: 2al6; Ceccarelli et al., 2006) was used, from which the structure for residues 113-117 was extracted, corresponding to the amino acid sequence Leu-Ala-His-Pro-Pro, as shown in Figure 6.2. This also tests whether the apo conformation of the FAK peptide corresponds to its active state. In the following text, this peptide is referred to as "LAHPP query" molecule. To account for structural

flexibility in the screening query, structures reflecting eight rotameric positions of Leu residue in the initial peptide sequence were created using the backbone-dependent rotamer function (Shapovalov & Dunbrack, 2011) in PyMOL. The eight rotameric structures of LAHPP were checked to ensure there were no unfavorable van der Waals overlaps between the rotated side chains and other protein atoms (Figure 6.1). The other potentially rotatable residue, His, is tightly packed within the LAHPP native structure, and thus no alternative rotamers were possible without steric clashes. In addition, the N- and C-termini of the peptide structure were capped to neutralize them while providing correct valence, reflecting their state within the intact protein. Next, to enable chemical matching of polar atoms in the peptide during the drug-like screening, partial atom charges were computed and assigned to the LAHPP query molecule using molcharge from OpenEye QUACPAC (version 1.6.3.1; https://www.eyesopen.com/quacpac; OpenEye Scientific Software, Santa Fe, NM) with the AM1BCC force-field (Jakalian et al., 2002). After the partial atom charge assignment, protons have been removed from the C- and N-terminal nitrogen atoms of the peptide chain (which would be zwitterionic for a peptide but are uncharged in the intact protein); the charges of these nitrogen atoms have been set to -0.55, mimicking the charge of interior amide groups.

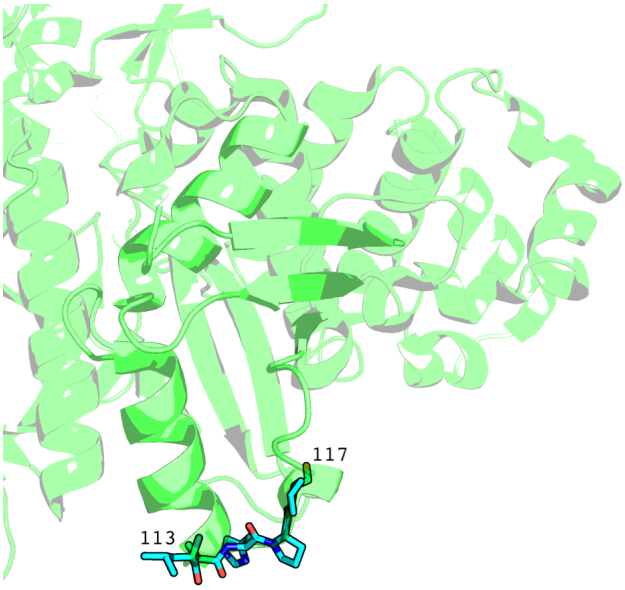


Figure 6.1: LAHPP query molecule from the FAK FERM domain. The FAK FERM domain (PDB entry: 2al6; Ceccarelli et al., 2006) is shown as cartoon representation in green, and the peptide target sequence, the LAHPP query molecule (residues 113-117 of FAK FERM), is shown in stick representation (cyan).

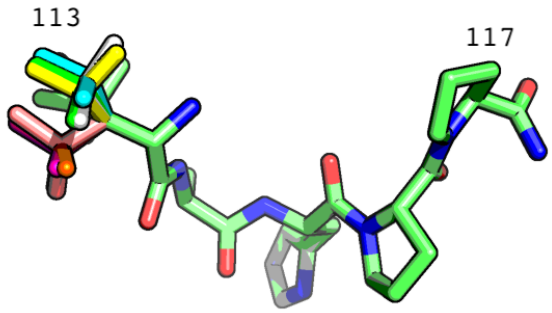


Figure 6.2: LAHPP query rotamers. Structure of the LAHPP query molecule extracted from the FAK FERM crystal structure (PDB entry: 2al6; Ceccarelli et al., 2006) showing the eight Leu113 rotamers. Hydrogen atoms are omitted for clarity.

Histidine protonation. In biological environments, the two imidazole nitrogens ($N\delta1$ and $N\epsilon2$) of a histidine residue can assume three different protonation states: protonation of both $N\delta1$ and $N\epsilon2$, protonation of $N\delta1$ only, and protonation of $N\epsilon2$ only (Li & Hong, 2011). In this work, only $N\delta1$ of histidine residues in the screening query molecules were protonated, because this state corresponds to the most prevalent tautomer of histidine at pH 7 (Bachovchin & Roberts, 1978; Figure 6.3). However, it shall be noted that a change in the protonation state would have little effect on the structural overlay, since the delta and epsilon nitrogens are negative in charge (ranging between -0.3 and -0.67); their charge will dominate the 3D overlay.

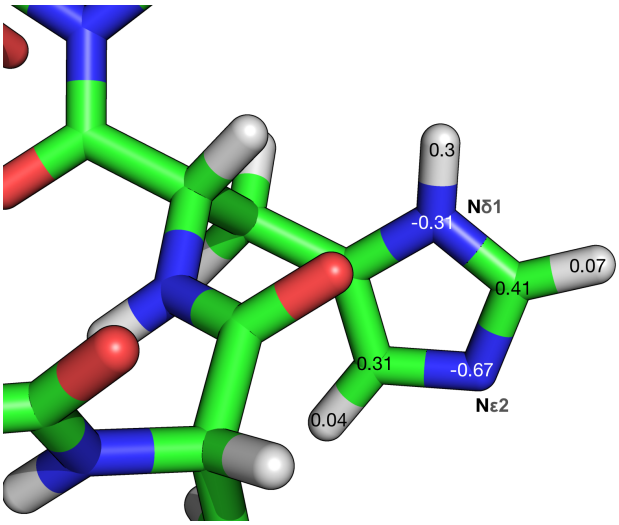


Figure 6.3: Partial charge assignment and protonation of the 115His residue in the LAHPP query peptide.

Constructing a screening query from an AKT1-binding two-site mutant peptide derived from FAK FERM. A second peptide query molecule was modeled for the computational screening of small molecules as AKT1 inhibitor candidates. This six-residue peptide (Ala-Ala-His-Pro-Ser-Glu) will be referred to as the AAHPSEE query molecule in the following text. This query peptide was a two-site mutant based on the LAHPPEE wild-type sequence (residues 113 to 119) in human FAK FERM (NCBI accession number: NP_722560.1; Jang et al., 2017). In the wildtype structure, the seven-residue peptide consists of a helical turn flexibly connected to a helical turn. One possibility is that together they form a continuous helical epitope upon interaction with AKT1. To test this, we designed the AAHPSEE sequence to have a higher degree of helicity whether a peptide with a greater degree of helicity, based on the very high helicity propensity for poly-Ala sequences and the ability of PS to form a less bent helix than PP. in the N-terminal (AAH) region, based on Sequery (Collawn et al., 1990) and Superpositional Structure Assignment (SSA; Craig et al., 1998) analysis were used to evaluate the actual helicity of sequences matching AAHPSEE in the Protein Data Bank (Zeng et al., 2017; Craig et al., 1998; Prevelige Jr & Fasman, 1989). If strong helicity in this region of FAK is important for AKT1 interaction, then the AAHPSEE peptide should would result in more effective competition with the wildtype FAK FERM domain for binding to AKT1. The peptide structure was built as an alpha helix in PyMOL (DeLano, 2002), with a low-energy (favorable) conformation chosen for the serine side chain in a similar orientation to the wild-type proline residue, and then it was energy-minimized by using the YASARA energy minimization server with YAMBER forcefield (Krieger et al., 2009). The resulting structure is shown in Figure 6.4. This peptide was used as a template to discover the most similar and somewhat more rigid small organic molecules (as more drug-like inhibitor candidates) for testing as inhibitors of the FAK-AKT1 interaction. After energy minimization, the AAHPSEE query molecule was handled similarly to the previously described LAHPP query preparation regarding the partial atom charge assignment and histidine treatment.

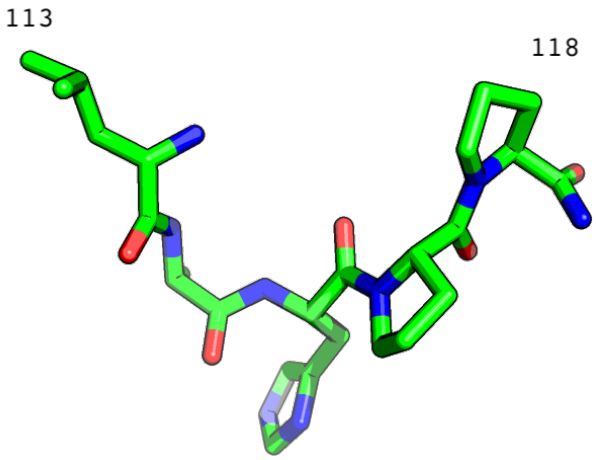


Figure 6.4: AAHPSEE query molecule. Structure of the AAHPSEE two-site mutant query molecule modeled after the LAHPPEE wild-type sequence in the FAK FERM domain.

6.2.2 Screening database

The 3D structure files of the molecules used in this work were downloaded from ZINC (Irwin & Shoichet, 2005) in MOL2 format. Only compounds listed as "off-the-shelf available" with drug-like properties ("Drugs Now" subset in ZINC) were considered. The total size of the screening database used in this work was 10,639,555 molecules before additional filtering criteria were applied. All molecules present in this database possessed the drug-like properties defined by Lipinski's Rule of 5 (Lipinski et al., 1997), which are:

- molecular weight between 150 and 500 g/mol;
- calculated octanol-water partition coefficient less than or equal to 5;
- 5 or fewer hydrogen bond donors and 10 or fewer hydrogen bond acceptors;
- polar surface area less than 150 Å^2;
- fewer than 8 rotatable bonds.

The single 3D conformations of the molecules downloaded from ZINC in MOL2 format have partial atom charges already assigned via semi-empirical quantum mechanical computations in AMSOL (Cramer & Truhlar, 1992).

6.2.3 Screening protocol

Conformer generation and sampling. ZINC provides 3D structures of molecules that were automatically generated from ASCII strings describing the structure of these molecules (*simplified molecular-input line-entry system*, abbreviated as SMILES; Weininger, 1988). Hence, being randomly generated, the single available 3D structure for each molecule is unlikely to be the same as the bioactive conformation that forms a complex with the binding partner. To increase the probability of matching the known conformation for LAHPP and the modeled helical conformation for the AAHPSEE peptide mimic, up to 200 favorable (low-energy) conformers of each ZINC molecule were generated using Omega (version 2.4.1; https://www.eyesopen.com/omega; OpenEye Scientific Software, Santa Fe, NM; Hawkins & Nicholls, 2012) with default settings. Using the default RMSD-based clustering, these conformations are both favorable in energy and distinct from each other.

Molecular overlays. To identify structural mimics of the FAK FERM peptide queries, the 3D structures of the drug-like database molecules were overlaid with the query molecules using ROCS (version 2.4.6; https://www.eyesopen.com/rocs; OpenEye Scientific Software, Santa Fe, NM; Hawkins et al., 2007) with default settings. The 3D overlays were then scored by the degree of volumetric similarity (ShapeTanimoto) and chemical similarity (ColorTanimoto) using an equal weighting of those two components (TanimotoCombo). The TanimotoCombo score ranges between 2 (perfect overlap) and 0 (no overlap). One important property of the Tanimoto metric is that a perfect score can only be achieved if the two molecules overlap entirely in contrast to the Tversky metric, which provides perfect scores for substructure matches.

PAINS removal. PAINS are pan-assay interference compounds, denoting compounds with chemical properties that interfere with many standard bioassays, such as aggregating around the protein, producing false spectroscopic signals, or reacting with common reagents. These compounds result in false positive assay results across many biological systems, and thus are removed from screening

results before assays are performed. After the top 500 LAHPP and top 500 AAHPSEE mimics were identified via ligand-based virtual screening, the corresponding SMILES strings of these candidates were obtained from ZINC. To check whether any of these molecules have been categorized as PAINS (pan-assay interference compounds) and hence have a high likelihood of yielding false positive assay results, these SMILES strings were entered into the *PAINS-Remover* server available at http://cbligand.org/PAINS/ (Baell & Holloway, 2010). Consequently, molecules flagged as PAINS were removed from further analyses (for instance, ZINC02157352, ZINC02368133, ZINC02102847, ZINC02786049, and ZINC01692153).

Candidate prioritization. The final prioritization of molecules for experimental assays was based on visual inspection in PyMOL (DeLano, 2002). Molecules were selected to represent different scaffolds to provide alternatives in case compounds lack bioavailability or are unstable. Further selection criteria included good hydrophobic group and polar group correspondence to the query peptide as well as proper overall matching of the surface-exposed moieties of the peptide that could interact with AKT1.

Comparing molecules based on physicochemical properties. Based on preliminary results from the Basson lab, additional molecules were selected as negative controls for further tests of molecules that were identified as potentially active. These additional molecules were similar in physicochemical properties and composition to molecules prioritized for and tested in experimental assays but should show little correspondence with the molecules to be tested when overlaid in 3D. To compare molecules based on physicochemical properties such as molecular weight, number of rotatable bonds, etc., the physicochemical information of these were downloaded for all 10,639,555 "Drugs Now" (commercially available off-the-shelf) molecules from ZINC (see *Methods: Screening database*). To bring all features onto the same scale, each feature was standardized to have zero mean and unit variance properties across the database. The standardized features were computed as follows:

$$x_i' = \frac{x_i - \bar{x_i}}{s_{x_i}},$$

where x_i represents the ith feature, $\bar{x_i}$ is the feature average, and s_{x_i} is the standard deviation of the ith feature.

To compute the pair-wise distance between a query molecule (x) and a molecule (y) from the 10,639,555 Drugs Now database, molecule (x), the Euclidean distance metric was used on the standardized feature columns over the m features:

$$d(x, y) = \sqrt{\sum_{j=1}^{m}(x_j - y_j)^2}.$$

6.2.4 Diagrams and graphs

The 3D structures of molecules were visualized and rendered using PyMOL (version 1.8; https://pymol.org/; DeLano, 2002). Data plots were created using matplotlib (version 2.0.2; https://matplotlib.org; Hunter, 2007) and Python (https://www.python.org; Van Rossum, 2007). Flowcharts and other graphics were created using PowerPoint (version 15.39; https://products.office.com/en-us/powerpoint) and OmniGraffle (version 7.5; https://www.omnigroup.com/omnigraffle).

6.3 Results and Discussion

6.3.1 Discovering small molecule mimics of FAK FERM peptides that block AKT1-FAK interaction

To identify small molecule mimics of the LAHPPEE peptide in the FAK FERM domain as inhibitors of the protein-protein interactions between FAK FERM and AKT1, we designed a 3D epitope-based based virtual screening protocol. This work describes the first instance of using ligand-based virtual screening to identify mimics of an intact protein epitope. The screening workflow is summarized in Figure 6.5, and the results are described in the following subsections.

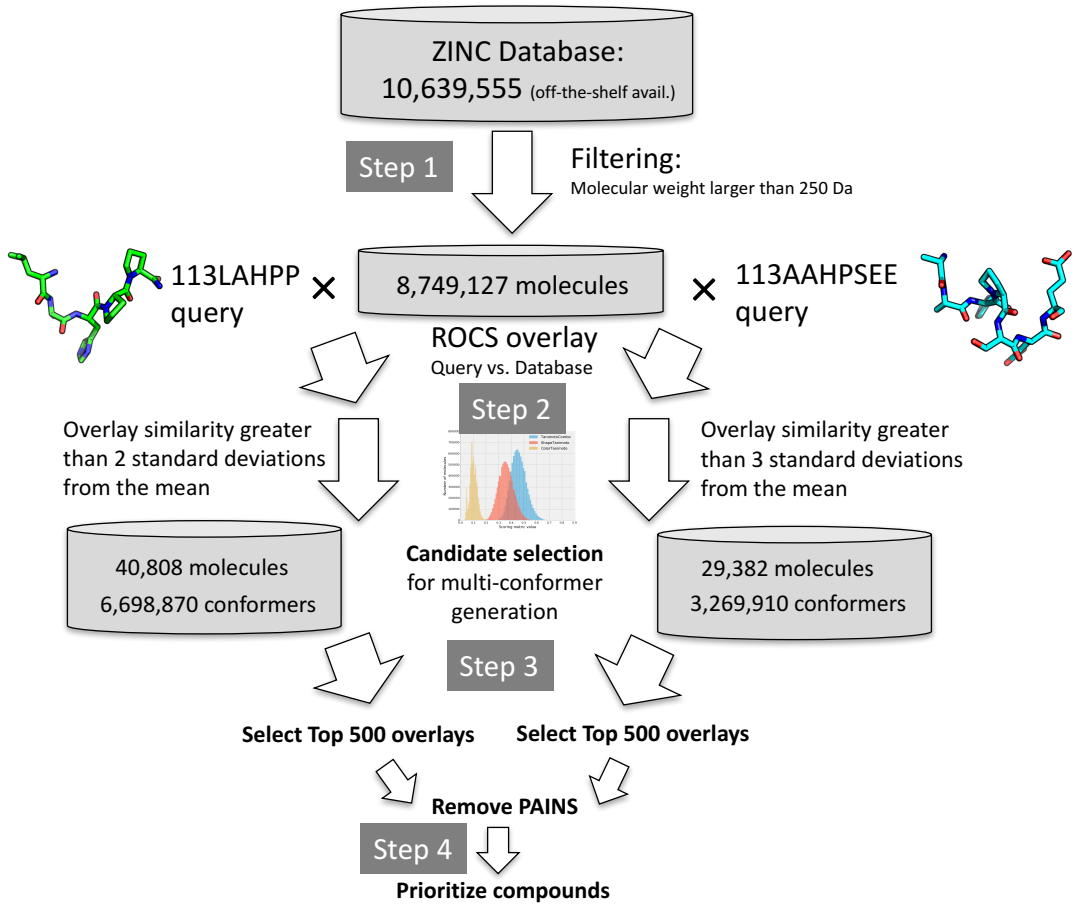


Figure 6.5: Flowchart summarizing the individual steps in the ligand-based screening for identifying FAK FERM peptide mimics.

6.3.2 Step 1: Pre-filtering the screening database

The initial screening database downloaded from ZINC (Irwin & Shoichet, 2005) consisted of MOL2 files representing the 3D structures of 10,639,555 commercially available molecules with drug-like properties, according to Lipinski's Rule of 5 (Lipinski et al., 1997), were downloaded from ZINC (Irwin & Shoichet, 2005). Considering the relatively large sizes of the five-residue LAHPP and seven-residue AAHPSEE peptides, a weight cut-off was applied so that only those molecules with a molecular weight above 250 Da were considered. The removal of very small molecules that were insufficient in size to mimic the peptide queries resulted in 8,749,127 molecules that were to be considered in the remaining steps of the virtual screening pipeline.

6.3.3 Step 2: Single-conformer overlays for candidate filtering

Generating up to 200 favorable-energy conformers of 8,749,127 database molecules, and overlaying those with the two peptide query molecules, was not computationally feasible within the time-frame of this project. (Similar to a project in a pharmaceutical company, our collaborators were waiting for compounds to test and needed to test them within weeks rather than months.) Instead, the LAHPP and AAHPSEE queries were overlaid with the 8,749,127 single-conformer structures from ZINC to identify a smaller subset of the most promising compounds based on the structural overlay with the query molecule. The average TanimotoCombo similarity score of all 8,749,127 LAHPP overlays was 0.459, and only those molecules with a similarity score (TanimotoCombo) of 3 standard deviations above the mean were selected. This score-based selection criterion resulted in a subset of 40,808 molecules (Figure 6.6a). Since the larger peptide query, AAHPSEE, is naturally harder

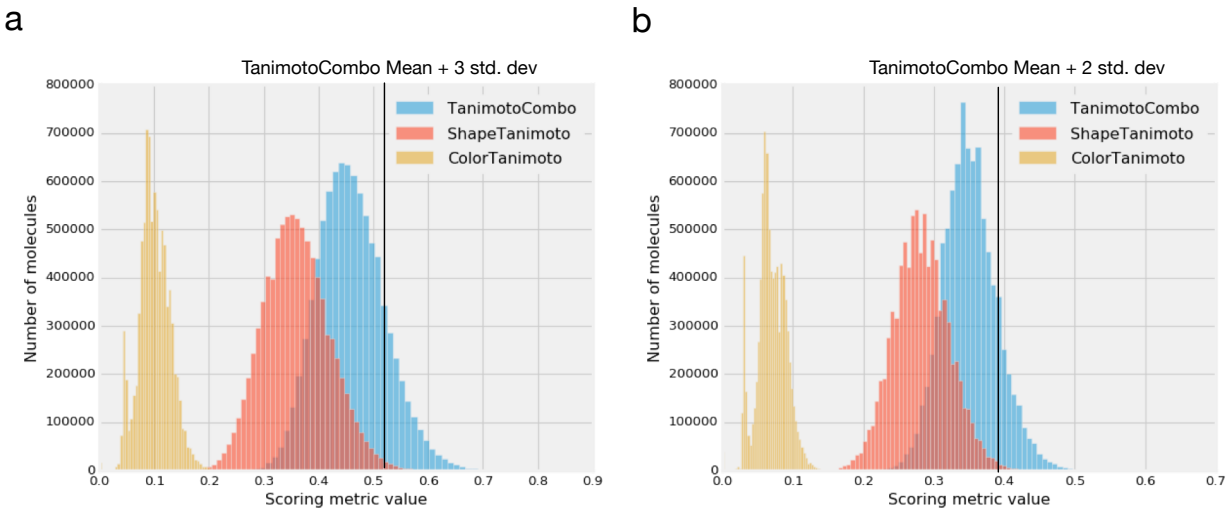


Figure 6.6: Similarity score distributions from single-conformer overlays. Subpanel (a) shows the distribution of similarity scores of the 8,749,127 ZINC molecules overlaid with LAHPP, and subpanel (b) shows the similarity scores of the those molecules overlaid with AAHPSEE.

to mimic with *small* molecules, the average TanimotoCombo similarity scores were expectedly lower compared to the LAHPP overlays, averaging 0.432. Consequently, a less stringent inclusion criterion was applied to obtain 29,509 molecules (Figure 6.6b) of candidate molecules for the next round in the screening workflow, by setting the TanimotoCombo cut-off to two standard deviations

above the mean (Figure 6.6b). The goal was to provide a sampling of compounds with good chemistry and volumetric match to the query peptides, for further evaluation by fully flexible alignment of all favorable conformers of the small molecules with the query peptides. Relatively low average TanimotoCombo scores were expected for the single conformer matches (considering that a perfect match corresponds to a TanimotoCombo score of 2), due to the focus on non-peptide matches, and also the presence of few molecules with molecular weight greater than 500 Da in the ZINC database screening. These results demonstrate that even 8.7 molecules cover only a tiny fraction of all chemically possible molecules of this weight.

6.3.4 Step 3: Multi-conformer overlays to select most similar peptide mimics

Considering that the conformations of the peptide queries LAHPP and AAHPSEE were modeled based on the crystal structure of the FAK FERM domain (see Methods), which is a known low-energy conformation that is likely similar to the structure FAK exhibits in complex with AKT1, conformations of the query peptides were not sampled. However, up to 200 low-energy conformation of the candidate subsets of 40,808 molecules (LAHPP query) and 29,509 molecules (AAHPSEE query) were sampled to improve the 3D alignment and scoring of similarity relative to the single-conformer overlays from the previous step. Following the multi-conformer overlays with the peptide queries, the top 500 molecule mimics of the LAHPP query and the top 500 molecules matching the AAHPSEE query were selected based on the TanimotoCombo similarity score.

6.3.5 Step 4: Prioritization for experimental assays

After removing PAINS, the top 500 molecule overlays with both the LAHPP query and the AAHPSEE query were visually inspected in PyMOL (DeLano, 2002) to select compounds that represented several different scaffolds (to provide alternatives should one class of compounds lack stability or bioavailability, for instance), were isosteric with the peptide backbone as well as side chains, matched the surface-exposed moeities of the peptide available to interact with AKT1, and exhibited polar group versus hydrophobic group correspondence to the query peptide throughout

the molecule (rather than matching one end of the peptide well and the other end poorly). Based on the inspection of these overlays, a diverse set of 11 LAHPP mimics were selected for assaying in the Basson lab (Figure 6.7a) as well as two AAHPSEE mimics (Figure 6.7b).

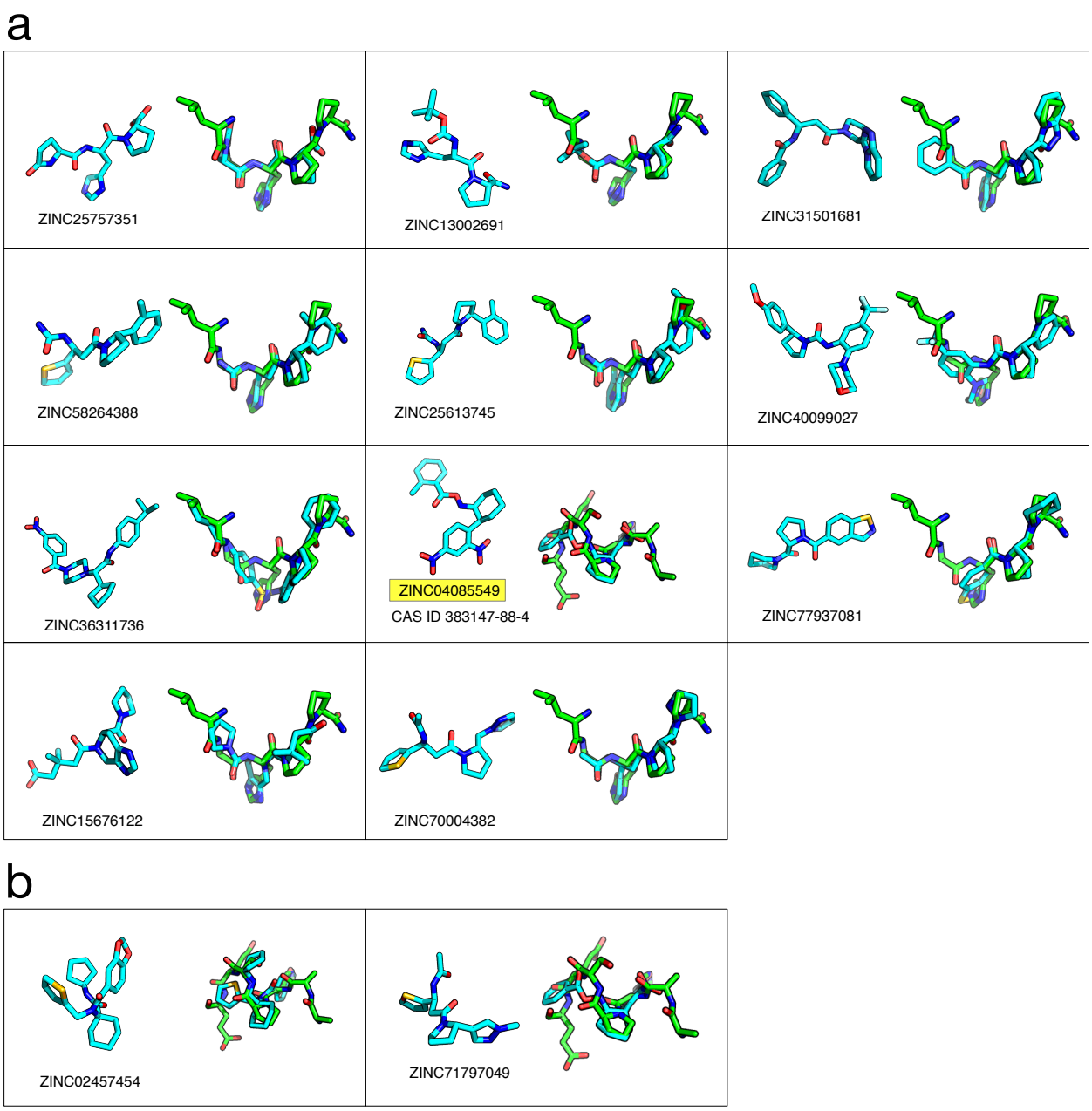


Figure 6.7: FAK peptide drug-like mimics prioritized for experimental assays. Top LAHPP (a) and AAHPSEE (b) mimics selected for experimental assays. The peptides are shown in green (LAHPP and AAHPSE) overlaid with the selected peptide mimics (cyan). The same peptide regions were not matched by all compounds since most the compounds are the size of about residues in the peptide.

6.3.6 Step 5: Selecting candidates for negative controls based on bearing similar physicochemical properties to the peptides

Based on preliminary experimental results from Dr. Basson's lab (measuring cellular adhesion *in vitro* and FAK phosphorylation with pressure at nanomolar concentrations) that looked promising for two of the compounds selected in step 4, ZINC31501681 (N-[(1S)-3-oxo-1-phenyl-3-[(2S)-2-([1,2,4]triazolo[4,3-a]pyridin-3-yl)pyrrolidin-1-yl]propyl]benzamid) and ZINC58264388 ([(1S)-3-[(2S)-2-(o-tolyl)pyrrolidin-1-yl]-3-oxo-1-(2-thienyl)propyl]urea), the need for appropriate negative controls drove the selection of additional molecules from ZINC based on showing physicochemical similarity to these two ZINC compounds (Figure 6.8). The idea was to consider molecules that were

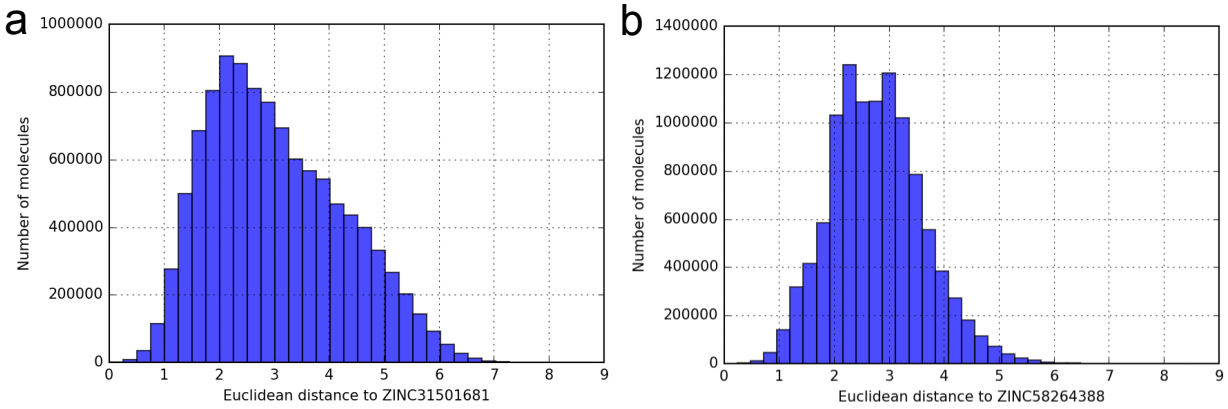


Figure 6.8: Distribution of physicochemical distances for [(1S)-3-[(2S)-2-(o-tolyl)pyrrolidin-1-yl]-3-oxo-1-(2-thienyl)propyl]urea (ZINC58264388) and N-[(1S)-3-oxo-1-phenyl-3-[(2S)-2-([1,2,4]triazolo[4,3-a]pyridin-3-yl)pyrrolidin-1-yl]propyl]benzamid (ZINC31501681). The histogram shows the distribution of distances of molecules in ZINC in six-dimensional space (LogP, NRB, tPSA, HBD, HBA, and MWT) from (a) ZINC31501681 and (b) ZINC58264388.

very similar in composition and properties to the two putative active compounds, while selecting molecules that were "scrambled," showing little correspondence with the two active compounds when overlayed in 3D. The properties of interest were obtained from the ZINC database at reference pH 7 (Table 6.1). Note that the first few matches with exactly the same molecular weight are, in fact, stereoisomers of ZINC31501681 and ZINC58264388.

Multiple compounds, based on those listed as most similar to the peptide query molecules in Table 6.2 and Table 6.3, were suggested as negative controls for experimental assays performed in the Basson lab, providing a selection based on current vendor availability (structures of this

Table 6.1: Physicochemical properties of N-[(1S)-3-oxo-1-phenyl-3-[(2S)-2-([1,2,4]triazolo[4,3-a]pyridin-3-yl)pyrrolidin-1-yl]propyl]benzamid (ZINC31501681) and [(1S)-3-[(2S)-2-(o-tolyl)pyrrolidin-1-yl]-3-oxo-1-(2-thienyl)propyl]urea (ZINC58264388).

	ZINC31501681	ZINC58264388
xlogP value (octanol/water partition coefficient) (LogP)	2.35	2.86
Number of rotatable bonds (NRB)	6	5
Polar surface area in $Å^2$ (tPSA)	80	75
Number of hydrogen bond donor atoms (HBD)	1	3
Number of hydrogen bond acceptor atoms (HBA)	7	5
Molecular weight in g/mol (MWT)	439.519	357.479

molecules are shown in Figure 6.9 and Figure 6.10. Also, compounds with a proline linked by a peptide bond to a phenyl group were avoided, because they could be true positives in experimental assays, mimicking the epitope in ZINC31501681 in its binding or interference with AKT interaction.

Table 6.2: Ten molecules most similar to N-[(1S)-3-oxo-1-phenyl-3-[(2S)-2-([1,2,4]triazolo[4,3-a]pyridin-3-yl)pyrrolidin-1-yl]propyl]benzamid (ZINC31501681) based on physicochemical properties.

	ZINC ID	LogP	NRB	tPSA	HBD	HBA	MWT	Distance
1	ZINC31501667	2.35	6	80	1	7	439.519	0
2	ZINC31501672	2.35	6	80	1	7	439.519	0
3	ZINC31501676	2.35	6	80	1	7	439.519	0
4	ZINC09263673	2.37	6	80	1	7	439.516	0.01
5	ZINC09327797	2.32	6	80	1	7	439.516	0.02
6	ZINC77973213	2.35	6	80	1	7	437.447	0.03
7	ZINC36398207	2.38	6	80	1	7	441.582	0.03
8	ZINC36398208	2.38	6	80	1	7	441.582	0.03
9	ZINC14543207	2.32	6	80	1	7	441.582	0.03
10	ZINC77973181	2.39	6	80	1	7	437.447	0.04

By excluding molecules that are most-similar ZINC31501681 or are stereoisomers, the following molecules were proposed as negative controls for

- ZINC31501681: ZINC14543207 (or isomers ZINC36398207 and ZINC36398208), ZINC77973213, or ZINC77973213 in that order, based on having high physicochemical similarity but little similarity in adjacent functional groups to ZINC31501681.
- ZINC58264388: One of the three closely related structures ZINC46869202, ZINC46869200,

Table 6.3: Ten molecules most similar to [(1S)-3-[(2S)-2-(o-tolyl)pyrrolidin-1-yl]-3-oxo-1-(2-thienyl)propyl]urea (ZINC58264388) based on physicochemical properties.

	ZINC ID	LogP	NRB	tPSA	HBD	HBA	MWT	Distance
1	ZINC58264389	2.86	5	75	3	5	357.479	0
2	ZINC58264391	2.86	5	75	3	5	357.479	0
3	ZINC58264388	2.86	5	75	3	5	357.479	0
4	ZINC58264390	2.86	5	75	3	5	357.479	0
5	ZINC09469794	2.80	5	75	3	5	357.841	0.04
6	ZINC58341565	2.91	5	75	3	5	355.482	0.04
7	ZINC58341566	2.91	5	75	3	5	355.482	0.04
8	ZINC46869202	2.86	5	74	3	5	355.825	0.05
9	ZINC46869200	2.86	5	74	3	5	355.825	0.05
10	ZINC46869359	2.88	5	74	3	5	355.825	0.05

or ZINC46869359 and ZINC58341565 or ZINC58341566 (again, two closely related structures).

6.4 Conclusions and Future Directions

Preliminary results from the experimental assays performed in the Basson lab indicate that at least one of the peptide mimics blocks FAK-AKT1 interaction *in vitro* and is showing promising preliminary *in vivo* results for inhibition of tumor growth in mice. Upon completion of the experiments, we will publish this approach and the screening and experimental results in collaboration with the Basson lab. From our perspective, we are excited because gaining positive experimental results within a relatively limited amount of screening and analysis time (2-3 months) provides proof-of-concept that a protein-protein interaction inhibitor can be discovered efficiently by using ligand-based screening methods to find mimics of an intact protein epitope. This has not been done before, and to our knowledge the use of physicochemical similarity to provide negative controls is also new and useful.

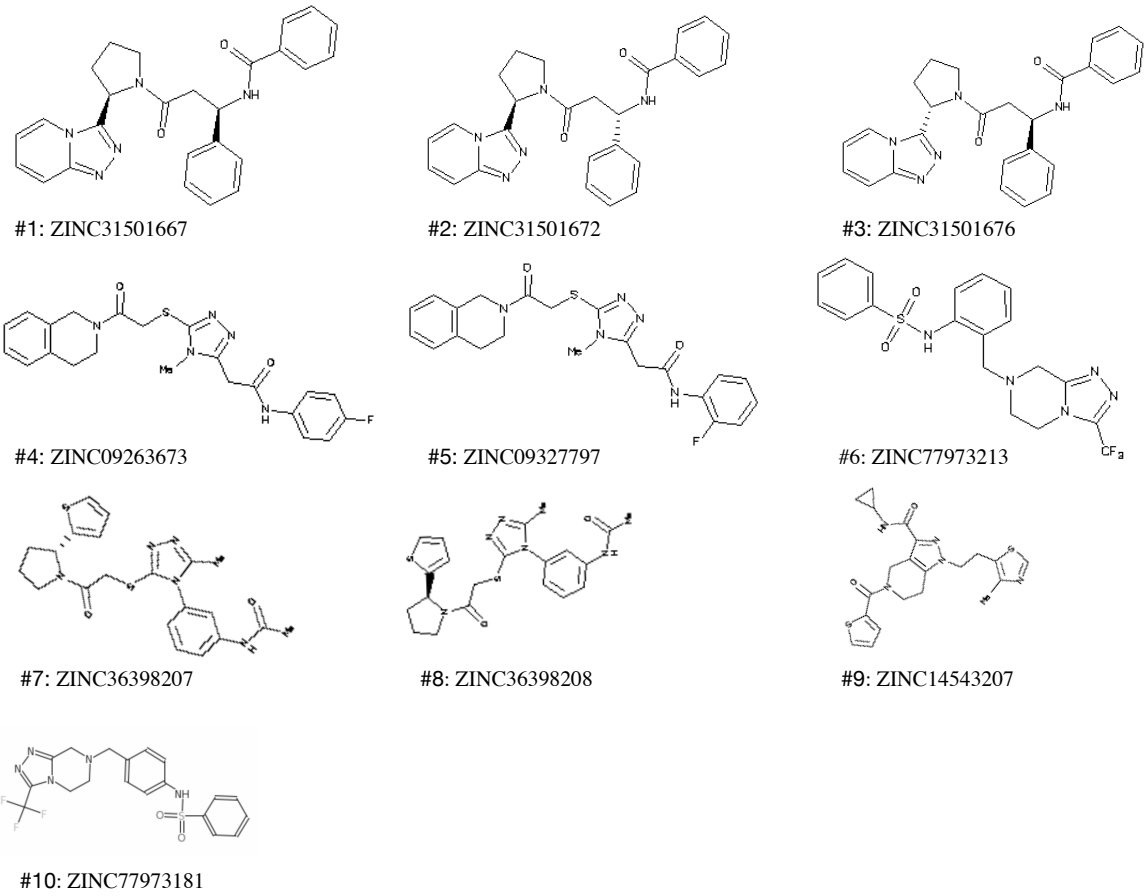
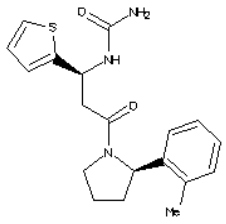
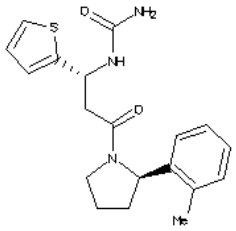


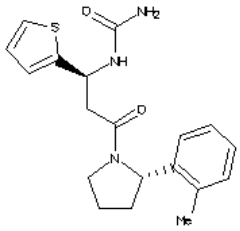
Figure 6.9: Top 10 ZINC molecules with physicochemical properties most similar to ZINC31501681.



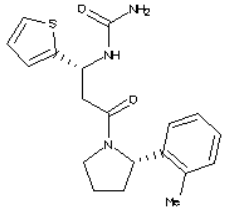
#1: ZINC58264389



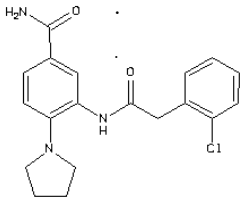
#2: ZINC58264391



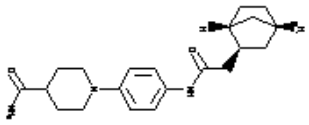
#3: ZINC58264388



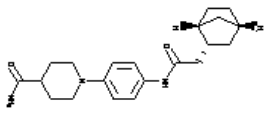
#4: ZINC58264390



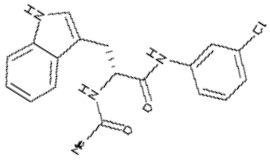
#5: ZINC09469794



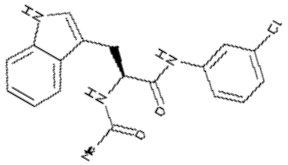
#6: ZINC58341565



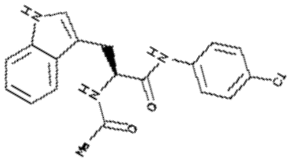
#7: ZINC58341566



#8: ZINC46869202



#9: ZINC46869200



#10: ZINC46869359

Figure 6.10: Top 10 ZINC molecules with physicochemical properties most similar to ZINC58264388.

CHAPTER 7

CONCLUSIONS

In this dissertation, I presented several different investigations and applications to derive new insights into fundamental principles that govern molecular recognition. This concluding chapter highlights the main contributions of this work. In particular, this thesis sought out to address the following questions:

- Are protein-ligand interfaces polarized, and if such is the case, what are the interaction patterns that characterize cognate ligand binding?
- Is protein-ligand binding interface rigidification a hallmark of protein-ligand binding, and if so, can it be captured computationally to predict native protein-ligand complexes?
- In the absence of structural data for a receptor, can the success rate of 3D ligand-based virtual screening to discover small molecule inhibitors be augmented by specific hypotheses about functional groups to discover potent inhibitors of biological processes?
- Can supervised machine learning be used to capture information in relatively small and noisy data sets from experimental assays to further characterize the molecular features that are involved in and essential for blocking biological signaling pathways?
- Is 3D ligand-based virtual screening a viable alternative to docking small molecules to protein-protein binding interfaces for identifying small molecule inhibitors of protein-protein interactions in large databases?

Throughout this thesis, many protein-ligand complexes were analyzed, computationally as well as visually. The inspection of these complexes led to an interesting observation that sparked the formulation of the following hypothesis: "donor groups on proteins are preferred in H-bonding to biological ligands." This hypothesis was tested in Chapter 2, and the detailed analysis of hydrogen bonding networks across a dataset of non-homologous proteins bound to their biological ligands revealed that proteins donate twice as many H-bonds as they accept from ligands. The opposite trend was observed for their biological ligands. Interestingly, this enrichment of hydrogen bond donors in proteins in forming non-covalent interactions with their native ligands cannot simply be

explained by atom type occurrences in binding sites: in fact, protein binding sites have almost twice as many electron lone pairs available to accept H-bonds versus polar hydrogens available to form H-bonds. The most likely explanation for this newly discovered phenomenon is that this intermolecular H-bonding polarity has evolved as a determinant for ligand selectivity. Further, the results presented in this chapter demonstrated that the general trends observed in the intermolecular H-bonds across the 136 non-homologous protein-small molecule complexes can be used to predict near-native protein-ligand interactions in individual complexes (Protein Recognition Index).

The novel insights and discoveries presented in this chapter will likely have implications for studies of protein-ligand interaction in general – for instance, guiding ligand design and protein mutagenesis. Moreover, by providing the software framework developed in this study to other researchers, scientists will be able to rigorously define H-bonds patterns in protein-ligand binding interfaces and also compute a Protein Recognition Index (PRI) to enable the prediction of small molecule binding relative to the cognate protein structure. For instance, future applications may include the testing of the hypothesis that mutating protein side chains and ligand atoms to optimize the PRI score improves ligand selectivity and binding affinity of receptor agonists or inhibitors versus cognate ligands. As described in this study, the PRI is uncorrelated to other scoring functions for protein-ligand docking and thus provides a new feature for the improvement of existing docking software as well. For instance, the PRI score, which is based on the chemical preference of atom types participating in H-bonding can be combined with the SiteInterlock score (Chapter 3), which provides an orthogonal method of predicting native protein-ligand complexes by measuring the coupling of interactions. Also, it would be interesting to investigate whether hot spot regions in protein-protein binding interfaces correspond to regions enriched in amino acids that frequently participate in H-bonding (for instance, Lys, Arg, Glu, and Asp), which can provide useful information for the identification of hotspot residues in uncharacterized protein-protein binding sites.

In Chapter 3, a computational study was carried out to test the hypothesis that complex formation rigidifies the protein-ligand binding interface. This hypothesis was based on ideas and

evidence from X-ray crystallographic experiments – that a number of proteins only crystallize in the presence of their cognate ligand presumably due to decreasing conformational variability that impedes crystallization – and thermofluor assays, where cognate ligand binding increases the melting temperature of proteins. The fact that proteins bound to their biological ligands have an increased chance to crystallize suggests that the bound form *stabilizes* the complex. To test the hypothesis that ligand binding rigidifies the protein interface, protein structures were compared before and after the computational deletion of the ligand from the binding pocket. In approximately 90% of the structures analyzed, increased rigidity was detected by the ProFlex method in the binding interface if a ligand was bound. Based on this observation, a rigidity-based scoring function (SiteInterlock score) was designed that was able to predict near-native protein-ligand complexes from a set of docking poses with an accuracy comparable to state-of-the-art docking scoring functions. Interestingly, SiteInterlock score performed robustly for both holo- and apo-complexes, in contrast to other widely used scoring functions. All in all, the results suggest that protein-ligand interfacial rigidification and the resulting SiteInterlock detection of cognate binding is a robust and reliable scoring function with performance comparable to existing scoring functions.

More interestingly though, and in addition to developing yet another scoring function for protein-ligand pose selection in docking studies, correlation analysis showed that predictions by SiteInterlock score are virtually uncorrelated to other scoring functions, which suggests that SiteInterlock captures information that is not considered in other methods. For instance, existing scoring functions regard individual interactions between proteins and their ligands as additive terms, whereas SiteInterlock considers the coupling between interactions. The rigidity index computed by SiteInterlock thus provides new information that cannot only be used as a standalone feature for pose selection but also has the potential to improve existing methods – for example, through the development of new ensemble scoring functions (a combination of multiple, different scoring functions to obtain better predictive performance), which is an interesting topic for future research. Another interesting future direction is using the rigidity information as computed in this work to predict hot spots in protein binding sites, which is currently investigated by Jiaxing Chen, a

fellow graduate student in Dr. Leslie Kuhn's lab. Altogether, this is the first instance where the rigidification of binding interfaces has been studied theoretically and computationally as a predictor of protein-ligand complex formation. And not only does this work provide new insights into how proteins and ligands interact, but a software package for computing the SiteInterlock score has been made available to other researchers under an open source license.

While Chapter 2 and 3 explored interfacial interactions, Chapter 4 focused on the analysis of functional group patterns in ligands that are linked to biological activity, for cases when the structure is not known. Predicting the biological activity of ligands in the absence of the receptor structure is a common challenge in inhibitor and drug discovery, and Chapter 4 presented the development of user-friendly and freely available software that enables the hypothesis-driven discovery of active compounds from databases of millions of commercially available molecules.

Many methods exist that enable ligand-based virtual screening, which involves a similarity-based search of mimics of a known active ligand. However, a shortcoming of current implementations is that users can only perform a brute-force approach (screening and scoring all molecules) on a given molecular database. Without the possibility to incorporate experimental knowledge, such an approach can naturally lead to a high number of false positives and false negatives when applied to large databases. The hypothesis-based protocol developed and presented in Chapter 4, Screenlamp, provides a flexible, modular solution that lets researchers leverage information from experiments, such as the importance of specific functional groups required for biological activity, to augment the search of active compounds (e.g., inhibitor candidates). It also allows the incorporation of alternative workflows and tools. The Screenlamp software has been made available to other researchers under a permissive open-source license and provides a user-friendly interface for large-scale virtual screening of millions of molecules based on custom filtering criteria in combination with existing, well-validated methods for 3D conformer sampling and 3D ligand-based molecular overlays.

Applied to aquatic invasive species control, the hypothesis-based virtual screening via Screenlamp led to the discovery of two potent inhibitors of a GPCR-mediated signaling pathway. Experiments showed that these inhibitors nullified the biological response towards the cognate pheromone

ligand of that receptor. The analysis of the approximately 300 experimentally tested molecules revealed that this hypothesis-driven framework led to a selection of molecules that are more bioactive than selections that would have been obtained by screening for overall shape similarity and chemical similarity alone. The successful use of Screenlamp in this project should motivate other researchers to incorporate such hypothesis-driven protocols in inhibitor studies, for instance, in the early stages of drug discovery.

While the Screenlamp toolkit presented in Chapter 4 provides a flexible framework for hypothesis-based screening, experimental knowledge is required to formulate such hypotheses. In the study presented in Chapter 5, protocols for supervised machine learning were developed, to automate the discovery of functional group patterns that are associated with biological activity. While existing methods focus on the analysis of abstract pharmacophores, that is, specific encodings of molecular features that are not easily interpretable by humans, the methods presented in this thesis provide intuitive insights into the presence and the position of chemical groups that are characteristic of active molecules. For instance, the use of these techniques led to interesting insights, namely, that mimicking the sulfate group in a cognate pheromone ligand of a GPCR receptor was an important feature and accounted for 58% of the activity of of bioactive molecules in the invasive species control project described in Chapter 4. It is expected that those methods can provide useful guidance in the discovery and design of bioactive molecules in many other research applications. Also, it shall be noted that the main focus of these protocols was not the mere prediction of bioactive molecules but rather the identification of structural and chemical features of bioactivity. While only the positions of functional groups were provided as input, together with experimental data of the molecules' activity, future work may focus on improving the performance of predictive models even further, for instance, by including physicochemical properties of molecules, too.

Chapter 6 presented a new 3D virtual screening protocol to identify small molecule mimics of a protein epitope to block protein-protein interactions. This novel 3D epitope-based screening approach described in Chapter 6 can be understood as an extension and advancement of 3D ligand-

based virtual screening for the small molecule candidates presented in Chapter 4 and 5. Finding small molecule inhibitors of protein-protein interactions is a notoriously difficult task. One of the reasons that make this problem particularly challenging is the large surface area of protein-protein interactions that must be disrupted by a small molecule. Also, compared to small molecule binding pockets, protein-protein binding interfaces are relatively flat and mostly hydrophobic and provide fewer opportunities for ligand-selective interactions. This results in fewer favorable, polar interactions and higher entropy of the interacting molecules, due to enhanced mobility. Thus, the most promising approach for identifying or designing small molecule ligands that outcompete a native protein-protein binding partner is to target a smaller number of regions that contribute a relatively large fraction of the binding affinity, the so-called hot spots.

Based on experimental evidence provided by collaborators (Dr. Marc Basson's lab), a small peptide region involved in and required for the direct interaction between AKT1 and FAK was modeled as a template for ligand-based virtual screening. This 3D epitope-based screening of more than 10 million drug-like molecules led to the selection of a small set of diverse molecules as candidate inhibitors of the interaction between two protein kinases, FAK and AKT1, which is involved in cancer metastasis. Preliminary results from binding assays and cancer adhesion studies in mice (carried out by collaborators in Dr. Marc Basson's lab) indicate that one of the discovered molecules is showing a promising inhibitory effect. While further experimental studies are underway, the 3D epitope-based approach presented in Chapter 6 appears to be a promising alternative to computationally more intensive alternatives such as docking small molecules to protein-protein binding interfaces, which is commonly known to have high false positive and false negative rates. Further, the described method does not require knowledge about the three-dimensional structure of the complex, only the 3D structure and binding epitope of one of the two proteins, which makes it applicable to more protein-protein interfaces.

Overall, this thesis presents several new insights into the concepts of molecular recognition, including the polarity of H-bonds in protein-ligand interfaces and the rigidification of interaction interfaces upon small molecule binding. Furthermore, new methods have been developed for

identifying active molecules as inhibitors or agonists of biological processes where the three-dimensional structure of the protein binding partner is unknown. These methods proved to be successful in real-world applications and led to pioneering work in aquatic invasive species control. Beyond the intellectual merits of this work, the computational tools developed in this work are being made available to other researchers under open source licenses, so that they can lead to further advancements in the fields of experimental and computational biology, and drug discovery.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahmed, Aqeel, Richard D Smith, Jordan J Clark, James B Dunbar Jr & Heather A Carlson. 2014. Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures. *Nucleic Acids Research* 43(D1). 465–469.

Aiello, Anna, Sabina Carbonelli, Giuseppe Esposito, Ernesto Fattorusso, Teresa Iuvone & Marialuisa Menna. 2000. Novel bioactive sulfated alkene and alkanes from the Mediterranean ascidian Halocynthia papillosa. *Journal of Natural Products* 63(11). 1590–1592.

Allen, Bryce K, Saurabh Mehta, Stewart WJ Ember, Ernst Schonbrunn, Nagi Ayad & Stephan C Schürer. 2015a. Large-scale computational screening identifies first in class multitarget inhibitor of EGFR kinase and BRD4. *Scientific Reports* 5.

Allen, Frank H. 2002. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B: Structural Science* 58(3). 380–388.

Allen, William J, Trent E Balius, Sudipto Mukherjee, Scott R Brozell, Demetri T Moustakas, P Therese Lang, David A Case, Irwin D Kuntz & Robert C Rizzo. 2015b. DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry* 36(15). 1132–1156.

Almela, Maria Jesus, Sonia Lozano, Joël Lelièvre, Gonzalo Colmenarejo, José Miguel Coterón, Janneth Rodrigues, Carolina Gonzalez & Esperanza Herreros. 2015. A new set of chemical starting points with Plasmodium falciparum transmission-blocking potential for antimalarial drug discovery. *PloS One* 10(8). e0135139.

Altschul, Stephen F, Warren Gish, Webb Miller, Eugene W Myers & David J Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3). 403–410.

Arkin, Michelle R & James A Wells. 2004. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery* 3(4). 301–317.

Bachovchin, William W & John D Roberts. 1978. Nitrogen-15 nuclear magnetic resonance spectroscopy. The state of histidine in the catalytic triad of alpha-lytic protease. Implications for the charge-relay mechanism of peptide-bond cleavage by serine proteases. *Journal of the American Chemical Society* 100(26). 8041–8047.

Baell, Jonathan B & Georgina A Holloway. 2010. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry* 53(7). 2719–2740.

Bahar, I, A Wallqvist, DG Covell & RL Jernigan. 1998. Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry* 37(4). 1067–1075.

Bahar, Ivet, Ali Rana Atilgan & Burak Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design* 2(3). 173–181.

Balgir, Praveen P & Maleeka Sharma. 2017. Biopharmaceutical potential of ACE-inhibitory peptides. *Journal of Proteomics & Bioinformatics* 10(7). 171–177.

Ballesteros, Juan A & Harel Weinstein. 1995. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods in Neurosciences* 25. 366–428.

Basson, Marc D, Bixi Zeng & Shouye Wang. 2015. Akt1 binds focal adhesion kinase via the Akt1 kinase domain independently of the pleckstrin homology domain. *Journal of Physiology and Pharmacology* 66(5). 701–710.

Becker, George C. 1983. *Fishes of Wisconsin*. Madison, WI, USA: University of Wisconsin Press.

Bengio, Yoshua & Yves Grandvalet. 2004. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research* 5. 1089–1105.

Benkert, Pascal, Marco Biasini & Torsten Schwede. 2011. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27(3). 343–350.

Benton, Richard, Silke Sachse, Stephen W Michnick & Leslie B Vosshall. 2006. Atypical membrane topology and heteromeric function of Drosophila odorant receptors in vivo. *PLoS Biologyogy* 4(2). 240–257.

Berman, Helen M, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov & Philip E Bourne. 2000. The protein data bank. *Nucleic Acids Research* 28(1). 235–242.

Bessac, Bret F & Sven-Eric Jordt. 2010. Sensory detection and responses to toxic gases: mechanisms, health effects, and countermeasures. *Proceedings of the American Thoracic Society* 7(4). 269–77.

Bianchi, Antonio, Claudia Giorgi, Paolo Ruzza, Claudio Toniolo & E James Milner-White. 2012. A synthetic hexapeptide designed to resemble a proteinaceous p-loop nest is shown to bind inorganic phosphate. *Proteins: Structure, Function, and Bioinformatics* 80(5). 1418–1424.

Boogaard, Michael A, Terry D Bills & David A Johnson. 2003. Acute toxicity of TFM and a TFM/niclosamide mixture to selected species of fish, including lake sturgeon (Acipenser fulvescens) and mudpuppies (Necturus maculosus), in laboratory and field exposures. *Journal of Great Lakes Research* 29. 529–541.

Bordoli, Lorenza, Florian Kiefer, Konstantin Arnold, Pascal Benkert, James Battey & Torsten Schwede. 2008. Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols* 4(1). 1–13.

Brandts, John F & Lung Nan Lin. 1990. Study of strong to ultratight protein interactions using differential scanning calorimetry. *Biochemistry* 29(29). 6927–6940.

Brant, Cory O. 2015. *Characterization of sea lamprey pheromone components*: Michigan State University dissertation.

Brant, Cory O, Mar Huertas, Ke Li & Weiming Li. 2016. Mixtures of two bile alcohol sulfates function as a proximity pheromone in sea lamprey. *PloS One* 11(2). e0149508.

Breiman, Leo. 2001. Random forests. *Machine learning* 45(1). 5–32.

Breiman, Leo, Jerome Friedman, Charles J Stone & Richard A Olshen. 1984. *Classification and regression trees*. Pacific Grove, CA: Wadsworth.

Buhrow, Leann, Carrie Hiser, Jeffrey R Van Voorst, Shelagh Ferguson-Miller & Leslie A Kuhn. 2013. Computational prediction and in vitro analysis of potential physiological ligands of the bile acid binding site in cytochrome c oxidase. *Biochemistry* 52(40). 6995–7006.

Burns, Aaron C, Peter W Sorensen & Thomas R Hoye. 2011. Synthesis and olfactory activity of unnatural, sulfated 5beta-bile acid derivatives in the sea lamprey (Petromyzon marinus). *Steroids* 76(3). 291–300.

Capuccini, Marco, Laeeq Ahmed, Wesley Schaal, Erwin Laure & Ola Spjuth. 2017. Large-scale virtual screening on public cloud resources with Apache Spark. *Journal of Cheminformatics* 9(1). 15.

Ceccarelli, Derek FJ, Hyun Kyu Song, Florence Poy, Michael D Schaller & Michael J Eck. 2006. Crystal structure of the FERM domain of focal adhesion kinase. *Journal of Biological Chemistry* 281(1). 252–259.

Chamberlin, Donald D & Raymond F Boyce. 1974. SEQUEL: A structured English query language. In Gene Altshuler, Randall Rustin & Bernard Plagman (eds.), *Proceedings of the 1974 acm sigfidet (now sigmod) workshop on data description, access and control*, 249–264. ACM.

Chang, Andy J, Fabian E Ortega, Johannes Riegler, Daniel V Madison & Mark A Krasnow. 2015. Oxygen control of breathing by an olfactory receptor activated by lactate. *Nature* 527(7577). 240–244.

Chodera, John D & David L Mobley. 2013. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annual Review of Biophysics* 42. 121.

Chook, Yuh Min, Joseph V Gray, Hengming Ke & William N Lipscomb. 1994. The monofunctional chorismate mutase from bacillus subtilis: structure determination of chorismate mutase and its complexes with a transition state analog and prephenate, and implications for the mechanism of the enzymatic reaction. *Journal of Molecular Biology* 240(5). 476–500.

Coleman, David E & Stephen R Sprang. 1999. Structure of Giα1·GppNHp, autoinhibition in a Gα protein-substrate complex. *Journal of Biological Chemistry* 274(24). 16669–16672.

Collawn, James F, Martin Stangel, Leslie A Kuhn, Victor Esekogwu, Shuqian Jing, Ian S Trowbridge & John A Tainer. 1990. Transferrin receptor internalization sequence YXRF implicates a tight turn as the structural recognition motif for endocytosis. *Cell* 63(5). 1061–1072.

Collyer, CA & DM Blow. 1990. Observations of reaction intermediates and the mechanism of aldose-ketose interconversion by d-xylose isomerase. *Proceedings of the National Academy of Sciences* 87(4). 1362–1366.

Colominas, Carles, Francisco J Luque & Modesto Orozco. 1996. Tautomerism and protonation of guanine and cytosine. Implications in the formation of hydrogen-bonded complexes. *Journal of the American Chemical Society* 118(29). 6811–6821.

Cozzini, Pietro, Glen E Kellogg, Francesca Spyrakis, Donald J Abraham, Gabriele Costantino, Andrew Emerson, Francesca Fanelli, Holger Gohlke, Leslie A Kuhn, Garrett M Morris, Modesto Orozco, Thelma A Pertinhez, Menico Rizzi & Christoph A Sotriffer. 2008. Target flexibility: an emerging consideration in drug discovery and design. *Journal of Medicinal Chemistry* 51(20). 6237–6255.

Craig, Lisa, Paul C Sanschagrin, Annett Rozek, Steve Lackie, Leslie A Kuhn & Jamie K Scott. 1998. The role of structure in antibody cross-reactivity between peptides and folded proteins. *Journal of Molecular Biology* 281(1). 183–201.

Cramer, Christopher J & Donald G Truhlar. 1992. AM1-SM2 and PM3-SM3 parameterized SCF solvation models for free energies in aqueous solution. *Journal of Computer-Aided Molecular Design* 6(6). 629–666.

Davis, Ian W, Andrew Leaver-Fay, Vincent B Chen, Jeremy N Block, Gary J Kapral, Xueyi Wang, Laura W Murray, W Bryan Arendall III, Jack Snoeyink, Jane S Richardson et al. 2007. Molprobity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research* 35(Suppl. 2). W375–W383.

Dawson, Natalie L, Tony E Lewis, Sayoni Das, Jonathan G Lees, David Lee, Paul Ashford, Christine A Orengo & Ian Sillitoe. 2016. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research* 45(D1). 289–295.

DeLano, Warren L. 2002. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter On Protein Crystallography* 40. 82–92.

Dill, Ken A. 1997. Additivity principles in biochemistry. *Journal of Biological Chemistry* 272(2). 701–704.

Dixon, Steven L, Alexander M Smondyrev, Eric H Knoll, Shashidhar N Rao, David E Shaw & Richard A Friesner. 2006. PHASE: A new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *Journal of Computer-Aided Molecular Design* 20(10-11). 647–671.

Drwal, Malgorzata N & Renate Griffith. 2013. Combination of ligand- and structure-based methods in virtual screening. *Drug Discovery Today: Technologies* 10(3). 395–401.

Ericsson, Ulrika B, B Martin Hallberg, George T DeTitta, Niek Dekker & Pär Nordlund. 2006. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical Biochemistry* 357(2). 289–298.

Eswar, Narayanan, David Eramian, Ben Webb, Min-Yi Shen & Andrej Sali. 2008. Protein structure modeling with MODELLER. *Structural Proteomics: High-Throughput Methods* 145–159.

Falchi, Federico, Fabiana Caporuscio & Maurizio Recanatini. 2014. Structure-based design of small-molecule protein–protein interaction modulators: the story so far. *Future Medicinal Chemistry* 6(3). 343–357.

Fan, Hao, Dina Schneidman-Duhovny, John J Irwin, Guangqiang Dong, Brian K Shoichet & Andrej Sali. 2011. Statistical potential for modeling and ranking of protein–ligand interactions. *Journal of Chemical Information and Modeling* 51(12). 3078–3092.

Ferrara, Philippe, Holger Gohlke, Daniel J Price, Gerhard Klebe & Charles L Brooks. 2004. Assessing scoring functions for protein-ligand interactions. *Journal of Medicinal Chemistry* 47(12). 3032–3047.

Ferri, F, P Pudil, M Hatef & J Kittler. 1994. Comparative study of techniques for large-scale feature selection. *Pattern Recognition in Practice IV* 1994. 403–413.

Folk, Mike, Gerd Heber, Quincey Koziol, Elena Pourmal & Dana Robinson. 2011. An overview of the HDF5 technology suite and its applications. In Julia Stoyanovich (ed.), *Proceedings of the edbt/icdt 2011 workshop on array databases*, 36–47. ACM.

Friedman, Jerome, Trevor Hastie & Robert Tibshirani. 2001. *The elements of statistical learning* Springer Series in Statistics. New York, NY: Springer.

Gatica, Edgar A & Claudio N Cavasotto. 2012. Ligand and decoy sets for docking to G protein-coupled receptors. *Journal of Chemical Information and Modeling* 52(1). 1–6.

Geppert, Hanna, Martin Vogt & Juergen Bajorath. 2010. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of Chemical Information and Modeling* 50(2). 205–216.

Gerstein, M & Werner Krebs. 1998. A database of macromolecular motions. *Nucleic Acids Research* 26(18). 4280–4290.

Ghosh, Sutapa, Aihua Nie, Jing An & Ziwei Huang. 2006. Structure-based virtual screening of chemical libraries for drug discovery. *Current Opinion in Chemical Biology* 10(3). 194–202.

Gohlke, Holger, Leslie A Kuhn & David A Case. 2004. Change in protein flexibility upon complex formation: Analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins: Structure, Function and Bioinformatics* 56(2). 322–337.

Goodman, Joshua L, Mark D Pagel & Martin J Stone. 2000. Relationships between protein structure and dynamics from a database of NMR-derived backbone order parameters. *Journal of Molecular Biology* 295(4). 963–978.

Gunner, MR, Mohammad A Saleh, Elizabeth Cross, Michael Wise & Others. 2000. Backbone dipoles generate positive potentials in all proteins: origins and implications of the effect. *Biophysical Journal* 78(3). 1126–1144.

Hagey, Lee R, Peter R Møller, Alan F Hofmann & Matthew D Krasowski. 2010. Diversity of bile salts in fish and amphibians: Evolution of a complex biochemical pathway. *Physiological and Biochemical Zoology* 83(2). 308–321.

Halgren, Thomas A. 1996. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* 17(5-6). 490–519.

Hansen, Gretchen JA & Michael L Jones. 2008. A rapid assessment approach to prioritizing streams for control of Great Lakes sea lampreys (Petromyzon marinus): a case study in adaptive management. *Canadian Journal of Fisheries and Aquatic Sciences* 65(11). 2471–2484.

Hawkins, Paul CD & Anthony Nicholls. 2012. Conformer generation with OMEGA: learning from the data set and the analysis of failures. *Journal of Chemical Information and Modeling* 52(11). 2919–2936.

Hawkins, Paul CD, A Geoffrey Skillman & Anthony Nicholls. 2007. Comparison of shape-matching and docking as virtual screening tools. *Journal of Medicinal Chemistry* 50(1). 74–82.

Hawkins, Paul CD, A Geoffrey Skillman, Gregory L Warren, Benjamin A Ellingson & Matthew T Stahl. 2010. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling* 50(4). 572–84.

Hayashi, Naoko, Hiroshi Egami, Mikio Kai, Yuji Kurusu, Sadamu Takano & Michio Ogawa. 1999. No-touch isolation technique reduces intraoperative shedding of tumor cells into the portal vein during resection of colorectal cancer. *Surgery* 125(4). 369–374.

Hespenheide, Brandon M, AJ Rader, MF Thorpe & Leslie A Kuhn. 2002. Identifying protein folding cores from the evolution of flexible regions during unfolding. *Journal of Molecular Graphics and Modelling* 21(3). 195–207.

Hou, Tingjun, Junmei Wang, Youyong Li & Wei Wang. 2011. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling* 51(1). 69–82.

Hu, Guoping, Guanglin Kuang, Wen Xiao, Weihua Li, Guixia Liu & Yun Tang. 2012. Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *Journal of Chemical Information and Modeling* 52(5). 1103–1113.

Hughes, Gordon. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* 14(1). 55–63.

Hunter, JD. 2007. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* 9(3). 90–95.

Hussain, A, LR Saraiva, DM Ferrero, G Ahuja, VS Krishna, SD Liberles & SI Korsching. 2013. High-affinity olfactory receptor for the death-associated odor cadaverine. *Proceedings of the National Academy of Sciences USA* 110(48). 19579–19584.

Ippolito, Joseph A, Richard S Alexander & David W Christianson. 1990. Hydrogen bond stereochemistry in protein structure and function. *Journal of Molecular Biology* 215(3). 457–471.

Irwin, John J & Brian K Shoichet. 2005. ZINC—a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* 45(1). 177–82.

Jacobs, Donald J & Bruce Hendrickson. 1997. An algorithm for two-dimensional rigidity percolation: the pebble game. *Journal of Computational Physics* 137(2). 346–365.

Jacobs, Donald J, Leslie A Kuhn & Michael F Thorpe. 2002. Flexible and rigid regions in proteins. In *Rigidity theory and applications*, 357–384. Springer.

Jacobs, Donald J, Andrew J Rader, Leslie A Kuhn & Michael F Thorpe. 2001. Protein flexibility predictions using graph theory. *Proteins: Structure, Function and Bioinformatics* 44(2). 150–165.

Jakalian, Araz, David B Jack & Christopher I Bayly. 2002. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* 23(16). 1623–1641.

Jang, Bohee, Hyejung Jung, Sojoong Choi, Young Hun Lee, Seung-Taek Lee & Eok-Soo Oh. 2017. Syndecan-2 cytoplasmic domain up-regulates matrix metalloproteinase-7 expression via the protein kinase Cγ–mediated FAK/ERK signaling pathway in colon cancer. *Journal of Biological Chemistry* 292(39). 16321–16332.

Jhoti, Harren, Onkar MP Singh, Malcolm P Weir, Robert Cooke, Peter Murray-Rust & Alan Wonacott. 1994. X-ray crystallographic studies of a series of penicillin-derived asymmetric inhibitors of HIV-1 protease. *Biochemistry* 33(28). 8417–8427.

Johnson, David K & John Karanicolas. 2016. Ultra-high-throughput structure-based virtual screening for small-molecule inhibitors of protein-protein interactions. *Journal of Chemical Information and Modeling* 56(2). 399–411.

Johnson, Nicholas S, Michael J Siefkes, C Michael Wagner, Gale Bravener, Todd Steeves, Michael Twohey & Weiming Li. 2015. Factors influencing capture of invasive sea lamprey in traps baited with a synthesized sex pheromone component. *Journal of Chemical Ecology* 41(10). 913–923.

Johnson, Nicholas S, Sang-Seon Yun & Weiming Li. 2014. Investigations of novel unsaturated bile salts of male sea lamprey as potential chemical cues. *Journal of Chemical Ecology* 40(10). 1152–1160.

Johnson, Nicholas S, Sang-Seon Yun, Henry T Thompson, Cory O Brant & Weiming Li. 2009. A synthesized pheromone induces upstream movement in female sea lamprey and summons them into traps. *Proceedings of the National Academy of Sciences* 106(4). 1021–1026.

Jones, Eric, Travis Oliphant & Pearu Peterson. 2001. SciPy: Open source scientific tools for Python. *http://www.scipy.org* .

Kabsch, Wolfgang & Christian Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12). 2577–2637.

Kain, Pinky, Sean Michael Boyle, Sana Khalid Tharadra, Tom Guda, Christine Pham, Anupama Dahanukar & Anandasankar Ray. 2013. Odour receptors and neurons for DEET and new insect repellents. *Nature* 502(7472). 507–512.

Katritch, Vsevolod, Vadim Cherezov & Raymond C Stevens. 2012. Diversity and modularity of G protein-coupled receptor structures. *Trends in Pharmacological Sciences* 33(1). 17–27.

Keskin, Ozlem, Buyong Ma & Ruth Nussinov. 2005. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology* 345(5). 1281–1294.

Koes, David, Kareem Khoury, Yijun Huang, Wei Wang, Michal Bista, Grzegorz M Popowicz, Siglinde Wolf, Tad A Holak, Alexander Dömling & Carlos J Camacho. 2012. Enabling large-scale design, synthesis and validation of small molecule protein-protein antagonists. *PloS One* 7(3). e32839.

Kohavi, Ron. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence* 14(12). 1137–1143.

Korn, Alex P & David R Rose. 1994. Torsion angle differences as a means of pinpointing local polypeptide chain trajectory changes for identical proteins in different conformational states. *Protein Engineering, Design and Selection* 7(8). 961–967.

Krieger, Elmar, Tom Darden, Sander B Nabuurs, Alexei Finkelstein & Gert Vriend. 2004. Making optimal use of empirical energy functions: force-field parameterization in crystal space. *Proteins: Structure, Function, and Bioinformatics* 57(4). 678–683.

Krieger, Elmar, Roland L Dunbrack, Rob WW Hooft & Barbara Krieger. 2012. Assignment of protonation states in proteins and ligands: combining pK a prediction with hydrogen bonding network optimization. *Methods in Molecular Biology: Computational Drug Discovery and Design* 819. 405–421.

Krieger, Elmar, Keehyoung Joo, Jinwoo Lee, Jooyoung Lee, Srivatsan Raman, James Thompson, Mike Tyka, David Baker & Kevin Karplus. 2009. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins: Structure, Function and Bioinformatics* 77(Suppl. 9). 114–122.

Kruse, Andrew C, Aaron M Ring, Aashish Manglik, Jianxin Hu, Kelly Hu, Katrin Eitel, Harald Hübner, Els Pardon, Celine Valant, Patrick M Sexton & Others. 2013. Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* 504(7478). 101–106.

Kubinyi, Hugo, Gerd Folkers & Yvonne C Martin. 2006. *3D QSAR in drug design: recent advances*, vol. 3. Springer Science & Business Media.

Kuhn, Leslie A, Craig A Swanson, Michael E Pique, John A Tainer & Elizabeth D Getzoff. 1995. Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins: Structure, Function, and Bioinformatics* 23(4). 536–547.

Laskowski, RA, MW MacArthur, DS Moss & JM Thornton. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* 26(2). 283–291.

Li, Qingliang & Salim Shah. 2017. Structure-based virtual screening. *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics* 111–124.

Li, Shenhui & Mei Hong. 2011. Protonation, tautomerization, and rotameric structure of histidine: A comprehensive study by magic-angle-spinning solid-state NMR. *Journal of the American Chemical Society* 133(5). 1534–1544.

Li, Weiming, Alexander P Scott, Michael J Siefkes, Honggao Yan, Qin Liu, Sang-Seon Yun & Douglas A Gage. 2002. Bile acid secreted by male sea lamprey that acts as a sex pheromone. *Science.* 296(5565). 138–141.

Li, Weiming, Peter W Sorensen & Daniel D Gallaher. 1995. The olfactory system of migratory adult sea lamprey (Petromyzon marinus) is specifically and acutely sensitive to unique bile acids released by conspecific larvae. *The Journal of General Physiology* 105(5). 569–587.

Libants, Scot, Kevin Carr, Hong Wu, John H Teeter, Yu-Wen Chung-Davidson, Ziping Zhang, Curt Wilkerson & Weiming Li. 2009. The sea lamprey Petromyzon marinus genome reveals the early origin of several chemosensory receptor families in the vertebrate lineage. *BMC Evolutionary Biology* 9. 180.

Liberles, Stephen D. 2014. Mammalian pheromones. *Annual Review of Physiology* 76. 151–75.

Lipinski, CA, F Lombardo, BW Dominy & PJ Feeney. 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* 23. 3–25.

Liu, Nan, Jeffrey R Van Voorst, John B Johnston & Leslie A Kuhn. 2015. CholMine: Determinants and prediction of cholesterol and cholate binding across nonhomologous protein structures. *Journal of Chemical Information and Modeling* 55(4). 747–759.

Liu, Wei, Eugene Chun, Aaron A Thompson, Pavel Chubukov, Fei Xu, Vsevolod Katritch, Gye Won Han, Christopher B Roth, Laura H Heitman, Adriaan P IJzerman & Others. 2012. Structural basis for allosteric regulation of GPCRs by sodium ions. *Science* 337(6091). 232–236.

Louppe, Gilles. 2014. *Understanding Random Forests: From Theory to Practice*: University of Liege dissertation.

Lucas, Martyn C, Damian H Bubb, Mun Ho Jang, Kyong Ha & Jerome EG Masters. 2009. Availability of and access to critical habitats in regulated rivers: Effects of low-head barriers on threatened lampreys. *Freshwater Biology* 54(3). 621–634.

Lundstrom, Kenneth. 2009. An overview on GPCRs and drug discovery: structure-based drug design and structural biology on GPCRs. *G Protein-Coupled Receptors in Drug Discovery* 51–66.

Lyne, Paul D. 2002. Structure-based virtual screening: An overview. *Drug Discovery Today* 7(20). 1047–1055.

Mathews, Irimpan I, Mark D Erion & Steven E Ealick. 1998. Structure of human adenosine kinase at 1.5 A resolution. *Biochemistry* 37(45). 15607–15620.

Maxwell, J Clerk. 1864. On the calculation of the equilibrium and stiffness of frames. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 27(182). 294–299.

McDonald, D Gordon & Cynthia S Kolar. 2007. Research to guide the use of lampricides for controlling sea lamprey. *Journal of Great Lakes Research* 33. 20–34.

McDonald, Ian & Janet M Thornton. 1994. *Atlas of side-chain and main-chain hydrogen bonding*. http://www.biochem.ucl.ac.uk/bsm/atlas: Biochemistry and Molecular Biology Department, University College London.

McGaughey, Georgia B, Robert P Sheridan, Christopher I Bayly, J Chris Culberson, Constantine Kreatsoulas, Stacey Lindsley, Vladimir Maiorov, Jean-Francois Francois Truchon & Wendy D Cornell. 2007. Comparison of topological, shape, and docking methods in virtual screening. *Journal of Chemical Information and Modeling* 47(4). 1504–1519.

McKinney, Wes. 2010. Data structures for statistical computing in Python. In Jarrod Millman & Stefan van der Walt (eds.), *Proceedings of the 9th python in science conference*, 51–56.

Merz Jr, Kenneth M. 2010. Limits of free energy computation for protein-ligand interactions. *Journal of chemical theory and computation* 6(5). 1769–1776.

Miller, Bill R & Adrian E Roitberg. 2013. Design of e-pharmacophore models using compound fragments for the trans-sialidase of Trypanosoma cruzi: screening for novel inhibitor scaffolds. *Journal of Molecular Graphics and Modelling* 45. 84–97.

Milligan, Graeme, Trond Ulven, Hannah Murdoch & Brian D Hudson. 2014. G-protein-coupled receptors for free fatty acids: nutritional and therapeutic targets. *The British Journal of Nutrition* 111(1). 3–7.

Mirza, Shaher Bano, Ramin Ekhteiari Salmas, M Qaiser Fatmi & Serdar Durdagi. 2016. Virtual screening of eighteen million compounds against dengue virus: Combined molecular docking and molecular dynamics simulations study. *Journal of Molecular Graphics and Modelling* 66. 99–107.

Mitchell, John BO. 2014. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4(10). 468–481.

Morris, Garrett M, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell & Arthur J Olson. 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* 30(16). 2785–2791.

Muegge, Ingo & Prasenjit Mukherjee. 2016. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery* 11(2). 137–148.

Mueller, Andreas C & Sarah Guido. 2017. *Introduction to machine learning with Python: a guide for data scientists*. Sebastopol, CA: O'Reilly Media.

Murakami, Midori & Tsutomu Kouyama. 2008. Crystal structure of squid rhodopsin. *Nature* 453(7193). 363–367.

Murgueitio, Manuela S, Philipp Henneke, Hartmut Glossmann, Sandra Santos-Sierra & Gerhard Wolber. 2014. Prospective virtual screening in a sparse data scenario: design of small-molecule TLR2 antagonists. *ChemMedChem* 9(4). 813–822.

Nagamine, Nobuyoshi, Takayuki Shirakawa, Yusuke Minato, Kentaro Torii, Hiroki Kobayashi, Masaya Imoto & Yasubumi Sakakibara. 2009. Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening. *PLoS Computational Biology* 5(6). e1000397.

Neudert, Gerd & Gerhard Klebe. 2011. DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. *Journal of Chemical Information and Modeling* 51(10). 2731–2745.

Niesen, Frank H, Helena Berglund & Masoud Vedadi. 2007. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols* 2(9). 2212–2221.

Niimura, Yoshihito, Niimura & Y. 2009. On the origin and evolution of vertebrate olfactory receptor genes: Comparative genome analysis among 23 chordate species. *Genome Biology and Evolution* 1(2006). 34–44.

Nittinger, Eva, Therese Inhester, Stefan Bietz, Agnes Meyder, Karen T Schomburg, Gudrun Lange, Robert Klein & Matthias Rarey. 2017. Large-scale analysis of hydrogen bond interaction patterns in protein-ligand interfaces. *Journal of Medicinal Chemistry* 60(10). 4245–4257.

Palumbi, Stephen R. 2001. Humans as the world's greatest evolutionary force. *Science* 293(5536). 1786–1790.

Panigrahi, Sunil K & Gautam R Desiraju. 2007. Strong and weak hydrogen bonds in the protein–ligand interface. *Proteins: Structure, Function, and Bioinformatics* 67(1). 128–141.

Pantoliano, Michael W, Eugene C Petrella, Joseph D Kwasnoski, Victor S Lobanov, James Myslik, Edward Graf, Ted Carver, Eric Asel, Barry A Springer, Pamela Lane & FR Salemme. 2001. High-density miniaturized thermal shift assays as a general strategy for drug discovery. *Journal of Biomolecular Screening* 6(6). 429–440.

Pauling, Linus. 1960. *The nature of the chemical bond and the structure of molecules and crystals: an introduction to modern structural chemistry*, vol. 18. Ithaca, New York: Cornell University Press.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg & Others. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12. 2825–2830.

Pérez-Nueno, Violeta I, David W Ritchie, Obdulia Rabal, Rosalia Pascual, Jose I Borrell & Jordi Teixidó. 2008. Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. *Journal of Chemical Information and Modeling* 48(3). 509–533.

Prakash, Balaji, Louis Renault, Gerrit JK Praefcke, Christian Herrmann & Alfred Wittinghofer. 2000. Triphosphate structure of guanylate-binding protein 1 and implications for nucleotide binding and GTPase mechanism. *The EMBO Journal* 19(17). 4555–4564.

Prevelige Jr, Peter & Gerald D Fasman. 1989. Chou-Fasman prediction of the secondary structure of proteins. In *Prediction of protein structure and the principles of protein conformation*, 391–416. Springer.

Rader, AJ, Brandon M Hespenheide, Leslie A Kuhn & Michael F Thorpe. 2002. Protein unfolding: rigidity lost. *Proceedings of the National Academy of Sciences USA* 99(6). 3540–3545.

Raschka, S, J Bemister-Buffington & LA Kuhn. 2016a. Detecting the native ligand orientation by interfacial rigidity: SiteInterlock. *Proteins: Structure, Function and Bioinformatics* 84(12). 1888–1901.

Raschka, Sebastian. 2015. *Python Machine Learning*. Birmingham, UK: Packt Publishing.

Raschka, Sebastian. 2017a. BioPandas: Working with molecular structures in pandas DataFrames. *The Journal of Open Source Software* 2(14). 1–3.

Raschka, Sebastian. 2017b. rasbt/mlxtend: Version 0.7.0. doi:10.5281/zenodo.816309. https://doi.org/10.5281/zenodo.816309.

Raschka, Sebastian, David Julian & John Hearty. 2016b. *Python: Deeper Insights into Machine Learning*. Birmingham, UK: Packt Publishing.

Raschka, Sebastian & Vahid Mirjalili. 2017. *Python Machine Learning, 2nd Ed.* Birmingham, UK: Packt Publishing.

Raschka, Sebastian, Anne M Scott, Nan Liu, Santosh Gunturu, Mar Huertas, Weiming Li & Leslie A Kuhn. 2017. Enabling the hypothesis-driven prioritization of ligand candidates in big databases: Screenlamp and its application to GPCR inhibitor discovery. *In Revision*.

Raymer, Michael L, William F Punch, Erik D Goodman, Leslie A Kuhn & Anil K Jain. 2000. Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation* 4(2). 164–171.

Raymer, Michael L, Paul C Sanschagrin, William F Punch, Sridhar Venkataraman, Erik D Goodman & Leslie A Kuhn. 1997. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *Journal of Molecular Biology* 265(4). 445–464.

Renault, Louis, Bernard Guibert & Jacqueline Cherfils. 2003. Structural snapshots of the mechanism and inhibition of a guanine nucleotide exchange factor. *Nature* 426(6966). 525–530.

Ripphausen, Peter, Britta Nisius & Juergen Bajorath. 2011. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today* 16(9). 372–376.

Rosenbaum, Daniel M, Søren GF Rasmussen & Brian K Kobilka. 2009. The structure and function of G-protein-coupled receptors. *Nature* 459(7245). 356–363.

Rush III, Thomas S, J Andrew Grant, Lidia Mosyak & Anthony Nicholls. 2005. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *Journal of Medicinal Chemistry* 1489–1495.

Russel, Daniel, Keren Lasker, Ben Webb, Javier Velázquez-Muriel, Elina Tjioe, Dina Schneidman-Duhovny, Bret Peterson & Andrej Sali. 2012. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biology* 10(1). e1001244.

Saeys, Yvan, Iñaki Inza & Pedro Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19). 2507–2517.

Sander, Chris & Reinhard Schneider. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function and Bioinformatics* 9(1). 56–68.

Sandru, A, S Voinea, E Panaitescu & A Blidaru. 2014. Survival rates of patients with metastatic malignant melanoma. *Journal of Medicine and Life* 7(4). 572.

Scott, John W & Pamela E Scott-Johnson. 2002. The electroolfactogram: a review of its history and uses. *Microscopy research and technique* 58(3). 152–160.

Scott, William Beverley & Edwin John Crossman. 1973. Freshwater fishes of Canada. *Fisheries Research Board of Canada Bulletin* 184.

Shan, Shu-ou & Daniel Herschlag. 1996. The change in hydrogen bond strength accompanying charge rearrangement: Implications for enzymatic catalysis. *Proceedings of the National Academy of Sciences* 93(25). 14474–14479.

Shapovalov, Maxim V & Roland L Dunbrack. 2011. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 19(6). 844–858.

Sheridan, Robert P, Georgia B McGaughey & Wendy D Cornell. 2008. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *Journal of Computer-Aided Molecular Design* 22(3-4). 257–265.

Shiratsuchi, Hiroe & Marc D Basson. 2004. Extracellular pressure stimulates macrophage phagocytosis by inhibiting a pathway involving FAK and ERK. *American Journal of Physiology-Cell Physiology* 286(6). 1358–1366.

Stillman, TJ, PJ Baker, KL Britton & DW Rice. 1993. Conformational flexibility in glutamate dehydrogenase: role of water in substrate recognition and catalysis. *Journal of Molecular Biology* 234(4). 1131–1139.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(1). 307.

Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4). 323.

Sukuru, Sai Chetan K, Thibaut Crepin, Youli Milev, Liesl C Marsh, Jonathan B Hill, Regan J Anderson, Jonathan C Morris, Anjali Rohatgi, Gavin O'Mahony, Morten Grøtli & Others. 2006. Discovering new classes of Brugia malayi asparaginyl-tRNA synthetase inhibitors and relating specificity to conformational change. *Journal of Computer-Aided Molecular Design* 20(3). 159–178.

Tang, Karen ES & Ken A Dill. 1998. Native protein fluctuations: the conformational-motion temperature and the inverse correlation of protein flexibility with protein stability. *Journal of Biomolecular Structure and Dynamics* 16(2). 397–411.

Taylor, Robin & Olga Kennard. 1984. Hydrogen-bond geometry in organic crystals. *Accounts of Chemical Research* 17(9). 320–326.

Thomas, Jeffrey W, Marion A Cooley, Jill M Broome, Ravi Salgia, James D Griffin, Christian R Lombardo & Michael D Schaller. 1999. The role of focal adhesion kinase binding in the regulation of tyrosine phosphorylation of paxillin. *Journal of Biological Chemistry* 274(51). 36684–36692.

Tinberg, Christine E, Sagar D Khare, Jiayi Dou, Lindsey Doyle, Jorgen W Nelson, Alberto Schena, Wojciech Jankowski, Charalampos G Kalodimos, Kai Johnsson, Barry L Stoddard & David Baker. 2013. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* 501(7466). 212–216.

Trott, Oleg & Arthur J Olson. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* 31(2). 455–461.

Van Der Walt, Stefan, S Chris Colbert & Gael Varoquaux. 2011. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering* 13(2). 22–30.

Van Rossum, Guido. 2007. Python programming language. In *Usenix annual technical conference*, vol. 41, 36.

Van Voorst, Jeffrey R, Yiying Tong & Leslie A Kuhn. 2012. ArtSurf: a method for deformable partial matching of protein small-molecule binding sites. In *Proceedings of the acm conference on bioinformatics, computational biology and biomedicine*, 36–43. ACM.

Vedadi, M, FH Niesen, A Allali-Hassani, OY Fedorov, PJ Finerty, GA Wasney, R Yeung, C Arrowsmith, LJ Ball, H Berglund, R Hui, BD Marsden, P Nordlund, M Sundstrom, J Weigelt & AM Edwards. 2006. Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proceedings of the National Academy of Sciences* 103(43). 15835–15840.

Velazquez-Campoy, Adrian, Matthew J Todd & Ernesto Freire. 2000. HIV-1 protease inhibitors: enthalpic versus entropic optimization of the binding affinity. *Biochemistry* 39(9). 2201–2207.

Venkatakrishnan, AJ, Xavier Deupi, Guillaume Lebon, Christopher G Tate, Gebhard F Schertler & M Madan Babu. 2013. Molecular signatures of G-protein-coupled receptors. *Nature* 494(7436). 185–194.

Verma, Jitender, Vijay M Khedkar & Evans C Coutinho. 2010. 3D-QSAR in drug design-a review. *Current Topics in Medicinal Chemistry* 10(1). 95–115.

Walker, Strother H & David B Duncan. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54(1-2). 167–179.

Wang, Renxiao, Luhua Lai & Shaomeng Wang. 2002. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design* 16(1). 11–26.

Wang, S & MD Basson. 2011. Akt directly regulates focal adhesion kinase through association and serine phosphorylation: implication for pressure-induced colon cancer metastasis. *AJP: Cell Physiology* 300(3). C657–C670.

Warne, Tony, Maria J Serrano-Vega, Jillian G Baker, Rouslan Moukhametzianov, Patricia C Edwards, Richard Henderson, Andrew GW Leslie, Christopher G Tate & Gebhard FX Schertler. 2008. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* 454(7203). 486–91.

Warren, Gregory L, Thanh D Do, Brian P Kelley, Anthony Nicholls & Stephen D Warren. 2012. Essential considerations for using protein–ligand structures in drug discovery. *Drug Discovery Today* 17(23). 1270–1281.

Weininger, David. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* 28(1). 31–36.

Wolpert, David H. 1996. The lack of a priori distinctions between learning algorithms. *Neural Computation* 8(7). 1341–1390.

Word, J Michael, Simon C Lovell, Jane S Richardson & David C Richardson. 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology* 285(4). 1735–1747.

Wright, Peter E & H Jane Dyson. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology* 293(2). 321–331.

Yamaguchi, Kazuya, Yukihiro Takagi, Shinichirou Aoki, Manabu Futamura & Shigetoyo Saji. 2000. Significant detection of circulating cancer cells in the blood by reverse transcriptase–polymerase chain reaction during colorectal cancer resection. *Annals of Surgery* 232(1). 58.

Yan, Xin, Chenzhong Liao, Zhihong Liu, Arnold T Hagler, Qiong Gu & Jun Xu. 2016. Chemical structure similarity search for ligand-based virtual screening: Methods and computational resources. *Current Drug Targets* 17(14). 1580–1585.

Yang, Chao-Yie, Renxiao Wang & Shaomeng Wang. 2005. A systematic analysis of the effect of small-molecule binding on protein flexibility of the ligand-binding sites. *Journal of Medicinal Chemistry* 48(18). 5648–5650.

Zahiri, Javad, Joseph Hannon Bozorgmehr & Ali Masoudi-Nejad. 2013. Computational prediction of protein–protein interaction networks: algorithms and resources. *Current Genomics* 14(6). 397–414.

Zavodsky, Maria I, Paul C Sanschagrin, Rajesh S Korde & Leslie A Kuhn. 2002. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *Journal of Computer-Aided Molecular Design* 16(12). 883–902.

Zavodszky, Maria I, Ming Lei, MF Thorpe, Anthony R Day & Leslie A Kuhn. 2004. Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins: Structure, Function and Bioinformatics* 57(2). 243–261.

Zavodszky, Maria I, Anjali Rohatgi, Jeffrey R Van Voorst, Honggao Yan & Leslie A Kuhn. 2009. Scoring ligand similarity in structure-based virtual screening. *Journal of Molecular Recognition* 22(4). 280–292.

Zeng, Bixi, Dinesh Devadoss, Shouye Wang, Emilie E Vomhof-dekrey, A Kuhn & Marc D Basson. 2017. Inhibition of pressure-activated cancer cell adhesion by FAK-derived peptides. *Oncotarget* .

Zhou, Hongyi & Yaoqi Zhou. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science* 11(11). 2714–2726.

Zoete, Vincent, Antoine Daina, Christophe Bovigny & Olivier Michielin. 2016. SwissSimilarity: a web tool for low to ultra high throughput ligand-based virtual screening. *Journal of Chemical Information and Modeling* 56(8). 1399–1404.