Title

Pharmacological organization of proteins: Towards an understanding of biological symbolism and protein relationships from the perspective of ligands

Author

Lin, Henry

Publication Date

2014

Peer reviewed|Thesis/dissertation

Pharmacological organization of proteins: Towards an understanding of biological symbolism and protein relationships from the perspective of ligands

by

Henry Lin

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

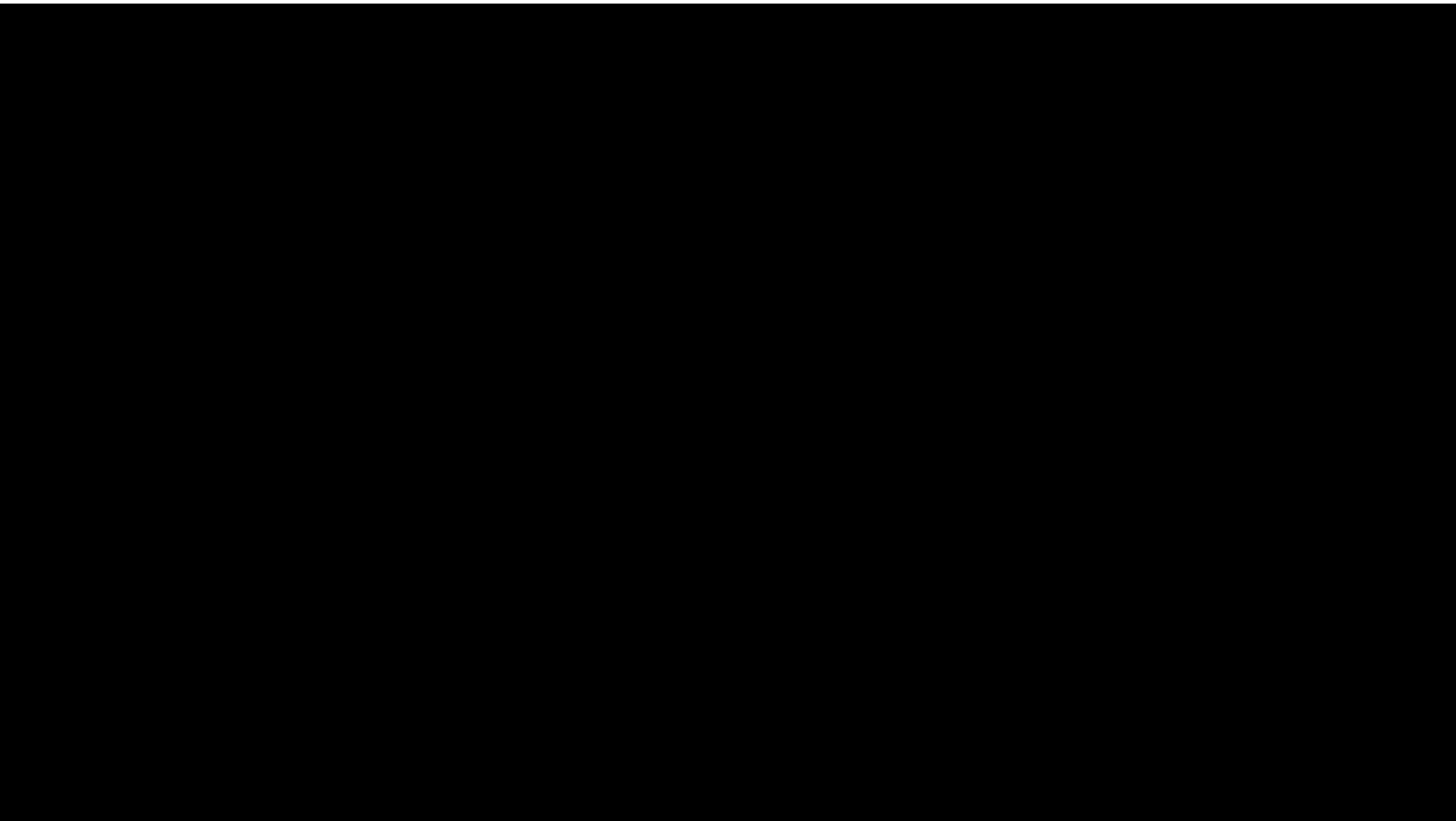
Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



To my family and Elizabeth

Acknowledgements

First and foremost I thank my advisor, Brian Shoichet, throughout all the years of graduate school at UCSF. I was constantly challenged by Brian and learned a great deal about how to think like a scientist. His constant encouragement and excitement for my projects gave me the motivation to keep pushing forward, even when it felt like I was swimming upstream. Brian's ability to have a vision and get excited about ideas was contagious and was easy for me to latch on and ride the wave.

I also thank my lab mates over the years that always knew what I was going through in my projects. John Irwin, Pascal Wassam, Teague Sterling and Therese Demers were always available to help with all my computer, ZINC, software and server needs. Without administrative and moral support from Allison Doak and Yvonne Munchua, this thesis would not have been possible, or would've taken me at least half a year longer trying to figure out all the logistics and administrative things in the lab. Michael Keiser's guidance and help with and invention of SEA were instrumental in me joining the lab and enabling the technology to make my thesis happen. Ryan Coleman and Michael Mysinger were always available to help with DOCK, help me get set up with scripts and answer my many mundane questions about DOCK and how to run it. Over the years, there were a lot of subgroup meetings that I was a part of including, Team GPCR, Team chemoinformatics, Team EFI, Team DOCK. The feedback and suggestions I received during these meetings from my lab mates provided me with a research atmosphere and insight that helped make my projects advance.

My thesis committee, Professors Patricia Babbitt, Andrej Sali and Matthew Jacobson, were always supportive in my projects and provided orthogonal perspectives at times to help make my projects and thesis stronger. Additionally, they made sure I would graduate in a timely manner and not drag things out forever. Thank you for also making sure Brian would let me graduate on time!

I am grateful to all my collaborators over the years. They provided not "just a pair of hands" to do experiments, but also support, energy and expertise on how to direct the project and steps to move the project forward. I give special thanks to Professor Bryan Roth and his group at the Psychoactive Drug Screening Center for performing most of the experiments in this thesis. Flori Sassano was my main point of contact and made sure all my requested experiments were performed and even then some. Professor Brian Kobilka and his graduate student Aashish Manglik gave me the opportunity to dock to the μ opioid receptor structure even before they released the structure to the public.

I am forever indebted to my family and friends. My parents, Tony and Judy, and brother, Dr. George, have always known how to supported me without ever knowing what I do in the lab or at work. My fiancé and future wife, Elizabeth, has given me the drive, support and motivation these past couple of years to make sure this thesis would be completed. I thank my friends, both from before and new at UCSF, for all the good times.

I dedicate this thesis to my family and Elizabeth.

The text of Chapter 1 is a reprint of the material as it appears in:

Lin, H.*, Sassano, M.F.*, et al. A Pharmacological Organization of G Protein-coupled Receptors. *Nat. Methods* **10**(2), 140-146 (2013). *Co-first authors.

It appears here with permission from the authors.

Abstract

Pharmacological organization of proteins: Towards an understanding of biological symbolism and protein relationships from the perspective of ligands

Henry Lin

Relating proteins based on the similarity of the ligands that bind to and modulate through them is a field that has its basis in classical pharmacology when ligands were tested for phenotypic effects on whole tissues, organs or organisms. We take a ligand-focused view of biology to build protein-protein relationships by describing proteins not by their sequence, structure or function, but by their ligand sets. A systematic approach to relate proteins together both within a family and between families are described. The engine we use to drive a pharmacological organization of proteins is the Similarity Ensemble Approach (SEA) using ligands from ChEMBL to create sets for each protein target as input. We use the output from SEA as a similarity metric to relate proteins pharmacologically.

Family A G protein-coupled receptors (GPCRs) was a family of proteins we first aimed to reorganize pharmacologically. Compared to a sequence based organization of GPCRs, the ligand based organization had a much different arborization of its dendrogram with some GPCRs moving away from their “commonly” related GPCRs while some “distantly” related GPCRs were brought together because they shared

common ligands. The pharmacological organization led us to predict for testing GPCRs that were similar based on pharmacology but different from a sequence view. We confirmed three new pairs of GPCRs that were now linked by a new, shared ligand where they previously had no known shared ligands.

Moving beyond GPCRs, we sought a deeper biological relationship between proteins and protein families we related pharmacologically by finding new protein links that not only could potentially share a ligand, based on our predictions, but also shared a phenotype, function, or are implicated in similar diseases. The proteins we related in this manner were also not from the same sequence or structural family but different from classical bioinformatics metrics. The assertion is that because there is a limited number of endogenous signaling molecules, their evolution is almost frozen and therefore, cells and their proteins must evolve and adapt around signaling molecules on top of functioning through them. Relating proteins pharmacologically and phenotypically is explored.

Table of Contents

List of Tables

List of Figures

Introduction

In 2007, I was contemplating what to study for graduate school and read an article from the Shoichet Lab at UCSF relating proteins not by their sequence, structure or function, but by the ligands that bind to each of the proteins[1]. My first thoughts were, "oh, isn't this just ligand similarity? We find the nearest neighbors for molecules all the time with the Tanimoto coefficient[2]." It was a very relatable paper for me in industry and I quickly saw the applications of the method but after reading through the discussion, I was intrigued by the references to the "molecular evolution of biological function" idea and attempts to explain drug side-effects by the predicted off-targets using the method. The paper in many ways was different than the docking papers that I was more familiar with, which ultimately led me to wanting to attend UCSF for graduate school and join the Shoichet Lab.

After officially joining the Shoichet Lab at UCSF in June 2009, I had a brainstorming session with Brian Shoichet on potential thesis projects for me to work on. The projects were entitled, "crazy goals" and were essentially combining DOCK[3, 4] with SEA to reorganize biology and proteins from a pharmacological point of view and find new functions for proteins using newly discovered ligands. What sounded like a challenging and ambitious set of projects was then later refined and revised in a subsequent meeting about two years later where the projects became further challenging and ambitious but perhaps more cohesive and focused and was renamed "crazy goals (II)", which thus became this thesis. In this introduction, I describe

the motivation, some background and a guide to the chapters that ultimately make up this thesis.

I. Motivation and background

An initial driving idea for this thesis was the idea of polypharmacology. Ligand recognition by multiple proteins, otherwise known as polypharmacology, is emerging as an approach to treat complex diseases by having a single molecule bind to multiple targets implicated in the same disease. Unfortunately, polypharmacology is also often the culprit of adverse drug reactions and undesired side-effects[5-7]. What is actually two sides of the same coin—polypharmacology—precise target combinations could be gained by identifying as many targets for a particular drug candidate as early as possible. Although this represents a considerable challenge to medicinal chemists, there are documented cases where targeting certain combinations was possible[8, 9]. Taking a step back, endogenous ligands in organisms have always exhibited polypharmacologic behavior. For instance, acetylcholine and serotonin signal through G-Protein Coupled Receptors (GPCRs) and ion channels and are trafficked through cell membranes by specific transporters[7]. All three protein types belong to largely different families from a bioinformatics perspective, since these proteins do not have structural and sequence similarity, but still recognize and function through the same neurotransmitters. Thus, it is not surprising that synthetic molecules and drugs also exhibit polypharmacology.

Within a target family such as kinases, efforts have been made to prospectively target specific kinases that are all implicated in a particular disease[8]. Sequence based metrics have often been used to suggest selectivity of a ligand for a set of targets. For specific instances,

these principles held but there were often many exceptions[6]. The polypharmacology of ligands seemed random or would just occur by chance and somehow still governed by the biophysics of protein ligand interactions. In this thesis, I aim to understand the biological driving force of the polypharmacology ligands and suggest how a limited number of endogenous ligands can regulate so many different biological functions, signaling pathways and phenotypes in an organism.

The key idea we use is to relate proteins not by their sequence, structure or function, but by the ligands that modulate the proteins or what we call pharmacologically[1]. The idea stems from the earlier days of drug discovery and pharmacology where ligands were tested for phenotypic effects on whole tissues, organs, or organisms[10]. The concept of target based drug discovery was not as developed or prevalent and the focus was on the small molecule ligands themselves. SEA uses the older pharmacological principles but instead of labeling phenotypes onto the ligands, they are annotated with targets that they modulate. With the SEA engine, target-target relationships based on the similarity of the ligands that bind to each target can be calculated.

But how do these protein-protein relationships compare to the more familiar sequence similarity metrics? We aimed to answer this question by comparing a sequence based organization to a pharmacological based organization for a protein family, namely Family A GPCRs [11] . We were motivated by a paper by Gloriam DE et. al.[12] that realigned the sequences of Family A GPCRs using structural and mutational, both public and in-house. The sequence alignment also suggested a ligand binding site for all the GPCRs that we felt we could use to

compare directly with ligand similarity using SEA. Surprisingly, the sequence based organization was drastically different than the pharmacological one in many areas. There were some similarities within a protein sub-family but also some dramatic shifts of some proteins. Other than the reorganization of GPCRs from a family perspective, we also had predictions for directly testing proteins that were similar based on pharmacology but different from a binding site sequence view. Three of the predictions were confirmed with ligands binding to new GPCRs so we expanded the protein-protein relationship beyond GPCRs to *non*-GPCR targets once again using SEA and ligand binding data for the *non*-GPCR targets.

Once again we were surprised at how many *non*-GPCR targets, ~485 with significant E-values, were closely associated with the GPCRs by ligand similarity. We subsequently made predictions for ligands that already interact with *non*-GPCRs and we tested and confirmed their binding to GPCR targets suggested by SEA. A huge influence for this thesis actually came when we submitted the paper and got four positive reviews but was still challenged by an editor of *Nature Methods*, Allison Doerr. She alerted us to a "serious editorial concern (reflected in some of the reviewer comments) about whether the method yields biologically relevant information, or whether the results of classifying proteins by ligand binding are only useful for determining drug off-target effects. To justify publication in *Nature Methods* a strong case for the basic research applications of the approach and the biological relevance of your findings (especially the GPCR:non-GPCR links) needs to be made, in our collective opinion." At first we were confounded by the request but with further thought, we felt it was a valid point and forced us to come up with a response that not only addresses their concerns, but also enlightened us to

take this project a step further by focusing in on deeper underlying structure and relationship of proteins that have chemical pressures against them to evolve.

A sequence based organization of proteins is a commonly accepted reflection of evolution but is limited to comparing just within a protein family due to secondary and tertiary structural constraints that define a protein family. A ligand based organization of protein can also be seen as a reflection of the chemical pressures from the relatively smaller number of signaling and endogenous molecules in an organism but is not limited to comparison within a protein family and can be expanded to proteins in different families as well. Proteins do not merely react to chemical perturbations, but respond and evolve to them. For example, long term exposure to morphine leads to lower efficacy due to opioid receptor desensitization. The μ-opioid receptor (MOR) will more quickly be phosphorylated and internalized into the cell sooner with more morphine treatment, thus decreasing the analgesic effect of morphine[13].

This significant development in our approach led us to follow up the Nature Methods paper in two separate focuses. One focus was to relate disparate proteins by predicted ligands instead of annotated ligands. The predicted ligands were generated by taking each protein's structure, namely the binding site, and docking a library of small molecules into it and using the outputted docking hit lists as the input into SEA. This technique would define proteins not by their sequence, or known ligands, but by the information generated from integrating the structure, binding site and small molecules with docking, namely the top hit list for each protein. Relating proteins by their docking hit lists addresses some of the limitations of the SEA method thus far by not needing annotated ligands. This would especially be useful for targets

that do not have any known function or have few known ligands. This was the essence of the "crazy goals".

The second focus was to branch out from "just so"[14] stories of finding ligands that share targets that are from different families to focusing on relating proteins not only pharmacologically, but also phenotypically. We were motivated to find a deeper biological relationship between proteins we related pharmacologically through choosing only those proteins that share a phenotype and investigate whether the related proteins were in the same signaling pathway or separate, but end up presenting with the same phenotype or outcome when perturbed. The idea stemmed out from a paper McGary ML, et. al.[15] that proposed a systematic way to map phenotypes from orthologous proteins of model organisms onto human diseases and phenotypes. It was then that Brian suggested adding on a phenotype layer to what became the standard SEA workflow for comparing all proteins from ChEMBL[16] pharmacologically. It was with this piece that the "crazy goals (II)" were planned out for this thesis.

As part of the training for learning how to use DOCK to extract docking hit lists for use in SEA, I was also given a docking project for the MOR[17]. What at first was what seemed like a straightforward project, it ended up being much more than finding just inhibitors but also biasing the signaling pathways that activate as a result of ligand binding towards the analgesia phenotype and not the constipation phenotype. I wanted to emphasize the pathway and systems aspect of the receptor to tie in with the systems pharmacology theme of this thesis.

II. Guide to the chapters

This thesis is made up of four major chapters based on published first-author or future co-first-author papers. The bulk of the chapters are yet to be published as they are still ongoing projects with projected future directions. Each chapter is prefaced with a short "gloss" that is both a summary and the research context that ties all the chapters together.

The first chapter provides the foundation for this thesis and presents how the sequence based organization of GPCRs differs from a pharmacological one and how the pharmacological one is not limited to within a protein family but can be used to relate proteins together between families. The examples illustrate how we can predict new ligands for proteins that did not share a ligand (to the best of our knowledge) whether within the GPCR family or between a GPCR and a *non*-GPCR. The big ideas of the paper are presented in the discussion on which we expand on in the other chapters. It was published in February 2013.

The second chapter is a brief digression from predicting protein relationships systematically but instead is an in depth study of just one GPCR, MOR, through docking. Nonetheless, we see how even one protein can respond and activate different downstream proteins and pathways just based on the ligand that is bound to it. This work is still ongoing as more mouse studies and other experiments are needed before publication.

Chapter three explores using docking hit lists as surrogates for annotated ligands for proteins and asks how well the relationships predicted using SEA with docking hit lists compares with annotated ligands. The concordance between the two is discussed and where there is discordance, predictions were made and tested for new shared ligands for targets that

do not (to the best of our knowledge) have known shared ligands. This work is also still ongoing and future directions are discussed with how to proceed with the project.

Overall, we were led by the results and a major section of the thesis is devoted to chapter four since many of the ideas presented in this introduction are thoroughly explained and illustrated through examples that worked. Here the phenotype layer is added onto the pharmacologic organization of proteins where we relate proteins that not only have similar, but not identical, ligands and have similar phenotypes. The biological implications of these methods are addressed as well as future directions with predictions and methods to be tested.

III. References

1. Keiser, M. J. *et al*. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197-206 (2007).

2. Rogers, D. J. & Tanimoto, T. T. A computer program for classifying plants. *Science* **132**, 1115-1118 (1960).

3. Lorber, D. M. & Shoichet, B. K. Hierarchical docking of databases of multiple ligand conformations. *Current topics in medicinal chemistry* **5**, 739-749 (2005).

4. Mysinger, M. M. & Shoichet, B. K. Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *Journal of Chemical Information and Modeling* **50**, 1561-1573 (2010).

5. Hopkins, A. L., Mason, J. S. & Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **16**, 127-136 (2006).

6. Lounkine, E. *et al*. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361-367 (2012).

7. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Reviews Drug Discovery* **3**, 353-359 (2004).

8. Knight, Z. A., Lin, H. & Shokat, K. M. Targeting the cancer kinome through polypharmacology. *Nature Reviews Cancer* **10**, 130-137 (2010).

9. Morphy, R., Kay, C. & Rankovic, Z. From magic bullets to designed multiple ligands. *Drug Discov. Today* **9**, 641-651 (2004).

10. Keiser, M. J., Irwin, J. J. & Shoichet, B. K. The chemical basis of pharmacology. *Biochemistry (N. Y.)* **49**, 10267-10276 (2010).

11. Lin, H., Sassano, M. F., Roth, B. L. & Shoichet, B. K. A pharmacological organization of G protein-coupled receptors. *Nature methods* **10**, 140-146 (2013).

12. Gloriam, D. E., Foord, S. M., Blaney, F. E. & Garland, S. L. Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design. *J. Med. Chem.* **52**, 4429-4442 (2009).

13. Bohn, L. M., Gainetdinov, R. R., Lin, F., Lefkowitz, R. J. & Caron, M. G. μ-Opioid receptor desensitization by β-arrestin-2 determines morphine tolerance but not dependence. *Nature* **408**, 720-723 (2000).

14. Kipling, R. in *Just so stories for little children* (Oxford University Press, 1998).

15. McGary, K. L. *et al*. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences* **107**, 6544-6549 (2010).

16. Gaulton, A. *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100-D1107 (2012).

17. Manglik, A. *et al*. Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature* **485**, 321-326 (2012).

Gloss to Chapter 1

The similarity ensemble approach (SEA)[1] systematically compares protein targets by their ligands. This method brings a bioinformatics and statistical engine to chemical similarity with new applications continuing to evolve. Its approach of beginning not with targets and seeking potential ligands but rather beginning with ligands and seeking potential targets rediscovers classical pharmacology. An iconic image in modern pharmacology is the sequence-based kinome tree[2]; it and related maps are used to guide ligand specificity. However, many drugs and reagents modulate multiple, sequence-unrelated targets, suggesting that sequence may be an imperfect guide to specificity.[3-5] Thus far, crossing of major target boundaries has been explored one drug at a time. In this chapter, we construct actionable, target family dendograms from ligand similarity that are based on multiple, sometimes thousands, of ligands being compared to each other in a systematic way.

We compared sequence- and ligand-based dendograms for the family A GPCRs, 146 of which have at least several annotated ligands. Compared to the sequence-based dendrogram, the ligand-based tree suffers major excursions. For example, the muscarinic receptors detach from the bioamines and shift toward the chemokines, while the β adrenergic receptors separate from the α adrenergic receptors and indeed other biogenic amine GPCRs. For targets that share eponymous ligands, and control some of the same physiology, this was surprising. Other rearrangements, though covering just as much distance, seem easier to reconcile with the biology they control.

To test the utility of the dendrograms and pharmacological reorganization of the GPCRs, pairs of GPCRs were chosen for direct testing. The lipid cannabinoid receptor 2 (CNR2) ligands were related to neuropeptide Y5 (NPY5R) ligands with a SEA E-value of 1.1×10^{-9}, though their sites share only 7% identity, or 3 residues. We identified a known CNR2 agonist with an EC_{50} of 110 nM[6] and upon experimental testing against NPY5R, it turned out to be an antagonist with a K_i of 8.5 nM. With this and other successes, we aimed to relate GPCRs with *non*-GPCRs again by ligand similarity.

For many of the GPCR:*non*-GPCR pairs, though the ligand sets were similar, there was no single identical ligand between them. For these we sought to predict and test ligands that would bind to both targets. An interesting class of association was one that linked targets implicated for similar clinical indications. Among these were the CB2 cannabinoid receptor (CNR2) and the enzyme Epoxide hydralase 2 (HYES). Inhibitors of either protein have been suggested for use as cardioprotectants, and HYES is in the degradative pathway for endocannabinoids as well[7]. A known HYES inhibitor was predicted to be a potential CNR2 ligand based on its E-value of 1.3×10^{-18} to the overall CNR2 set. On *in vitro* testing, we found it to be a dual CNR1 and CNR2 binder with K_i of 3.6 and 2.3 μM, respectively.

Similar to primary signaling molecules, we also found ligands that modulate multiple proteins in the same pathway even though the proteins themselves are from different families and function in different time domains. Our explanation for this phenomenon was that proteins face chemical pressures to evolve and adapt with new functions through mutations. Since there are relatively only a small number of signaling and endogenous molecules, proteins must

respond to the limited repertoire of ligands through mutations to bind and function through these molecules[8]. We assert that cellular regulation can still be controlled because the different proteins that bind the same molecules respond either faster (e.g. ion channels) or slower (e.g. nuclear hormone receptors) depending on the particular function and pathway.

I. References

1. Keiser, M. J. *et al*. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197-206 (2007).

2. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912-1934 (2002).

3. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Reviews Drug Discovery* **3**, 353-359 (2004).

4. Lounkine, E. *et al*. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361-367 (2012).

5. Keiser, M. J. *et al*. Predicting new molecular targets for known drugs. *Nature* **462**, 175-181 (2009).

6. Cheng, Y. *et al*. Discovery and optimization of a novel series of N-arylamide oxadiazoles as potent, highly selective and orally bioavailable cannabinoid receptor 2 (CB2) agonists. *J. Med. Chem.* **51**, 5019-5034 (2008).

7. Snider, N. T., Walker, V. J. & Hollenberg, P. F. Oxidation of the endogenous cannabinoid arachidonoyl ethanolamide by the cytochrome P450 monooxygenases: physiological and pharmacological implications. *Pharmacol. Rev.* **62**, 136-154 (2010).

8. Tomkins, G. M. The metabolic code. *Science* **189**, 760-763 (1975).

Chapter 1: A Pharmacological Organization of G Protein-coupled Receptors

Henry Lin[1,3], Maria F. Sassano[2,3], Bryan L. Roth[2]* and Brian K. Shoichet[1]*

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, California, USA.

[2]Department of Pharmacology, University of North Carolina Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA.

[3]These authors contributed equally to this work.

* Corresponding Authors:

Brian Shoichet: shoichet@cgl.ucsf.edu, phone: 415-514-4126, fax: 415-514-4126

Bryan Roth: bryan_roth@med.unc.edu, phone: 919-966-7535, fax: 919-843-5788

1.1 Abstract

Protein classification typically uses structural, sequence, or functional similarity. Here we introduce an orthogonal method that organizes proteins by ligand similarity, focusing here on the class A G protein-coupled receptor (GPCR) protein family. Comparing a ligand-based dendogram to a sequence-based one, we sought examples of GPCRs that were distantly linked by sequence but neighbors by ligand similarity. Experimental testing of compounds predicted to link three of these new pairs confirmed the predicted association, with potencies ranging from the low-nanomolar to low-micromolar. We then identified hundreds of *non*-GPCRs closely related to GPCRs by ligand similarity, including the CXCR2 chemokine receptor to Casein kinase I, the cannabinoid receptors to epoxide hydrolase 2, and the $\alpha_2$adrenergic receptor to phospholipase D. These, too, were confirmed experimentally. Ligand similarities among these targets may reflect a chemical integration in the time domain of molecular signaling.

Protein classification typically uses structural, sequence, or functional similarity. Here we introduce an orthogonal method that organizes proteins by ligand similarity, focusing here on the class A G protein-coupled receptor (GPCR) protein family. Comparing a ligand-based dendogram to a sequence-based one, we sought examples of GPCRs that were distantly linked by sequence but neighbors by ligand similarity. Experimental testing of compounds predicted to link three of these new pairs confirmed the predicted association, with potencies ranging from the low-nanomolar to low-micromolar. We then identified hundreds of *non*-GPCRs closely related to GPCRs by ligand similarity, including the CXCR2 chemokine receptor to Casein kinase I, the cannabinoid receptors to epoxide hydrolase 2, and the $\alpha_2$adrenergic receptor to

phospholipase D. These, too, were confirmed experimentally. Ligand similarities among these targets may reflect a chemical integration in the time domain of molecular signaling.

1.2 Introduction

Since the molecular biology revolution, proteins have been related to each other bioinformatically by either sequence or structural similarities.[1, 2] When we seek to understand the ligand recognition of a protein or the specificity of a drug or a reagent, we typically consider those proteins that are related structurally, functionally[3], or by sequence[4]. Correspondingly, methods and databases of protein families such as Pfam[5] and TRIBE-MCL[6] rely on multiple sequence alignments and machine learning to classify protein families.

Ligand recognition does not always respect such molecular biology metrics. For instance, acetylcholine and serotonin signal both through G protein-coupled receptors (GPCRs) and ion channels, which are unrelated by sequence or structure. Both ligands are also recognized by specific transporters, which are, in turn, are unrelated to GPCRs and ion channels. In addition, drugs like alosetron, which target the ionotropic serotonin receptors (HTR3), also modulate the metabotropic serotonin receptors (e.g. HTR2B, HTR4),[7, 8] while serotonergic GPCR-targeting drugs also modulate the serotonin transporter.[9] Ligands that modulate bile acid nuclear hormone receptor (NR1H4) also modulate the G protein-coupled bile acid receptor (GPBAR1)[10]. Inhibitors of enzymes, from reverse transcriptases to kinases to proteases, can also modulate GPCRs and nuclear hormone receptors.[9, 11-13]

We thus wondered how a quantitative ligand-based organization of pharmacological targets might differ from the more familiar sequence- and structure-based approaches. It is easy, after all, to build a “just-so” story with a few selected cases, such as acetylcholine and serotonin, but to understand whether a ligand-based relationship among targets will

substantially differ from a sequence-based one, the two schemes must be compared globally and quantitatively. Since sequence and structure comparisons are restricted to targets within a single, evolutionarily related target family, we will focus our attention on class A (rhodopsin-like) GPCRs. These targets are recommended by their abundance--about 700 genes in the human genome[14]— and the substantial number that have annotated ligands.

Here we ask the following questions: how different is a sequence-based organization of the class A GPCRs from one based on ligand similarity? Do the differences explain non-obvious aspects of target pharmacology and drug discovery? Can we use the ligand-based organization *prospectively*, to predict and test new associations among previously unrelated targets? Whereas we and others[15-17] have used ligand-based metrics to predict the activities of individual drugs against off-targets, this is, to our knowledge, the first effort to compare pharmacological relationships across an entire family of targets. The associations that emerge are startling: some GPCRs that are distant by sequence identity become neighbors by ligand similarity, while others that are neighbors by sequence are pushed far apart by the dissimilarity of their ligand sets. The ligand-based target similarities also suggest new associations among receptors that are, for the first time, predicted and demonstrated to share ligands. Because these associations are based on ligand similarities, they may be expanded to explore the polypharmacology between GPCRs and *non*-GPCRs, which are wholly unrelated by sequence and structure. An emergent property of these associations is that they recapitulate the activities of the cognate primary messengers, which also cross major target boundaries. This may reflect relationships in the time domain of molecular signaling, where ligand chemistry, not receptor sequence, is conserved.

1.3 Results

There were 146 class A GPCRs with at least six ligands in the ChEMBL database, which annotates ligands to targets based on literature reports.[18] On average each GPCR had 608 ligands, with a median of 380. Whereas this list captures a minority of the roughly 700 class A GPCR members[14], all of the major sub-families are included, such as the biogenic amine receptors, the peptide-receptors, the lipid-activated GPCRs, and receptors responding to protein ligands.

These GPCRs were organized by sequence and by ligand similarity. To focus on the part of the sequence most implicated in ligand binding, we only used those residues previously mapped to one of 43 orthosteric sites.[19] Sequence distances between any pair of targets was measured using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm, [20] and rendered using FigTree (http://tree.bio.ed.ac.uk/software/figtree/) (**Figure 1.1a**); notwithstanding the focus on sequence identity in the binding site, the relationships that emerge resemble those based on dendrograms using full receptor sequence identity (e.g., http://gpcr.scripps.edu/).

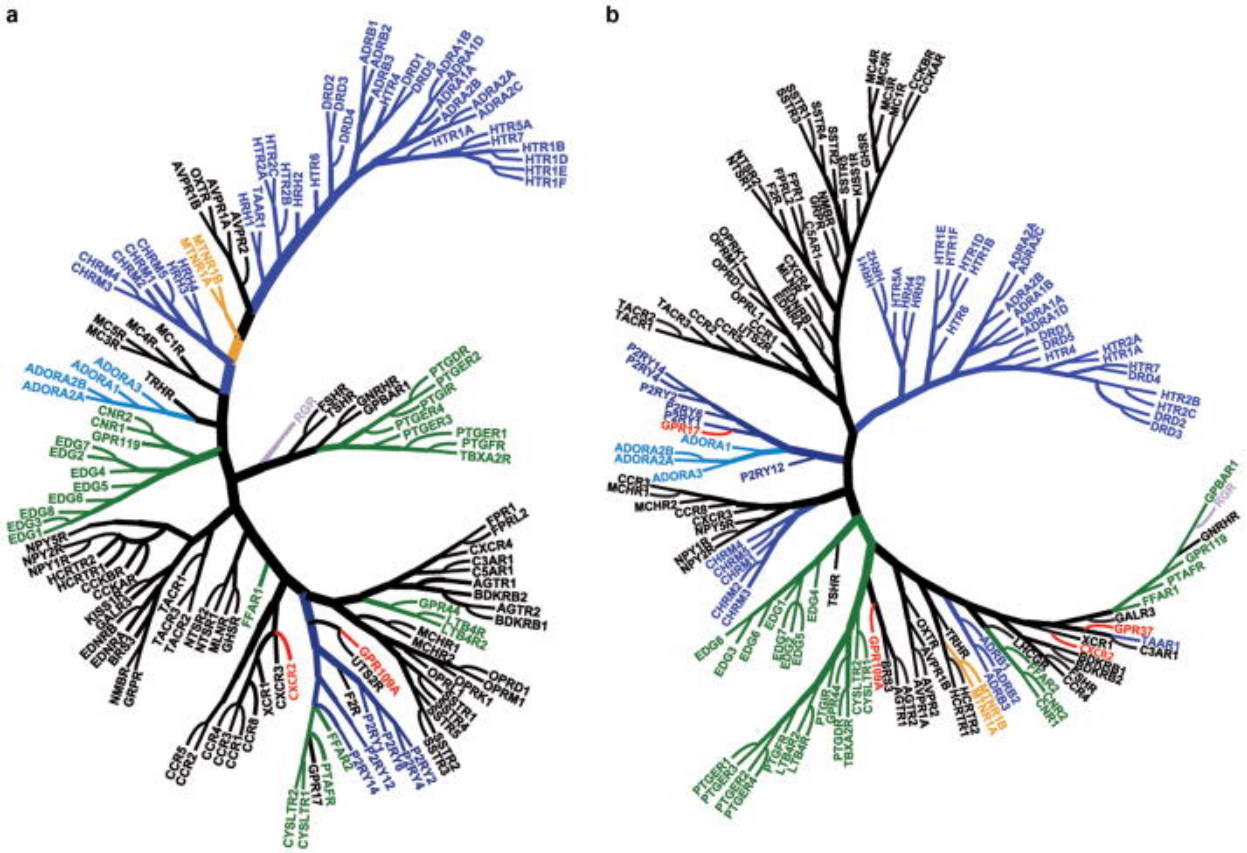


Figure 1.1
Dendrograms of human GPCRs with annotated ligands from ChEMBL.
Organization based on (a) sequence similarity in the binding site and (b) ligand set similarity based on SEA E-values. Color coding is based on chemistry of their endogenous ligands (i.e. Bioamines (blue), melatonins (gold), lipids (green), peptides (black), purinergics (dark blue), adenosines (light blue), orphans (red).).

To associate receptors by ligand similarity, ligands were represented by topological fingerprints, which are bit strings that reflect the presence or absence of chemotypes and their chemical environment in the ligand. The similarity of these fingerprints was compared for all pairs of molecules in each ligand-set for each pair of receptors, and the overall similarity of these sets was compared to an expected random similarity using machinery drawn from the sequence algorithm Basic Local Alignment Search Tool (BLAST). Similarly, an expectation value

(E-value) can be calculated, using the Similarity Ensemble Approach (Methods).[11, 21] Ligand-based dendrograms, too, were calculated using FigTree, with the distance between pairs of GPCRs quantified by the cosine angle of their SEA E-values (**Figure 1.1b**).

In the sequence-based dendrogram, the relationships among the GPCRs are as expected. The biogenic amine receptors, including the adrenergic, dopaminergic and serotonergic GPCRs largely cluster together, as do peptidic receptors such as the chemokines and melanocortins, as do the lipid responding GPCRs. At a finer resolution, some peculiar divergences and associations begin to appear. For instance, the cysteinyl leukotriene and leukotriene B4 receptors are separated from not only the other lipid-recognizing GPCRs, but also from each other, even though they are a part of the same 5-lipoxygenase pathway involved in airway inflammation.

Compared to the sequence-based organization, the ligand-based dendrogram seems victimized by almost grotesque rearrangements. The muscarinic receptors shift away from the other biogenic amine GPCRs and toward the chemokine receptors, with which they share very little orthosteric site sequence identity (9–21% identity across all subtypes). Equally perplexing, the β-adrenergic receptors separate from the α-adrenergic receptors and indeed other biogenic amine GPCRs, moving closer to the cannabinoid lipid receptors and melatonin receptors. Other rearrangements, though covering just as much distance, seem easier to reconcile with the biology they control. Thus, the cysteinyl leukotriene and leukotriene B4 receptors move much closer to each other than they were by sequence, and now cluster with other lipid GPCRs, consistent with their roles in the same leukotriene inflammatory pathway.

Though these rearrangements seem superficially perplexing, their basis may be grasped by comparing the ligands that bind to these targets. Many GPCRs that are dissimilar by orthosteric site sequence bind similar ligands, to the point where exactly the same ligands are sometimes shared between them (Supplementary Table 1.1). For example, the opioid and somatostatin receptors shift closer to the biogenic amine receptors. Despite their sequence differences, these peptidic receptors often bind aminergic molecules. The SSTR5 somatostatin 5 receptor and the HRH1 histamine H1 receptor, for instance, share only 33% sequence identity in their binding sites, even though their ligand sets resemble one another (E-value of 9.9×10^{-8}; **Supplementary Table 1.1**). Indeed, the two receptors are modulated by several identical ligands[22] (**Supplementary Table 1.1**).

Conversely, some receptors, like the muscarinic, the β-adrenergic, and the chemokine families, separate from apparently cognate GPCRs. Based on ligand similarity, the muscarinic receptors move closer to peptidic GPCRs, such as neuropeptide Y and chemokine receptors, and to lipid GPCRs, like sphingosine phosphate and prostaglandin receptors, and *away* from the biogenic amine receptors. Thus, whereas the CHRM1 muscarinic acetylcholine receptor M1 and the MCHR1 Melanin-containing hormone receptor 1 share only 26% sequence identity in the binding site, their SEA E-value is 8.3×10^{-7} and they share several sub-micromolar ligands[23] (**Supplementary Table 1.1**). Meanwhile the muscarinics share few ligands, and little ligand-set similarity, with most bioaminergic receptors. The separation of the β- and α-adrenergics is explained by the divergence of their ligand sets. The two classes of receptors share adrenaline and noradrenaline as primary messengers, and have sequence identities ranging from 49% to 63%, but once past the small catecholamines their ligands diverge: the β-

adrenergic ligands largely resemble isoproterenol, while the α adrenergic antagonists vary widely, often characterized by larger compounds with disparate scaffolds. Meanwhile, the chemokine receptors, which form an essentially contiguous family by sequence, are split into two groups by ligand similarity. One group, characterized by CXCR4, CCR1, CCR2, and CCR5, move closer to the biogenic amine receptors, while CCR3, CCR8 and CXCR3 move closer to the muscarinics and the neuropeptide Y receptors. For instance, though CCR5 and the CHRM2 muscarinic acetylcholine receptor M2 share only 16% sequence identity in the binding site, they share over 30 antagonists in several different ligand series (**Supplementary Table 1.1**).

Emboldened by these observations, we asked if the new associations *predict* crosstalk between targets not formerly known to share ligands. Many of the new neighbors in the ligand-based dendrogram share not even a single ligand, neither in ChEMBL nor in the literature, but nevertheless are highly related by the SEA E-values of their ligand lists. One such was the link between the OPRK κ opioid receptor and the HTR2B 5-HT_{2B} serotonin receptor ligands, which resemble each other with a SEA E-value of 9.9×10^{-8} though their sites share only 28% sequence identity. A SEA-screen of the ZINC database[24] suggested that compound **1** was similar to both the OPRK and HTR2B ligands. Upon *in vitro* testing, compound **1** had a K_i of 0.9 μM to HTR2B and 1.0 μM to OPRK (**Figure 1.2**, **Table 1.1**). We note that after these experiments were concluded, another series of compounds were found by some of us, in an unrelated project, that also inhibited both targets. The chemical series that did so is unrelated to that described here.[25]

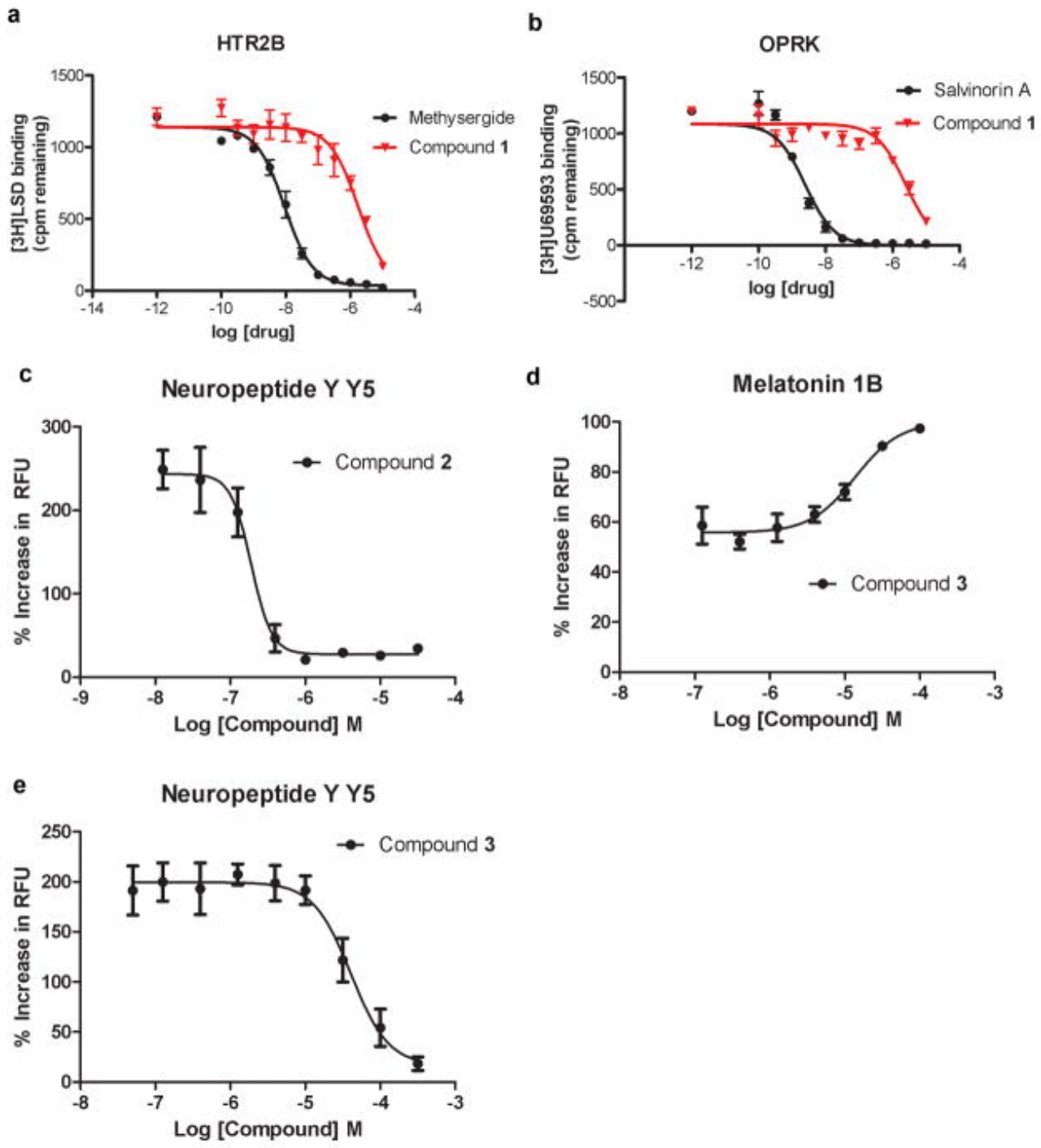
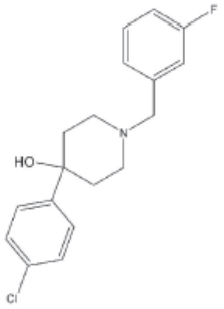
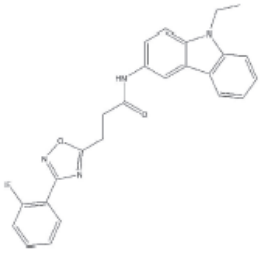
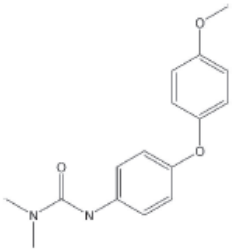


Figure 1.2
Dose-response curves of new GPCR cross-activities.
(**a–e)**Radioligand competition binding assay: compound **1** at HTR2B (**a**) and OPRK (**b**), compound **2** at NPY5R (**c**), compound **3** at MTR1B (**d**) and NPY5R (**e**). Data represent mean values ± s.e.m, performed on triplicate experiments.

Compound	GPCR1	GPCR1 affinity of predicted ligand (μM)	GPCR2	GPCR2 affinity of predicted ligand (μM)	Sequence identity (shared residues)	SEA *E*-value between ligand sets	Predicted ligand
1	5-HT_{2B} serotonin (HTR2B)	0.9[a]	κ opioid receptor (OPRK)	1.0[a]	28% (12)	9.88×10^{-8}	
2	Neuropeptide Y Y5 receptor (NPY5R)	0.0085[a]	Cannabinoid receptor 2 (CNR2)	0.14[b]	7% (3)	1.08×10^{-9}	
3	Neuropeptide Y receptor 5 (NPY5R)	1.9[a]	Melatonin 1B receptor (MTNR1B)	15[b]	21% (9)	5.33×10^{-13}	

[a] K_i. [b] EC_{50}.

Table 1.1
Predicted and confirmed ligand associations between GPCRs with low sequence identities.

If there have been many previous examples of ligand crosstalk between peptide and bioamine GPCRs, there are many fewer between peptide- and lipid-recognizing GPCRs. We were therefore interested to observe an association between the NPY5R neuropeptide Y receptor 5 and CNR2 cannabinoid receptor 2. Whereas their binding sites share only 7% identity, they had a SEA E-value of 1.1×10^{-9}. A particular CNR2 agonist, compound **2**, resembled NPY5R ligands and was commercially available (**Table 1.1**). Compound **2** was found bind to NPY5R with an IC_{50} of 190 nM (K_i = 8.5 nM), similar to its CNR2 potency (EC_{50} = 140 nM)

[26]. NPY5R was also linked to the MTNR1B Melatonin receptor 1B, in yet another GPCR sub-clade, with a SEA E-value of 5.3×10^{-13}. Here too, we found a particular MTNR1B agonist (EC_{50} = 14 μM), compound **3,** that we measured to antagonize NPY5R with a K_i of 1.9 μM.

We next asked how many of the GPCRs were strongly related by ligand similarity to a sequence-unrelated target. Interrogating all of the ChEMBL ligand sets, there were 485 *non*-GPCRs that resembled at least one GPCR in our dendrogram with an E-value of 1×10^{-10} or better (lower). Similarity values ranged from this level, for the ligand sets of the EDG7 lysophosphatidic acid GPCR and the enzyme Arachidonate 12-lipoxygenase (LOX12), to 3×10^{-314} for the ligand sets of the NTSR1 neurotensin 1 GPCR and Sortilin (SORT). These *non*-GPCR targets covered most protein families including ion channels, enzymes, kinases, and glycoproteins. Indeed, there were so many *non*-GPCR to GPCR links that clarity only allowed us to show up to two for any given GPCR (**Figure 1.3**).

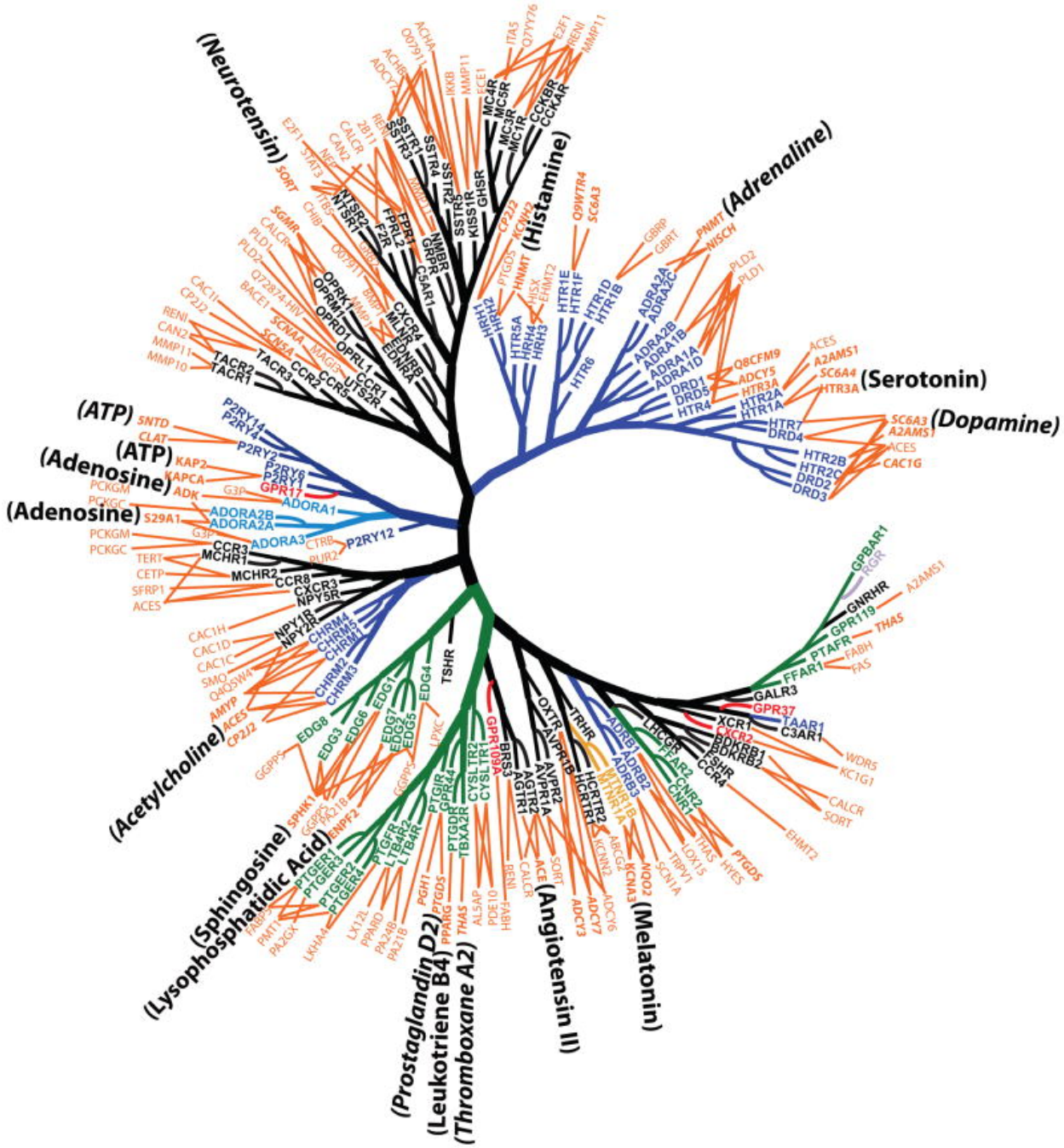


Figure 1.3

Non-GPCRs (orange) highly-related to particular GPCRs by ligand similarity (color code is as in **Figure 1.1**).

Bolded targets have known ligands that bind to both the GPCR and *non*-GPCR target. Links that share known messengers are labeled in black in parenthesis.

Here again, many highly-related pairs shared no single ligand between them, and for a few we predicted and tested ligands that would bind to both targets. We started with CXCR2 and casein kinase 1 (KC1G1), linked by a SEA E-value of 1.3×10^{-15}, and identified an inhibitor of the kinase that resembled the CXCR2 ligands. Compound **4** was tested and found to be an agonist for CXCR2 with an EC_{50} of 254 nM (**Table 1.2**, **Figure 1.4**). More ambitiously, we searched for a compound that can inhibit a GPCR and an enzyme in the same pathway. One such link was between the α2 adrenergic receptors and their downstream phospholipase D1 and D2 (PLD1 and PLD2) enzymes[27]. Compound **5**, a known phospholipase inhibitor, was tested against three α_2 adrenergic receptor subtypes and had a K_i of 556 nM to the α_{2c} sub-type (**Table 1.2**, **Figure 1.4**).

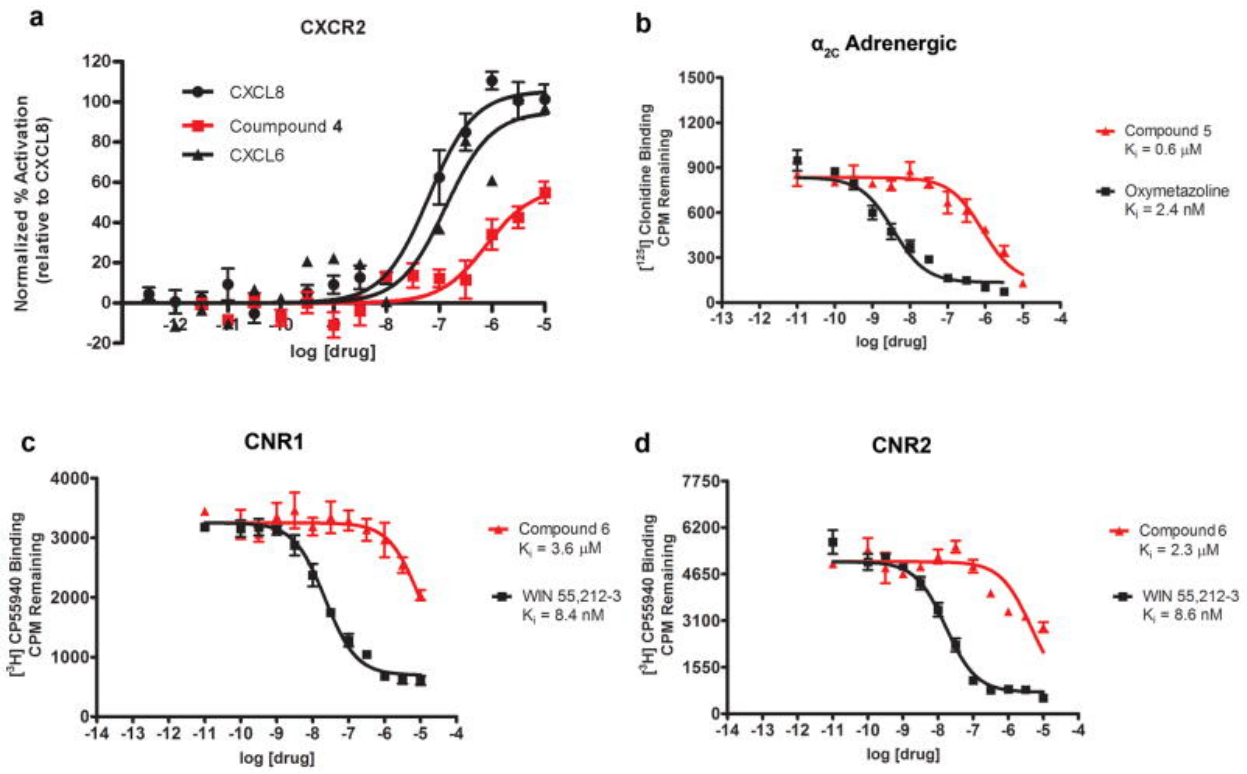
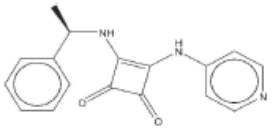
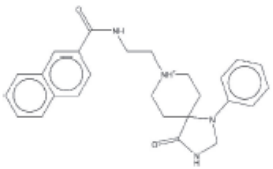
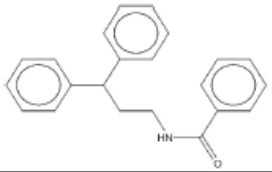


Figure 1.4

Dose-response curves of new GPCR cross-activities with non-GPCRs. Testing new GPCR cross-activities with non-GPCRs. β-Arrestin Recruitment Tango Assay: compound **4** at CXCR2 (**a**), competition binding assay: compound **5** at α_{2c} Adrenergic receptor (**b**), compound6 at CNR1 (c) and CNR2 (d). Data represent mean values ± s.e.m, performed on triplicate experiments.

Compound	GPCR	GPCR affinity (μM)	Non-GPCR	Non-GPCR affinity (μM)	SEA *E*-value	Ligand
4	C-X-C chemokine receptor type 2 (CXCR2)	0.78[a]	Casein kinase I isoform γ-1 (KC1G1)	0.25[c]	1.3×10^{-15}	
5	α_{2C}-adrenergic receptor (ADA2C)	0.6[b]	Phospholipase D2 (PLD2)	0.11[c]	9.22×10^{-15}	
6	Cannabinoid receptor 1 (CNR1)	3.6[b]	Epoxide hydrolase 2 (HYES)	0.005[c]	1.3×10^{-8}	
	Cannabinoid receptor 2 (CNR2)	2.3[b]				

[a]EC_{50}. [b]K_i. [c]IC_{50}.

Table 1.2

New GPCR associations with *non*-GPCR proteins linked by tested ligands.

Finally, we sought targets implicated not only in the same pathway, but also in a similar clinical indication. Among these were the cannabinoid receptors and the enzyme epoxide hydralase 2 (HYES), whose ligand sets have an E-value of 1.3×10^{-18}. Intriguingly, both proteins are cardioprotectant targets and both are in the endocannabinoid pathway (epoxide hydrolase 2 deactivate epoxidated endocanniboids).[28] We identified compound **6**, an HYES inhibitor, as a potential CNR2 cannabinoid receptor 2 ligand. On testing, compound **6** had K_i values of 3.6 and 2.3 μM against CNR1 and CNR2, respectively (**Table 1.2**, **Figure 1.4**).

1.4 Discussion

Relationships among targets are typically visualized by sequence-based family trees, and it is common to infer from these trees both on- and off-target pharmacology[29]. A key observation from this study is that when GPCRs are compared by ligand similarity, the arborization of the family tree changes dramatically. Targets that are neighbors by sequence are separated, while targets that are distant by sequence become neighbors. This is reflected in targets that unexpectedly respond to the same drugs and reagents, and can predict sequence-distant neighbors that will share ligands where none were previously known. The predicted and confirmed cross-activity of ligands against the κ opioid and serotonin receptors, the cannabinoid and neuropeptide Y receptors, and the neuropeptide Y and melatonin receptors, is doubly unexpected. These pairs of targets not only share little residue identity in their orthosteric sites, from 7% to 28%, but they cross target boundaries among the GPCRs: from peptide to biogenic amine, lipid to peptide, and peptide to neutral small molecule. More startling still is the observation that many *non*-GPCRs strongly resemble GPCRs by ligand similarity (**Figure 1.3**, **Table 1.2**). Whereas some of this undoubtedly reflects the conservatism of medicinal chemistry, it is impossible to look at the penumbra of *non*-GPCRs that are strongly associated with GPCRs (**Figure 1.3**) without wondering whether a more basic principle might be at work.

As sequence similarities reflect the action of evolution on proteins, the ligand-based dendrograms may reflect the chemical pressures *against which* the receptors have evolved. Many primary signaling molecules themselves target receptors unrelated by sequence or

structure. For instance, serotonin modulates both the HTR3 receptor, an ion channel, and the HTR1-2,4-7receptors, which are GPCRs. Acetylcholine targets the nicotinic receptors (ion channels) and the muscarinic receptors (GPCRs). Glutamate and GABA similarly both signal ionotropically and metabatropically. Leukotriene B4 activates GPCRs and the nuclear hormone PPARs. Estrogen binds to not only its eponymous nuclear hormone receptor but also to GPR30 (**Supplementary Table 2**).[30]

The promiscuity of primary signaling molecules reflects two constraints in biological signaling. First, cells respond to signals in multiple time domains: the millisecond, the second to minute, and the hour-to-day. To achieve this temporal resolution, they will often use ion channels, GPCRs, and nuclear hormone receptors, respectively. Second, these responses are evoked by a small repertoire of chemical messengers; once the machinery to synthesize, degrade, and regulate molecules like serotonin, acetylcholine, and estrogen is created, it is costly to change and becomes fixed.[31] On the other hand, it is relatively easy for evolution to repurpose an ion channel to recognize serotonin or acetylcholine, or a GPCR to recognize glutamate. Thus, the ability of receptors across major sequence and fold boundaries to recognize related ligands, which is captured in the ligand-based dendograms, may reflect a core chemo-evolutionary constraint in molecular signaling. If true, then probe and drug polypharmacology is neither epiphenomenal nor capricious, but reflects the evolution of signaling relationships in the time domain. Pragmatically, the associations among unrelated targets, revealed in the ligand-based dendograms, may suggest joint targets for a single molecule. Known examples are drugs that bind to both ionotropic and metabatropic serotonin receptors, like alosetron, or that bind to both muscarinic receptors and acetylcholinesterase,

like flaxedil (**Figure 1.3**). Meanwhile, the discovery that compound **6** modulates both cannabinoid GPCRs and epoxide hydrolase 2 is consistent with a role for this enzyme in the degradation pathway of the endocannabinoids, potentially arresting their signaling.[28]

Several weaknesses in this approach merit airing. Most prominently, a ligand-based view remains inference-based: targets for which no ligands are known are invisible to it, and even when ligands *are* known they can never be known perfectly, unlike the protein sequence. Mechanically, SEA remains imperfect, here, as previously,[7, 12, 32] the method had a 50% false-positive rate, with six of twelve predictions falsified by experiment (**Supplementary Table 3**). Pharmacologically, finding a ligand to modulate a GPCR and an enzyme *in vitro* does not guarantee intracellular enzyme inhibition *in vivo*, though GPCR activity of an enzyme inhibitor may be more likely. Also, we do not currently distinguish among agonists and antagonists, nor even between allosteric and orthosteric ligands; the conflation of these for a single receptor weakens the signal on which SEA operates. Meanwhile, in some protein families, such as the kinases, ligand-based and sequence-based dendrograms may resemble each other more closely than do the GPCRs, since the binding site environments are more similar and the proteins bind a single or closely related native ligand.

These cautions should not obscure the central observation from this study: a systematic and comprehensive ligand-based receptor organization differs startlingly from the more familiar sequence-based view. If this approach is weakened by ligand-based inference, it is also true that at least one other chemoinformatic approach, using only partially overlapping ligands and GPCRs, results in a dendrogram with receptor associations and disassociations that

resemble those observed here[29]. Pragmatically, ligand-based organizations of receptors offer a guide to the off-targets of tool and therapeutic molecules that is orthogonal to, but sometimes as illuminating as, the sequence-based view. More broadly, the association of 485 *non*-GPCRs with GPCRs by ligand similarity suggests a model for polypharmacology that reflects to the roles of primary messengers in cellular signaling. A virtue of this model is that it leads naturally to testable hypotheses, articulated through the very molecules that are the basis of the ligand-based organization. Some of these are suggested by the dendrograms investigated here (**Figure 1.3**).

1.5 Methods

I. Sequences and structural alignment

The initial transmembrane sequence alignments were downloaded and filtered for human sequences only. The 43 binding site residues described by Gloriam DE et. al.[19] were then extracted for all human sequences, maintaining the sequence alignments.

II. Annotated ligands

The ligands and affinity data were downloaded from ChEMBL (version 7) and filtered by their binding affinity values to create sets of ligands for targets if their IC_{50}, K_d, K_i or EC_{50} were 10μM or less. Ligands were also filtered by molecular weight (under 700), nitrogen count (fewer than eight) and oxygen count (fewer than eight) to remove large molecules and peptides. The ChEMBL database does not explicitly differentiate between agonists and antagonists for its ligands and here we combine both into the same ligand-set for each GPCR without differentiating their functional activity. 146 human GPCR sets and 2090 *non*-GPCR protein sets were assembled that each contained at least five annotated ligands and were used to compare using SEA.

III. Protein similarity calculations

Binding site sequence alignments were used to calculate relative distances between all 146 GPCRs that had ligand sets associated with them. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm was used in MEGA 3.1[33] to produce the pairwise distance

matrix between all GPCRs. Similarities ranged from just 5% to 88% identity in the binding sites with an average of 23%.

The ligand sets were also used to calculate relative distances between all 146 GPCRs by using SEA to obtain E-values between each GPCR. Each ligand was broken down into molecular fingerprints; here Extended Connectivity Fingerprints (ECFPs)[34] were used. Briefly, ECFPs are circular topological fingerprints that represent molecular structures by small atom neighborhoods or substructures, along with their physical chemical properties. The similarity between any pair of bit strings (molecules) is quantified by the bits they share in common divided by the total number of bits, via the Tanimoto coefficient (Tc)[35]. The sum of all Tc values over a certain cutoff between all the molecules in the two target-ligand sets is then calculated and compared to what we would expect for two sets of ligands, of the same set size, randomly drawn from ChEMBL. The ratio of the observed sum of Tc values to that expected at random is divided by the standard deviation of the random similarity to give a Z-score; when plotted against an extreme value distribution, this gives an expectation value (E-value). The E-values were then logged and used to calculate the pairwise cosine angle. The cosine angle was used as the distance metric, since E-values are not necessarily completely correlative with similarity, rather we use them as more of a binary measure with E-values less (better) than 1×10^{-5} as significant and anything greater (worse) taken as insignificant. Therefore, using the cosine angle, the magnitude of the E-value is not over-weighed such that E-values of 1×10^{-300} and 1×10^{-20} are treated about the same since they are both significant E-values. Similarity between GPCRs and *non*-GPCRs were calculated in the same way, using the annotated ligand sets as

surrogates for the protein to calculate SEA E-values. The two lowest E-values between each GPCR and the *non*-GPCRs were retained.

IV. Dendrograms

Using the similarity distance matrices of the binding site sequences and ligand sets, dendrograms were constructed using FigTree. The distance matrices were inputted in Newick format and a radial tree format was used for the layout. The spread was increased to better distinguish the proteins that are highly similar to each other. Nodes were further expanded out in Adobe Illustrator for legibility and color coded based on the chemistry of their endogenous ligands, e.g. peptide, bioamine, lipid, and so forth. The two *non*-GPCRs with the lowest E-value was drawn on using Adobe Illustrator and linked to their respective GPCRs. The *non*-GPCRs in bold and italicized represent those *non*-GPCRs that have a known shared ligand with the GPCR.

V. Radioligand competition binding assays

Standard techniques were used[36] at the NIMH Psychoactive Drug Screening Program.

VI. CXCR2 β-Arrestin Recruitment Tango Assay

Recruitment of β-arrestin to agonist-stimulated CXCR2 receptors was performed using a previously described "Tango"-type assay.[37] Briefly, HTLA cells stably expressing β-arrestin-TEV protease and a tetracycline transactivator-driven luciferase were plated in 10-cm dishes in DMEM containing 10% FBS and transientlytransfected (via calcium phosphate) with 10 μg of a CXCR2-V_2-TCS-tTA construct. The next day, cells were plated in white, clear-bottom, 384-well plates (Greiner; 15,000 cells/well, 50 μL/well) in DMEM containing 1% dialyzed FBS and

incubated overnight at 37°C. The following day, cells were challenged with 10 µL/well of reference agonist or CXCR2 test ligand (CXCL6 and CXCL8) at evenly distributed concentrations that ranged from 6 pM to 60 µM prepared in HBSS, 20 mM Hepes, pH 7.4, and 6% DMSO (final ligand concentrations are 1 pM to 10 µM, final DMSO concentration is 1%). After 18 h, the medium was removed and replaced with 1× BriteGlo reagent (Promega), and luminescence per well was read using a TriLux plate reader (1 s/well). Data were normalized to vehicle (0%) and reference compound (100%) controls and regressed using the sigmoidal dose-response function built into GraphPad Prism 5.0.

1.6 Acknowledgments

Supported by NIH grant GM71896 (to B.K.S. & J. Irwin) and by the National Institutes of Mental Health Psychoactive Drug Screening Program (to B.L.R.). We thank E. Gregori-Puigjane, M. Keiser and R. Coleman for reading this manuscript. B.K.S. is grateful to I.D. Kuntz and H. Bourne for illuminating conversations about time domains in molecular signaling.

1.7 Footnotes

I. Author Contributions

H.L performed the calculations. M.F.S performed experiments. B.L.R. reviewed experimental observations. H.L. and B.K.S drafted the manuscript. M.F.S and B.L.R extensively edited the manuscript. B.K.S. and B.L.R. were provoked by editorial comments to explore the chemo-evolutionary constraints and time-domain signaling implied by this work.

II. Competing Financial Interest

BKS declares a competing financial interest: he is the founder of SeaChange Pharmaceuticals, which uses chemoinformatics for target prediction. All other authors declare no competing interests.

1.8 References

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

2. Orengo, C. A. *et al*. CATH–a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1109 (1997).

3. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25-29 (2000).

4. Hopf, T. A. *et al*. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607-1621 (2012).

5. Bateman, A. *et al*. The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-D141 (2004).

6. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).

7. Lounkine, E. *et al*. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361-367 (2012).

8. Morphy, R., Kay, C. & Rankovic, Z. From magic bullets to designed multiple ligands. *Drug Discov. Today* **9**, 641-651 (2004).

9. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Reviews Drug Discovery* **3**, 353-359 (2004).

10. Watanabe, M. *et al*. Bile acids induce energy expenditure by promoting intracellular thyroid hormone activation. *Nature* **439**, 484-489 (2006).

11. Keiser, M. J. *et al*. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197-206 (2007).

12. Keiser, M. J. *et al*. Predicting new molecular targets for known drugs. *Nature* **462**, 175-181 (2009).

13. Hopkins, A. L., Mason, J. S. & Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **16**, 127-136 (2006).

14. Fredriksson, R., Lagerström, M. C., Lundin, L. & Schiöth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **63**, 1256-1272 (2003).

15. Bender, A. *et al*. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2**, 861-873 (2007).

16. Lengauer, T., Lemmen, C., Rarey, M. & Zimmermann, M. Novel technologies for virtual screening. *Drug Discov. Today* **9**, 27-34 (2004).

17. Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S. & Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **24**, 805-815 (2006).

18. Gaulton, A. *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100-D1107 (2012).

19. Gloriam, D. E., Foord, S. M., Blaney, F. E. & Garland, S. L. Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design. *J. Med. Chem.* **52**, 4429-4442 (2009).

20. Sokal, R. R. & Michener, C. D. in *A statistical method for evaluating systematic relationships* (University of Kansas, 1958).

21. Hert, J., Keiser, M. J., Irwin, J. J., Oprea, T. I. & Shoichet, B. K. Quantifying the relationships among drug classes. *Journal of chemical information and modeling* **48**, 755-765 (2008).

22. Guba, W. *et al*. From astemizole to a novel hit series of small-molecule somatostatin 5 receptor antagonists via GPCR affinity profiling. *J. Med. Chem.* **50**, 6295-6298 (2007).

23. Balavoine, F. *et al*. Design and synthesis of novel hydantoin-containing melanin-concentrating hormone receptor antagonists. *Bioorg. Med. Chem. Lett.* **17**, 3754-3759 (2007).

24. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling* **52**, 1757-1768 (2012).

25. Frankowski, K. J. *et al*. N-Alkyl-octahydroisoquinolin-1-one-8-carboxamides: Selective and Nonbasic κ-Opioid Receptor Ligands. *ACS medicinal chemistry letters* **1**, 189-193 (2010).

26. Cheng, Y. *et al*. Discovery and optimization of a novel series of N-arylamide oxadiazoles as potent, highly selective and orally bioavailable cannabinoid receptor 2 (CB2) agonists. *J. Med. Chem.* **51**, 5019-5034 (2008).

27. Scott, S. A. *et al*. Design of isoform-selective phospholipase D inhibitors that modulate cancer cell invasiveness. *Nature chemical biology* **5**, 108-117 (2009).

28. Snider, N. T., Walker, V. J. & Hollenberg, P. F. Oxidation of the endogenous cannabinoid arachidonoyl ethanolamide by the cytochrome P450 monooxygenases: physiological and pharmacological implications. *Pharmacol. Rev.* **62**, 136-154 (2010).

29. van der Horst, E. *et al*. A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization. *BMC Bioinformatics* **11**, 316 (2010).

30. Revankar, C. M., Cimino, D. F., Sklar, L. A., Arterburn, J. B. & Prossnitz, E. R. A transmembrane intracellular estrogen receptor mediates rapid cell signaling. *Science* **307**, 1625-1630 (2005).

31. Tomkins, G. M. The metabolic code. *Science* **189**, 760-763 (1975).

32. Gregori-Puigjané, E. *et al*. Identifying mechanism-of-action targets for drugs and probes. *Proceedings of the National Academy of Sciences* **109**, 11178-11183 (2012).

33. Kumar, S., Tamura, K. & Nei, M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in bioinformatics* **5**, 150-163 (2004).

34. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742-754 (2010).

35. Rogers, D. J. & Tanimoto, T. T. A computer program for classifying plants. *Science* **132**, 1115-1118 (1960).

36. Jensen, N. H. *et al*. N-desalkylquetiapine, a potent norepinephrine reuptake inhibitor and partial 5-HT1A agonist, as a putative mediator of quetiapine's antidepressant activity. *Neuropsychopharmacology* **33**, 2303-2312 (2007).

37. Chen, X. *et al*. Structure–Functional Selectivity Relationship Studies of β-Arrestin-Biased Dopamine D2 Receptor Agonists. *J. Med. Chem.* **55**, 7141-7153 (2012).

Gloss to Chapter 2

Morphine is the most widely prescribed drug in the United States with over 41 tons consumed each year since it is very effective at relieving severe pain and acts as an analgesic. The target of morphine is the μ-opioid receptor (MOR) and its structure was recently solved by x-ray crystallography in the laboratory of Brian Kobilka[1]. With recent success in our lab of docking to G protein-coupled receptors (GPCRs)[2-5], we aimed to find new inhibitors of MOR. The MOR binding site is a large, wide open cavity that binds a large variety of ligands from endogenous peptides called enkephalins and endorphins to small molecule ligands like morphine. The structure was solved with several crystal waters coordinating the protein to the morphanin antagonist, which complicates the docking in terms of knowing which waters to include for docking. A problem we encountered was getting low enrichment when docking known ligands into the binding site. We adjusted certain residues in terms of their electrostatics in a technique we call "tarting" that increases the electrostatics potential so certain ligands that have complementary matching charges with the protein residues will have a better overall calculated energy. It was then that we found the right combination of residue "tarting" and finally achieved positive enrichment of known ligands. We felt comfortable enough with the parameters to begin docking.

The first set of ligands we docked and tested resulted in 7 molecules that showed dose response binding out of 23 molecules we tested, 30.4% hit rate. Despite the high hit rate, the K_i range was between 2.3 μM and 13.8 μM so we did not have any tight binders relative to the potency of morphine and other opiates. Next we went through a round of SAR by catalog[6] to

find analogs of the best initial hits by docking all the analogs we could source and picked 15 analogs for experimental testing. The best analog had a K_i of 42 nM and we moved forward on the compound in collaboration with Bryan Roth's lab for functional studies. As it turns out, the compound was active in cells and quite potent as an agonist with an average EC_{50} of 177 ± 93 nM with 98 ± 4% efficacy.

With such unexpectedly positive results, we wanted to take the project a step further by interrogating the signaling pathways of MOR. Agonists of MOR are great at producing an analgesic effect but have a slew of side effects that come along with taking the drugs such as constipation, respiratory depression and euphoria[7]. Recent publications suggest that the side effects are a result of agonists signaling through β-arrestin and the analgesic effect is triggered by G-protein signaling[7, 8]. If we can identify a G-protein biased agonist that does not trigger β-arrestin binding then the idea is there would be the pain relief effect without the unwanted side effects. As it turns out, our ligand was G-protein biased and showed little effect signaling through β-arrestin. In the hot plate analgesia mouse model, our ligand showed significant analgesia over vehicle and the effect was reversible by an antagonist, naloxone. This indicated that the ligand's effect on the mice were through the MOR mechanism. Although the G-protein biased mechanism has not been fully validated by the community, the success of the current molecules in the clinic that also have these properties will be key. Nonetheless, we are optimistic in this pathway and mechanism and were pleasantly surprised that the molecule was discovered by docking.

I. References

1. Manglik, A. *et al*. Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature* **485**, 321-326 (2012).

2. Kolb, P. *et al*. Structure-based discovery of β2-adrenergic receptor ligands. *Proceedings of the National Academy of Sciences* **106**, 6843-6848 (2009).

3. Carlsson, J. *et al*. Structure-Based Discovery of A2A Adenosine Receptor Ligands. *Journal of Medicinal Chemistry* **53**, 3748-3755 (2010).

4. Carlsson, J. *et al*. Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nat Chem Biol* **7**, 769-778 (2011).

5. Mysinger, M. M. *et al*. Structure-based ligand discovery for the protein–protein interface of chemokine receptor CXCR4. *Proceedings of the National Academy of Sciences* **109**, 5517-5522 (2012).

6. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* **52**, 1757-1768 (2012).

7. Bohn, L. M. *et al*. Enhanced Morphine Analgesia in Mice Lacking β-Arrestin 2. *Science* **286**, 2495-2498 (1999).

8. DeWire, S. M. *et al*. A G Protein-Biased Ligand at the μ-Opioid Receptor Is Potently Analgesic with Reduced Gastrointestinal and Respiratory Dysfunction Compared with Morphine. *Journal of Pharmacology and Experimental Therapeutics* **344**, 708-717 (2013).

Chapter 2: Structure-based discovery of functionally selective Mu opioid receptor agonists

Henry Lin[1], Maria F. Sassano[2], Aashish Manglik[3], Kate White[2], Brian K. Kobilka[3]*, Brian K. Shoichet[1]*, and Bryan L. Roth[2]*

[1]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, USA, 94158

[2]Department of Pharmacology, National Institute of Mental Health Psychoactive Drug Screening Program, University of North Carolina, Chapel Hill, Chapel Hill, North Carolina, USA, 27759.

[3]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California, USA, 94305.

* Corresponding authors

Brian Kobilka, kobilka@stanford.edu,

Brian Shoichet: bshoichet@gmail.com, phone: 416-978-8861, fax: 416-978-8287

Bryan Roth: bryan_roth@med.unc.edu, phone: 919-966-7535, fax: 919-843-5788

2.1 Abstract

Mu opioid receptor (MOR) agonists such as morphine have long been used to treat pain, but potentially lethal side effects and addictive liabilities limit their use and effectiveness. These side effects are thought to be mediated by signaling through the arrestin pathway following receptor activation, whereas signaling through the G_i-protein pathway is thought to confer analgesia. The recent determination of the MOR structure, and of other opioid subtypes enables ligand discovery campaigns for new chemotypes which may confer G-protein biased signaling in a way that classic opioids have not. Over 3 million molecules were computationally docked against the structure of the MOR and 23 with new scaffolds were selected for testing. Seven had affinities from 2.3 to 13 μM. Optimization led to agonists with K_i values as low as 42 nM and 11- to >100-fold selectivity vs the other opioid receptors. The new scaffold appeared to confer new biology, as this agonist was strongly biased toward G-protein over arrestin signaling. This molecule substantially increased pain tolerance in a mouse model, in a manner reversible by the specific MOR antagonist naloxone. This agonist is the first G-protein-biased MOR-specific agonist discovered from structure based virtual screening that is active *in vivo*, and may serve as a probe to isolate the different components of MOR signaling, and as a lead to therapeutics without the dose-limiting side effects of the current opioid analgesics.

2.2 Introduction

The use of opioids in therapy dates to the Neolithic, and their effectiveness for pain treatment, as euphoragens, and their addictiveness has made them central to medicine, commerce, and conflict ever since. Their addictiveness and potentially lethal side effects, such as respiratory depression, have driven campaigns to improve them since the early 19th century, with the purification of morphine and codeine, and with the synthesis of heroin in the last part of that century. Indeed, these efforts might reasonably mark the beginning of modern pharmaceutical chemistry. These molecules, though more effective and more manageable than opium itself, retained the side effects of the plant product. The characterization of the opioid receptors into the μ, κ, and δ subtypes, a triumph of classical pharmacology, raised hopes for the design of subtype specific molecules without the liabilities of the opioid analgesics, but despite the introduction of potent and fast-acting drugs such as fentanyl and ultiva[1], molecules without the liabilities of the classic opioid analgesics remained difficult to find by empirical pharmaceutical chemistry. Recently, molecular studies have in fact suggested that the CNS-based analgesia relates to Mu opioid receptor (MOR) signaling through the G_i-protein pathway, while many of the pathophysiology of the opioid drugs, including respiratory depression and constipation, is conferred via arrestin pathway signaling. Both as molecular probes to understand signaling, and as therapeutic leads, agonists specific to the MOR with bias toward the G_i-protein pathway are much sought.[2, 3]

The recent determination of the crystal structures the μ, κ, δ, and nociceptin opioid receptors[4-7] provided an opportunity to seek μ agonists with new chemotypes and specificities.

Recent structure-based discovery campaigns, where large libraries of molecules are computationally docked against the structure of a receptor, have shown promise against G-protein coupled receptors (GPCRs). Hit rates—the number of experimentally active molecules divided by the number of docking-priortized molecules tested—have ranged from 17 to 70%, with potencies in the nanomolar range, including ligands with previously unknown scaffolds that nevertheless complement the receptors well. [8-16] As opioids are members of this family of membrane-bound receptors, we were encouraged to target the MOR in a docking screen. We sought ligands with new chemotypes, unrelated to previously known MOR agonists, with physical and chemical properties consistent with CNS activity *in vivo*. We hoped that the novel chemotypes of the new agonists would confer on them novel biology and signaling properties, something observed in campaigns against other targets.[17, 18]

2.3 Results

I. The μ opioid orthosteric site.

Most GPCR docking campaigns have targeted receptors responding to small endogenous signaling molecules, like adrenaline[13], histamine[11], dopamine[10, 19], and adenosine[12]. These receptors have relatively small, well-defined, and largely water-free binding sites that are well-suited to molecular docking, which scores candidate ligands for physical complementarity to the receptor. The μ-opioid receptor, conversely, is a peptide receptor with an open binding site where ordered waters dominate the interactions with the co-crystalized ligand ligand, β-funaltrexamine (β-FNA). These physical features typically result in lower hit rates and affinities in molecular docking campaigns[16, 18], and to larger ligands with physical properties less favorable for *in vivo* activity. Also, the receptor was crystallized in an inactive state, which previous GPCR work has suggested will bias toward antagonist discovery, and here we sought agonists. These features made ligand discovery against this target technically challenging. We were encouraged by two recent observations: the discovery of novel agonists specific for μ/δ receptor hetero-dimers in an empirical high-throughput screen[20], and the discovery of new agonists, with affinities in the 100 to 400 μM range, in a docking campaign against the κ opioid structure[21], a target freighted with the same challenges as the μ receptor.

II. Docking for μ-opioid receptor agonists.

Against the structure of the μ-opioid receptor[4], we screened over 3 million commercially available lead-like[22] compounds from the ZINC database[23, 24] using DOCK3.6[25, 26]. Each

compound was evaluated for complementarity to the receptor using an energy function that scores van der Waals and electrostatics interactions, and that corrects these energies for ligand desolvation by the receptor[27]. The best scoring configuration of each molecule was saved and ranked. As is common in docking[16, 18, 21] and high-throughput screens[28], the top scoring molecules, here those in the top 0.07% of the docking ranked list, were inspected for features not explicitly captured in the ranking. These included their novelty compared to the 5,215 MOR ligands annotated in ChEMBL16[29]—ECFP4-based topological[30] Tanimoto coefficients (Tc) [31] ranged from 0.28 to 0.31, consistent with the exploration of new scaffolds[32], their interactions with key polar residues such as Asp147 and Tyr148, and the absence of strained interactions internal to the small molecules themselves, a feature that is only considered in calculating ligand conformations for the ZINC flexibase[25, 33], and not further considered by DOCK3.6 itself. Ultimately, 23 high-scoring molecules, with ranks ranging from 237 to 2095, were selected and purchased for testing. All were monocations at physiological pH and docked to ion pair with the aminergic-recognizing residue Asp147. As far we know, none had previously been tested against opioid receptors. Of the 23, 7 molecules bound to the MOR with binding affinities (K_i) ranging from 2.3 μM to 14 μM by radio-ligand displacement in full dose-response (**Table 2.1**, **Figure 2.1e**).

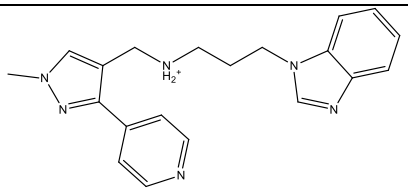
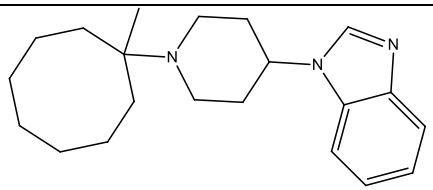
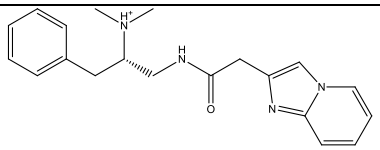
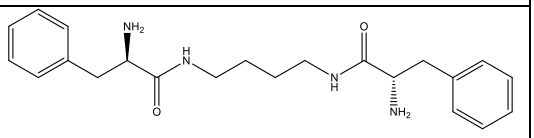
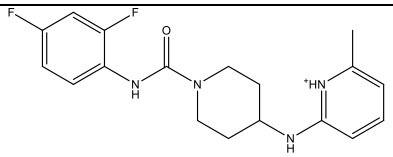
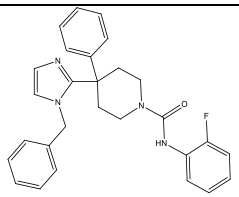
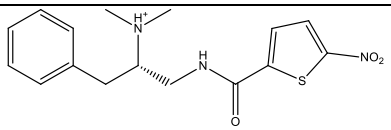
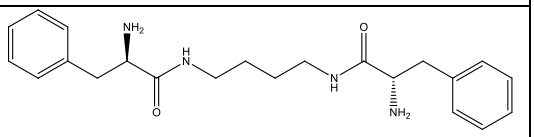
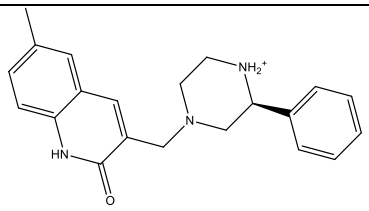
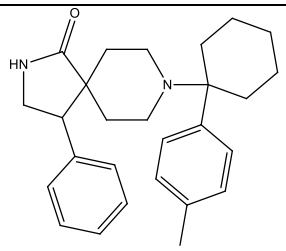
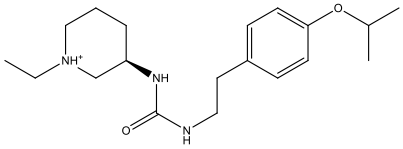
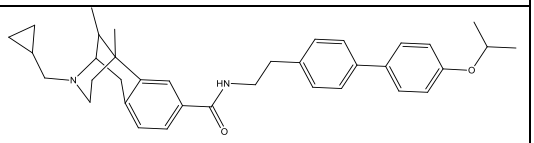
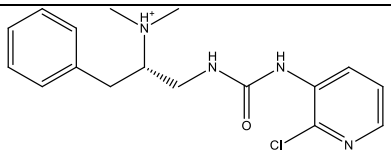
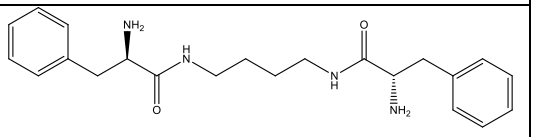
Cmpd #	Structure	T_c [a]	K_i (μM) [b]	Nearest ChEMBL MOR ligand
1		0.28	7.2	
2		0.28	5.8	
3		0.30	13.8	
4		0.30	2.3	
5		0.31	4.7	
6		0.30	10.0	
7		0.30	2.5	

Table 2.1

Docking-derived ligands from the screen against the μ-opioid receptor.

[a]The ECFP_4 Tanimoto similarity (T_c) to the most similar μ-opioid receptor ligand in the ChEMBL database.

[b]Measured affinity for the μ-opioid receptor.

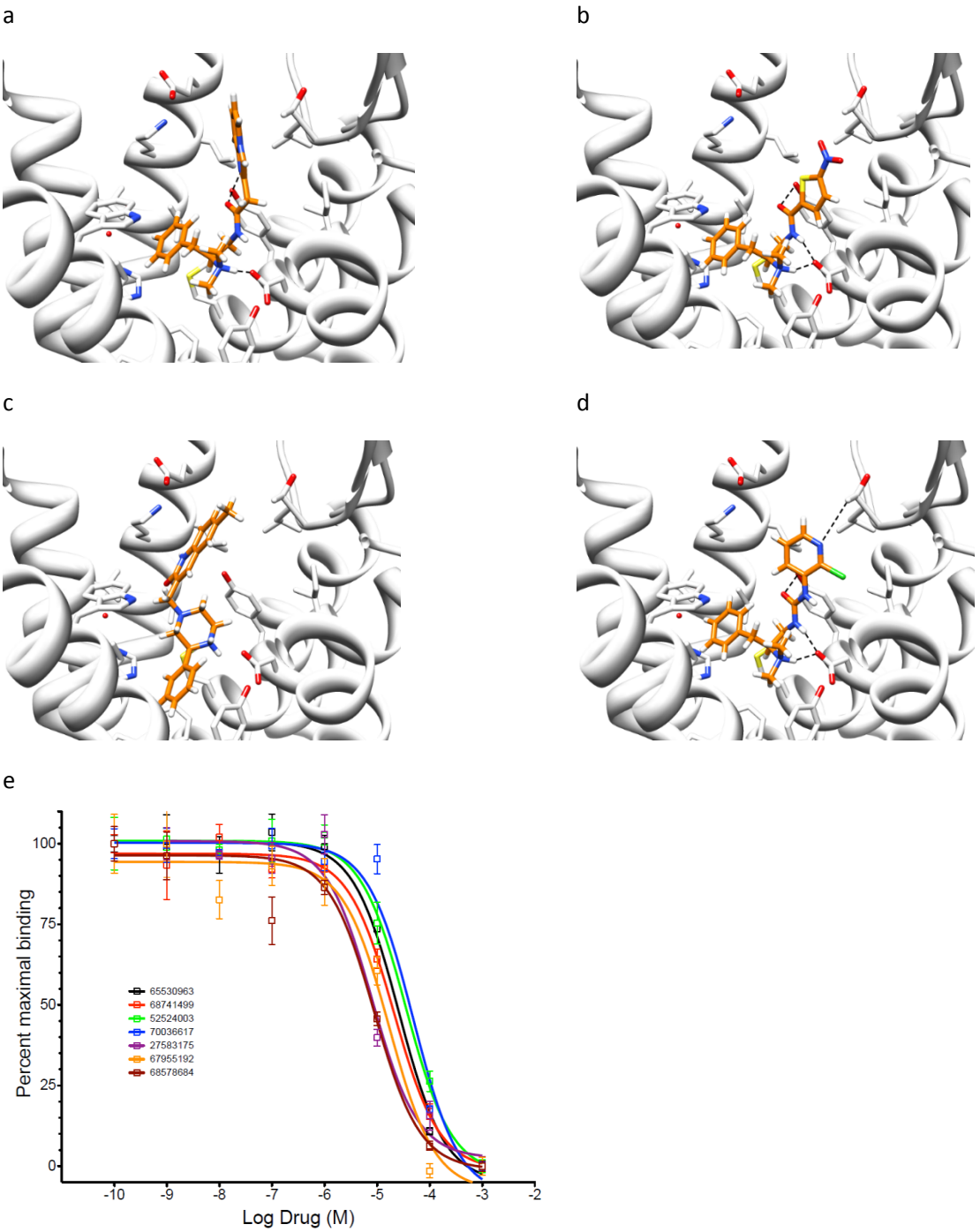


Figure 2.1

Predicted binding poses and dose response curves of initial hits from the docking screen against μ-opioid receptor.

Receptor carbons in gray, ligand carbons orange, nitrogens blue, oxygens red, sulfur yellow, chlorine green. Putative hydrogen-bonds shown as dotted lines (a) **2** (b) **4** (c) **5** (d) **7**. All figures visualized with Chimera[34]. (e) Dose response curves of initial hits from docking screen. Data represent mean values ± standard error, performed on triplicate experiments.

III. Structure-based optimization for increased affinity

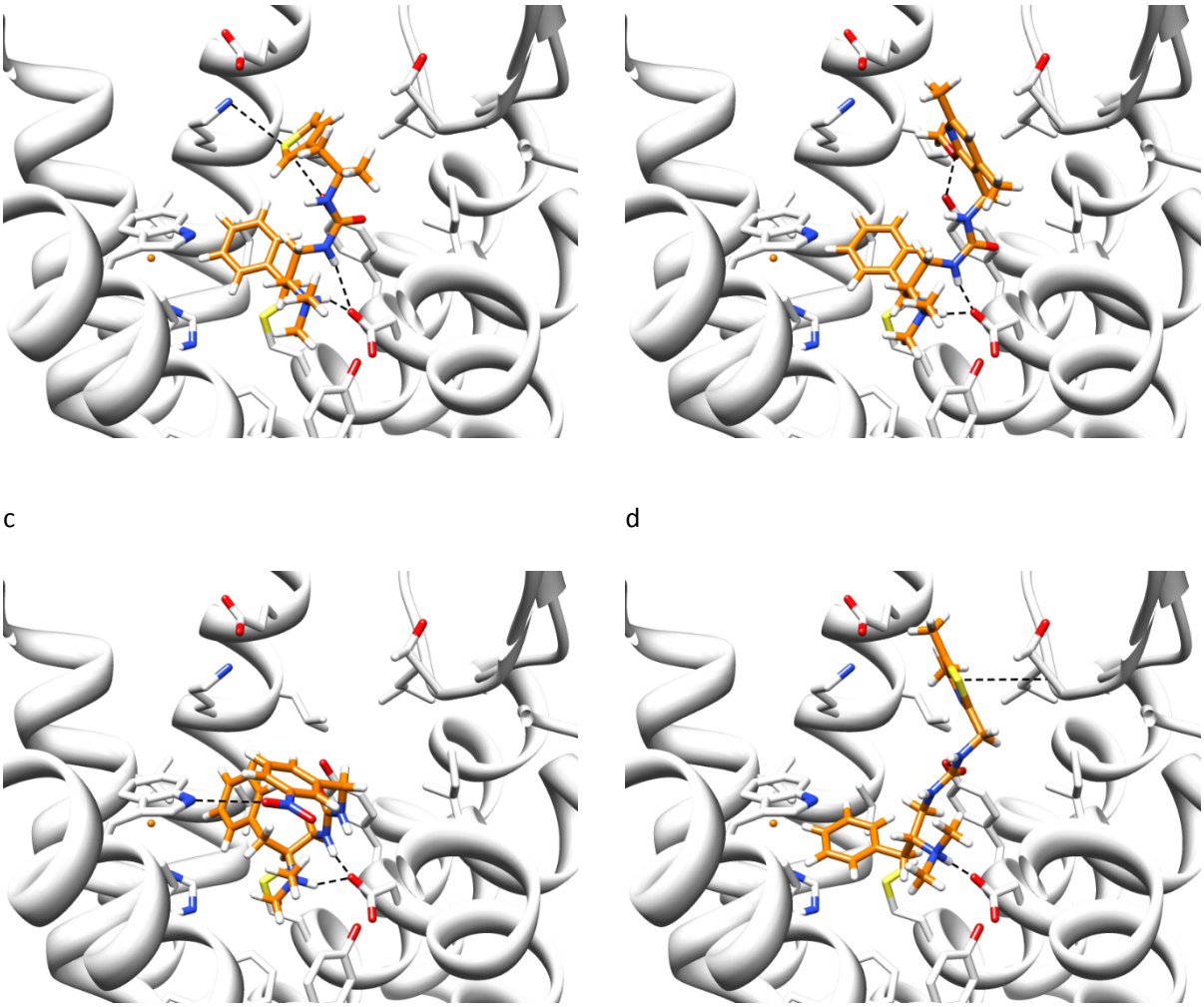
The new ligands appeared to engage with the receptor in new ways. Several used a urea or amide carbonyl to hydrogen-bond with Tyr148 (**Figure 2.1 a**, **b**, and **d**), and occasionally also engaged Thr218, interactions that are not observed in the crystallographic complex with β-FNA or not usually anticipated for the classic opioids. A common feature in the new ligands was the joint engagement of an aminergic nitrogen and an amide or urea nitrogen with the key recognition Asp147. While ion pairing with Asp147 is canonical to opioid recognition, using such a double hydrogen bond is not, and indeed ureas have little precedent among the opioid receptor ligands (there are only 50 among the over 5,215 opioid ligands in ChEMBL16[29]). On the other hand, none of the water-ligand interactions that were so common in the β-FNA complex were observed in the docked structures, simply because we only modeled two of the ordered waters in the docking calculations and did not have ligands with hydroxyl or carbonyl substituted rings in that area in ZINC.

Intrigued by these new interactions, we therefore sought analogs that would retain the key recognition groups but would optimize peripheral packing interactions, as there were several places in the docked complexes where pockets in the receptor structure were left uncomplemented by ligand groups. We searched for analogs of the most potent ligands (**2** and **7**), including those left out of the version of the ZINC lead-like library that we had initially docked, or that were outside of ZINC's lead-like definitions[24]; 500 such analogs were docked into the orthosteric site of MOR. Most of the analogs retained the same polar recognition groups as compounds **2** and **7**, but added packing substituents or extended further toward the

extracellular side of the receptor, where the opioid receptors have more sequence variability (**Figure 2.2**). We chose 15 top scoring analogs to purchase and experimentally tested them for binding to MOR and also the κ-Opioid Receptor (KOR), investigating selectivity. Of the 15 compounds, 7 bound to MOR with K_i between 42 nM and 4.7 μM (**Table 2.2, Figure 2.3**). Compared to their lead compound **7**, compounds **12** and **15** had 60-fold and 20-fold higher affinity, were 11- and 77-fold specific for MOR over KOR, respectively, while retaining good ligand efficiency (0.43 and 0.40, respectively).

a

b



c

d

Figure 2.2
Predicted binding poses for 4 analog ligands discovered in the docking screen against the μ-Opioid Recpetor; colored as in **Figure 2.1**. (a) **12** (b) **13** (c) **14** (d) **15**.

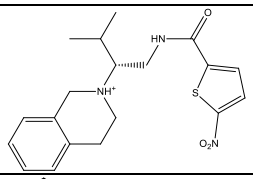
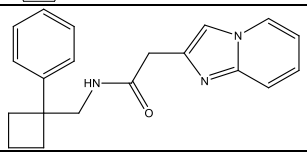
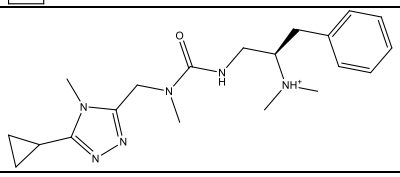
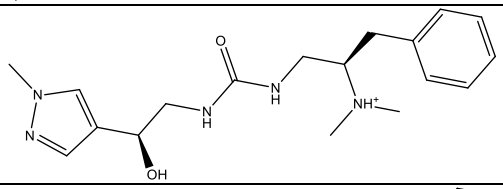
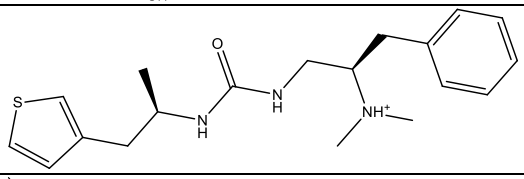
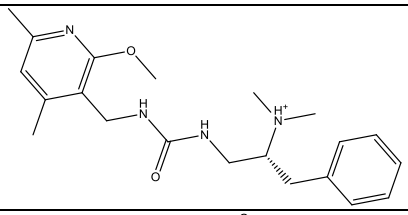
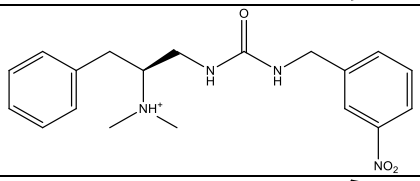
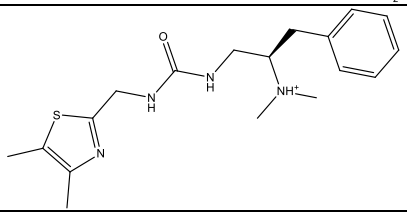
Cmpd #	Structure	MOR[a]	KOR[b]	KOR/MOR fold selectivity
8		0.817	0.455	0.56
9		<50%	1.360	<0.14
10		4.748	<50%	>2.1
11		1.858	<50%	>5.3
12		0.042	0.464	11.05
13		0.550	1.022	1.86
14		0.087	0.511	5.87
15		0.130	<50%	>76.9

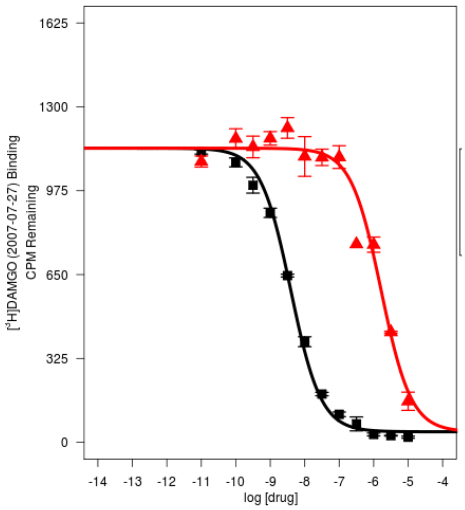
Table 2.2

Analogs of discovered ligands from the docking screen against the μ-opioid receptor and selectivity over κ-opioid receptor.

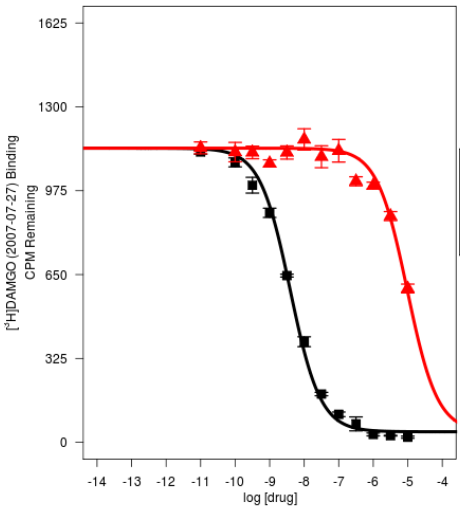
[a]Measured affinity (K_i in μM) for the μ-opioid receptor or % inhibition at 10 μM of compound.

[c]Measured affinity (K_i in μM)for the κ-opioid receptor or % inhibition at 10 μM of compound.

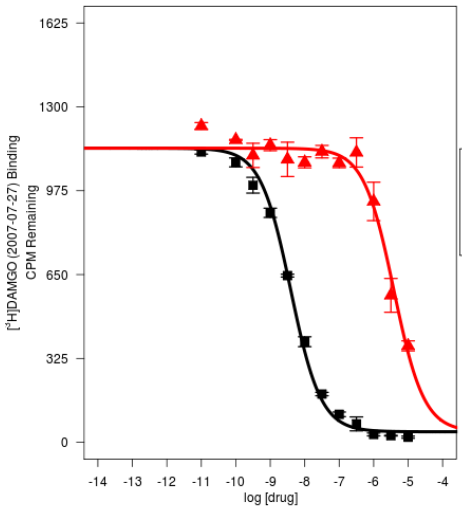
a


[³H]DAMGO (2007-07-27) Binding
CPM Remaining
1625
1300
975
650
325
0
-14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4
log [drug]

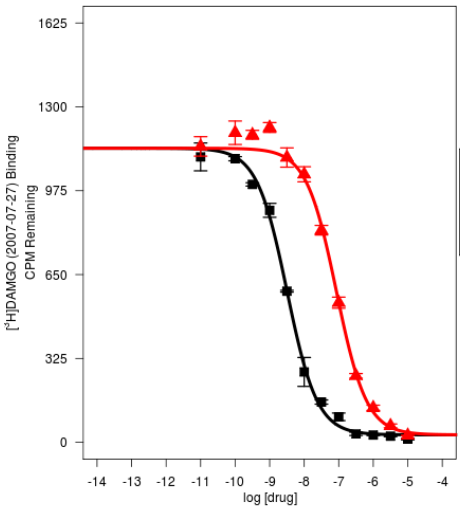
b


[³H]DAMGO (2007-07-27) Binding
CPM Remaining
1625
1300
975
650
325
0
-14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4
log [drug]

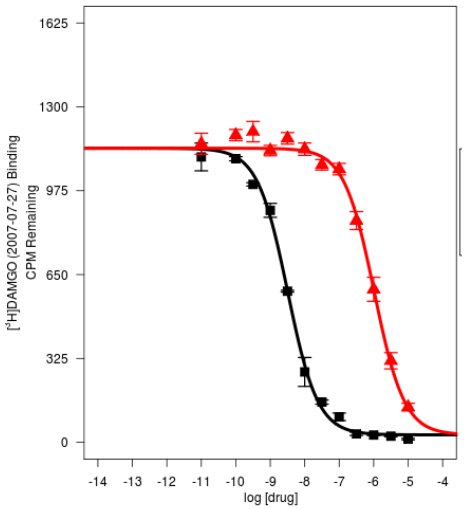
c


[³H]DAMGO (2007-07-27) Binding
CPM Remaining
1625
1300
975
650
325
0
-14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4
log [drug]

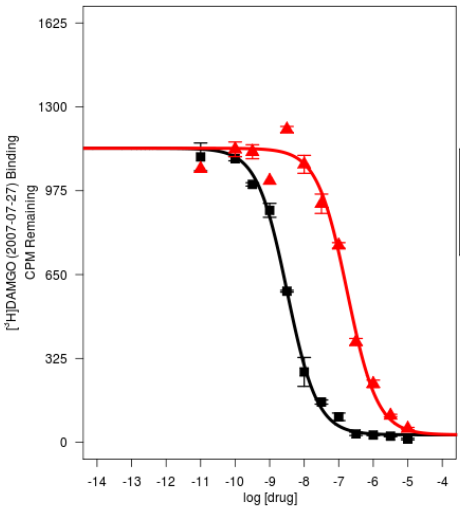
d


[³H]DAMGO (2007-07-27) Binding
CPM Remaining
1625
1300
975
650
325
0
-14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4
log [drug]

e


[³H]DAMGO (2007-07-27) Binding
CPM Remaining
1625
1300
975
650
325
0
-14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4
log [drug]

f


[³H]DAMGO (2007-07-27) Binding
CPM Remaining
1625
1300
975
650
325
0
-14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4
log [drug]

g

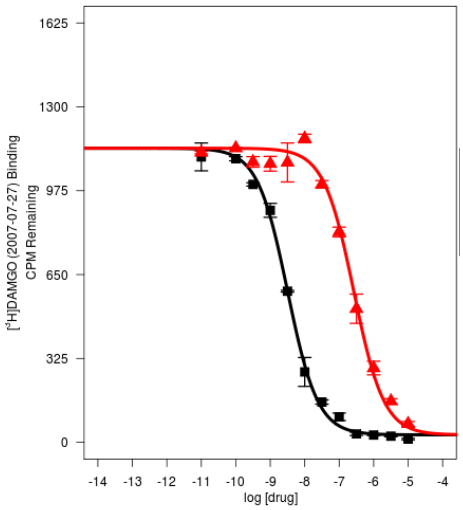


Figure 2.3

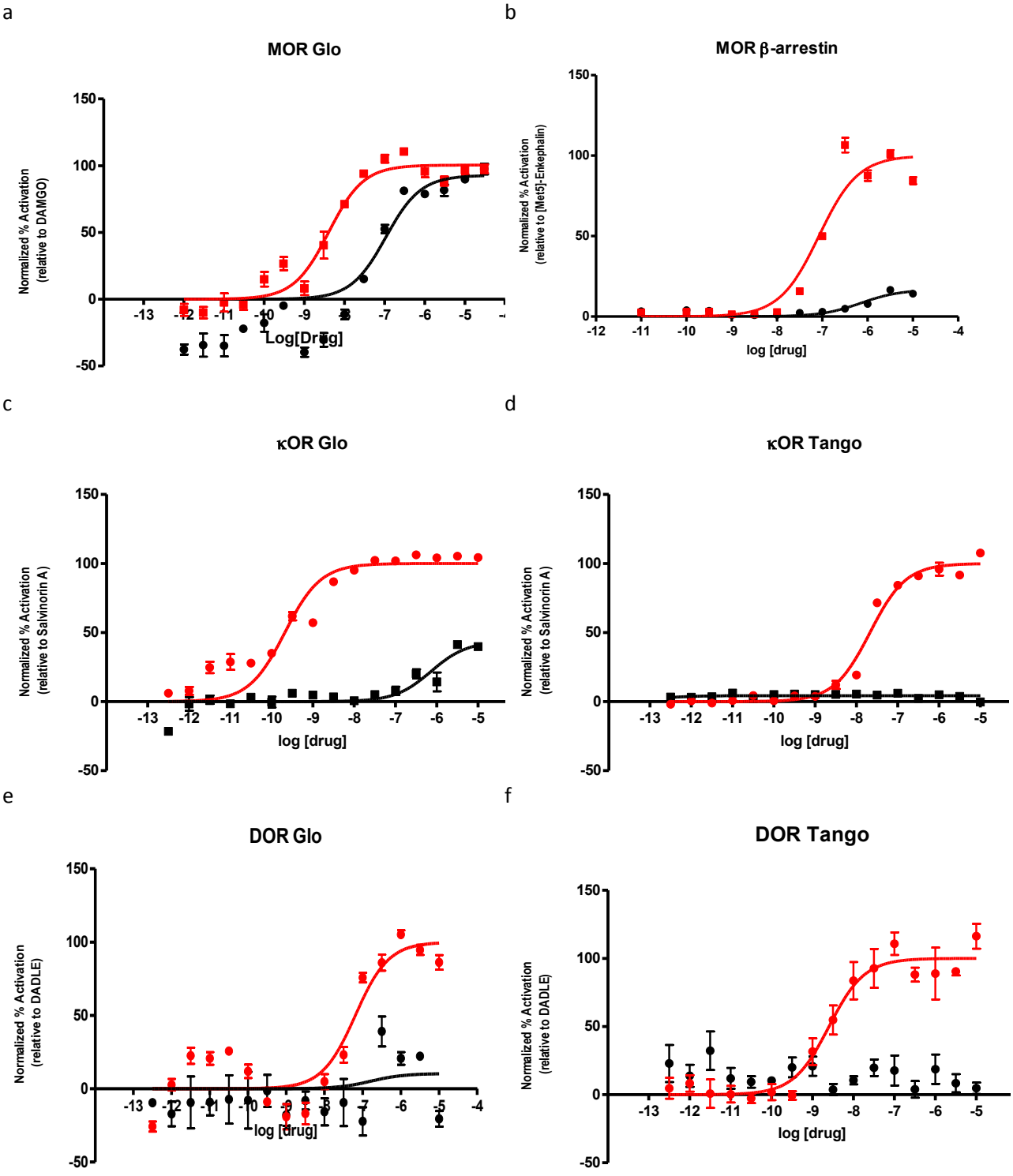
Radioligand ([3H]DAMGO) competition binding isotherms for analog compounds **8**, **10**-**15** (a-g).

Data for a reference compound (Morphine, black curve) are shown along with data for the test compound (red curve). Assays are performed using a final radioligand concentration between ($0.5 \times K_D$) and ($1 \times K_D$), where K_D equals the radioligand dissociation constant, which is determined for each crude membrane preparation by radioligand saturation binding analysis. Data represent mean values ± standard error, performed on triplicate experiments.

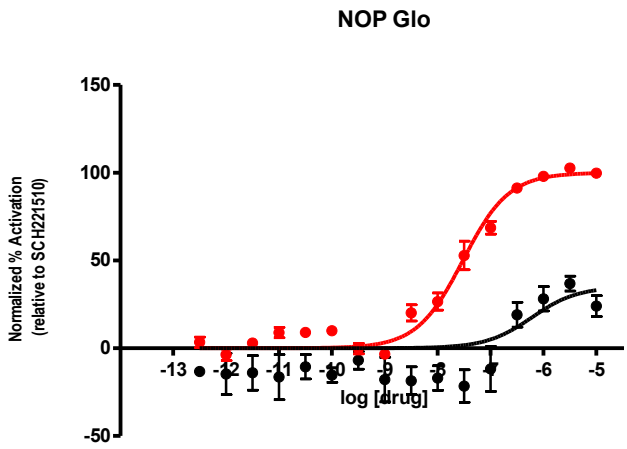
IV. Functional studies with entire opioid family

We investigated the more potent analogs for functional selectivity against the other opioid receptor subtypes. Compounds **8**, **12** to **14** activated G-protein signaling in G_i recruitment assay. Intriguingly, though consistent with a study against the KOR[21], all were agonists without measurable antagonism of MOR, even though the structure was solved with an antagonist bound. The most potent compound, **12**, was tested against MOR, KOR, δ-Opioid Receptor (DOR), Nociceptin Receptor (NOP) in both agonist and antagonist modes and for functional signaling through either G-protein or β-arrestin. This agonist had an average functional EC_{50} of 177 ± 93 nM with 98 ± 4% efficacy—versus the full agonist DAMGO—at MOR signaling through G-protein (Glosensor assay, **Figure 2.4**). In the β-arrestin recruitment assay at MOR, compound **12** had only 17% max efficacy and no measurable EC_{50} (no saturation up to 10 μM). Against KOR and NOP, **12** only achieved 50% and 40% efficacy and EC_{50} values of 1.0 μM and 1.5 μM, respectively, as G-protein agonists; no such activity was observed against the DOR (**Figure 2.4**). No antagonist activity was observed for all opioid receptors, and no β-arrestin activity for the NOP, DOR and KOR (Supplementary Figure 1). Since several opiod-active molecules also modulate other aminergic receptors (morphine itself has nanomolar activity against the α_2-adrenergic receptor), we tested **12** for activity against a representative set of these targets (Supplementary Table 1). Compound **12** had K_i values in the high nanomolar to low micromolar range against the 5-HT2a, 5-HT2B, 5-HT2C, α_{2b} adrenergic, and Dopamine D4, Histamine H1, Muscarinic M5 receptors, and the Norepinepherine, Serotonin and Dopamine transporters, and had less than 50% inhibition in a 10 μM primary screen against 27 other GPCR aminergic receptors tested. Compound **12** is a potent agonist of the μ opioid receptor with bias

toward G-protein vs the arrestin pathway, 11- to over 100-fold specificity vs. other opioid receptors, and only modest activity against other aminergic receptors.



g



h

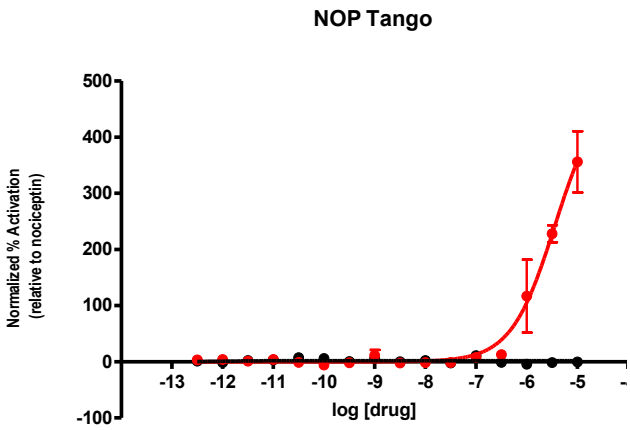


Figure 2.4

Functional assays for Compound **12** in agonist mode for MOR, KOR, DOR, NOP in G-protein activation (a,c,e,g) and β-arrestin recruitment (b,d,f,h)

V. In vivo analgesia

Encouraged by these observations, and by the favorable physical and chemical properties of **12** (mw < 350, lack of obvious toxophores or metabolic liabilities), we tested this molecule for *in vivo* analgesia in a mouse pain model. On treatment with 3mg/kg of compound **12** in the widely-used hot plate assay, mice showed a roughly 25% increased tolerance to pain compared to vehicle (N=21) (**Figure 2.5**). By comparison, 3mg/kg of morphine increases tolerance by about 50% (N=12) but not statistically significantly more than compound **12**. Co-administration of **12** with the specific μ antagonist naloxone completely reversed this effect (data not shown). These observations support an analgesic activity for **12** *in vivo* that is mediated via the μ opioid receptor.

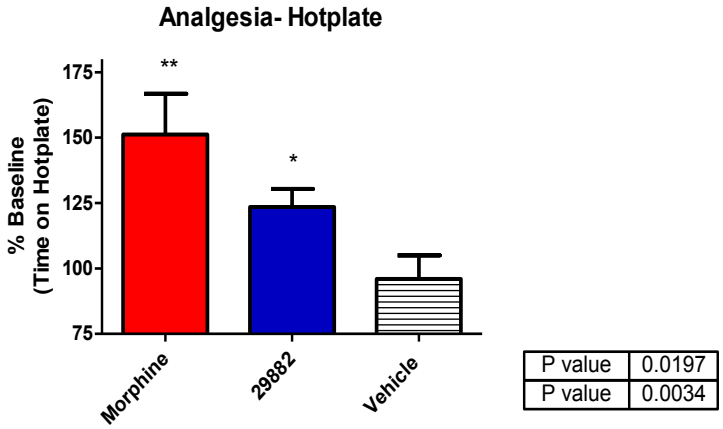


Figure 2.5
Mouse model for analgesia with Compound **12** and morphine.
The assay measure how long a mouse can stand on a hotplate without showing a pain response (hind paw lick, flutter, or splaying of hind pay toes). The percent increase in time spent on the hot plate compared to baseline (N=15) is shown. There is no statistical difference between Morphine (N=12) and Compound **12** (N=21).

2.4 Discussion

A promise of the structure-based approach is the discovery of novel molecules, unexplored by empirical methods and unrelated to natural and endogenous signaling molecules, but nevertheless able to complement the structure of the receptor. A hope is that the new chemotypes of such structure-derived ligands will confer on them new biological properties. Whereas structure-based design has made pivotal contributions to the development of over 15 marketed drugs and reagents[35, 36], including aliskiren, telapravir, nelfinavir, zanamivir, captopril and dorzolamide, rarely have *in vivo* active molecules with new properties emerged as rapidly as in this study. Compound **12** is dissimilar to other opioid ligands (ECFP4-based Tc value of 0.35 to its nearest molecule in ChEMBL16[29]), and is perhaps the most G-protein-biased of any extant μ opioid agonist. It has substantial specificity over other opioid sub-types and confers analgesia in a mouse pain model via its μ opioid activity. Compound **12** emerged from one round of computational optimization of molecules discovered from a docking screen. An opportunity to interact with the ordered waters—thought to be crucial to the recognition of the classic opioid drugs—was unmet by **12** based on its docked complex; addition of a single hydroxyl on the phenyl ring has an opportunity to greatly improved affinity due to predicted interactions with the ordered waters. This polar interaction exists with β-FNA and other morphanin ligands and is predicted from the docking poses.

The new activities of **12** are likely conferred by the new interactions that they make with the receptor. From the standpoint of specificity, the extension of both molecules into the region of the less conserved region of orthosteric site of the MOR, where the thiophene group

of the agonists bind in areas defined residues Lys233 and Tyr148, likely contributes to their specificity relative to the KOR, DOR and NOR receptors; it is in this region that the otherwise highly similar orthosteric sites of the four subtypes are distinct. The juxtaposition of an urea group proximal to the aminergic group in these molecules is a less common feature to opioid active molecules, and may contribute to the ligand efficiency of these molecules.

If structure-based discovery succeeded in illuminating novel scaffolds and chemotypes, and led to facile optimization to compounds **12**, it must be admitted that some of our results seem fortuitous. Arrestin vs. G-protein signaling likely reflects the stabilization of conformational sub-states over 30 Å from the orthosteric site and, except for an insistence on novel chemotypes, was not explicitly designed here. Indeed, even the seemingly more attainable goal of sub-type specific design, given the determination of the structures of all other major opioid receptors, was only specified in broad brush-stroke, by seeking analogs that extended into the less conserved part of the μ opioid site. Here, too, we largely relied on chemical novelty to confer new biological properties. Opportunities for more specific design of such remain open in this field. Finally, the observation that new molecules acted as agonists, despite docking to the inactive state of the receptor, flies in the teeth of most of our previous experience. [10, 13, 37] In our earlier studies, the function of docking-derived ligands exhibited high fidelity to the state of the receptor against which they were screened[9, 10, 13, 37]. Though the discovery of agonists here is consistent with an earlier docking screen against the KOR[21], which also found agonists, the structural bases of their agonism remains uncertain, except in broad stroke. A differentiating feature between agonists and antagonists is extension of bulky sub-structures below the basic nitrogen, into the region of the receptor defined by residues Tyr626,

Trp593 and Gly625. For example, the cyclopropyl of β-funaltrexamine, in the ”address” part of the ligand, is buried deep in the receptor of the structure, interacting with Tyr626 and Trp593, while morphine and other agonists lack such groups. Correspondingly, in their docked poses the new compounds do not extend into this region. This hypothesis remains to be tested.

These caveats should not obscure the main observations of this paper. A structure-based approach led rapidly to potent, sub-type specific and strongly G-protein biased ligand for the μ opioid receptor. Compound **12** conferred analgesia in a mouse pain model via specific action at the μ opioid receptor, without the potential side-effect liabilities of the classic opioid analgesics. These properties support their use as specific molecular probes for G-protein signaling at the μ-opioid receptor, and as lead molecules for new analgesics without the dose-limiting side-effects of the opioid drugs; compound **12** is freely available from the authors for research purposes. More broadly, this study supports the pragmatism of a structure-based approach, certainly against GPCRs. If the long promise of the structure-based program has only been fitfully realized, the *in vitro* results of multiple campaigns against GPCRs[9-13, 37, 38] and the *in vivo* results reported here, portend a general approach to the problem of novel tool and lead discovery for this pharmacologically central family of receptors.

2.5 Methods

I. Molecular docking

The μ-opioid receptor structure was used as input for receptor preparation with DOCK Blaster (http://blaster.docking.org) [26]. Forty-five matching spheres were used based on a truncated version of the crystalized ligand. The covalent bond and linker region of the morphanin antagonist was removed for sphere generation. The ligand sampling parameters were set with bin size, bin size overlap and distance tolerances of 0.4Å, 0.1 Å and 1.5 Å, respectively, for both the matching spheres and the docked molecules. Ligand poses were scored by summing the receptor-ligand electrostatics and van der Waals interaction energy, corrected for ligand desolvation. Receptor atom partial chargers were used from the united atom AMBER force field except for Lys233 and Try626, where the dipole moment was increased as previously described. The ZINC[23, 24] lead-like set of commercially available molecules were docked into the receptor using DOCK3.6 (http://dock.compbio.ucsf.edu/). Over 3 million compounds were docked and manually selected for experimental testing.

II. Analog selection

The top three hits from the primary screen, Compounds **4**, **5** and **7** were used to find analogs in ZINC with a similarity of greater than 0.7. Additionally, substructure searches were performed using the scaffolds in each of the three compounds. The searches yielded 500 purchasable compounds that were then docked as in primary screen. Analogs were then manually visualized for interactions and selected for experimental testing.

III. Radioligand competition binding assays

Standard techniques were used[39] at the NIMH Psychoactive Drug Screening Program.

IV. β-Arrestin recruitment Tango assay

Recruitment of β-arrestin to agonist-stimulated MOR receptors was performed using a previously described "Tango"-type assay[40].

V. Mouse analgesia model assay

This assay measures how long a mouse can stand on a 56°C hotplate without showing a pain response (hind paw lick, flutter, or splaying of hind pay toes). The baseline measurement is conducted the same day as drug or vehicle treatment, but several hours before hand. Animals were treated with 3mg/kg of Compound **12** or 3mg/kg of morphine and data was collected 20 minutes after subcutaneous injection.

2.6 References

1. Buniatian, A. A., Mizikov, V. M., Lovtsevich, N. V., Stamov, V. I. & Flerov, E. V. Esterase-metabolized remifentanil hydrochloride opioid (Ultiva-TM), a new step towards the solution of the problem of regulation of analgetic components of general anesthesia. *Anesteziol. Reanimatol.* **(2)**, 4-7 (2001).

2. Bohn, L. M. *et al*. Enhanced Morphine Analgesia in Mice Lacking β-Arrestin 2. *Science* **286**, 2495-2498 (1999).

3. DeWire, S. M. *et al*. A G Protein-Biased Ligand at the μ-Opioid Receptor Is Potently Analgesic with Reduced Gastrointestinal and Respiratory Dysfunction Compared with Morphine. *Journal of Pharmacology and Experimental Therapeutics* **344**, 708-717 (2013).

4. Manglik, A. *et al*. Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature* **485**, 321-326 (2012).

5. Wu, H. *et al*. Structure of the human κ-opioid receptor in complex with JDTic. *Nature* **485**, 327-332 (2012).

6. Granier, S. *et al*. Structure of the δ-opioid receptor bound to naltrindole. *Nature* **485**, 400-404 (2012).

7. Thompson, A. A. *et al*. Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature* **485**, 395-399 (2012).

8. Andrews, S. P. & Tehan, B. Stabilised G protein-coupled receptors in structure-based drug design: a case study with adenosine A2A receptor. *MedChemComm* **4**, 52-67 (2013).

9. Carlsson, J. *et al*. Structure-Based Discovery of A2A Adenosine Receptor Ligands. *Journal of Medicinal Chemistry* **53**, 3748-3755 (2010).

10. Carlsson, J. *et al*. Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nat Chem Biol* **7**, 769-778 (2011).

11. de Graaf, C. *et al*. Crystal Structure-Based Virtual Screening for Fragment-like Ligands of the Human Histamine H1 Receptor. *Journal of Medicinal Chemistry* **54**, 8195-8206 (2011).

12. Katritch, V. *et al*. Structure-Based Discovery of Novel Chemotypes for Adenosine A2A Receptor Antagonists. *Journal of Medicinal Chemistry* **53**, 1799-1809 (2010).

13. Kolb, P. *et al*. Structure-based discovery of β2-adrenergic receptor ligands. *Proceedings of the National Academy of Sciences* **106**, 6843-6848 (2009).

14. Kruse, A. C. *et al*. Muscarinic Receptors as Model Targets and Antitargets for Structure-Based Ligand Discovery. *Molecular Pharmacology* **84**, 528-540 (2013).

15. Langmead, C. J. *et al*. Identification of Novel Adenosine A2A Receptor Antagonists by Virtual Screening. *Journal of Medicinal Chemistry* **55**, 1904-1909 (2012).

16. Mysinger, M. M. *et al*. Structure-based ligand discovery for the protein–protein interface of chemokine receptor CXCR4. *Proceedings of the National Academy of Sciences* **109**, 5517-5522 (2012).

17. Tondi, D., Morandi, F., Bonnet, R., Costi, M. P. & Shoichet, B. K. Structure-Based Optimization of a Non-β-lactam Lead Results in Inhibitors That Do Not Up-Regulate β-Lactamase Expression in Cell Culture. *Journal of the American Chemical Society* **127**, 4632-4639 (2005).

18. Powers, R. A., Morandi, F. & Shoichet, B. K. Structure-Based Discovery of a Novel, Noncovalent Inhibitor of AmpC β-Lactamase. *Structure* **10**, 1013-1023 (2002).

19. Kalani, M. Y. S. *et al*. The predicted 3D structure of the human D2 dopamine receptor and the binding site and binding affinities for agonists and antagonists. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 3815-3820 (2004).

20. Gomes, I. *et al*. Identification of a μ-δ opioid receptor heteromer-biased agonist with antinociceptive activity. *Proceedings of the National Academy of Sciences* **110**, 12072-12077 (2013).

21. Negri, A. *et al*. Discovery of a Novel Selective Kappa-Opioid Receptor Agonist Using Crystal Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling* **53**, 521-526 (2013).

22. Oprea, T. Virtual Screening in Lead Discovery: A Viewpoint. *Molecules* **7**, 51-62 (2002).

23. Irwin, J. J. & Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chemical Information and Modeling* **45**, 177-182 (2004).

24. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* **52**, 1757-1768 (2012).

25. Lorber, D. M. & Shoichet, B. K. Hierarchical docking of databases of multiple ligand conformations. *Current topics in medicinal chemistry* **5**, 739-749 (2005).

26. Irwin, J. J. *et al*. Automated Docking Screens: A Feasibility Study. *Journal of Medicinal Chemistry* **52**, 5712-5720 (2009).

27. Mysinger, M. M. & Shoichet, B. K. Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *Journal of Chemical Information and Modeling* **50**, 1561-1573 (2010).

28. Hermann, J. C. *et al*. Metal Impurities Cause False Positives in High-Throughput Screening Campaigns. *ACS Medicinal Chemistry Letters* **4**, 197-200 (2012).

29. Gaulton, A. *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100-D1107 (2012).

30. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742-754 (2010).

31. Rogers, D. J. & Tanimoto, T. T. A computer program for classifying plants. *Science* **132**, 1115-1118 (1960).

32. Muchmore, S. W. *et al*. Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping. *Journal of Chemical Information and Modeling* **48**, 941-948 (2008).

33. Stahl, M. & Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *Journal of Medicinal Chemistry* **44**, 1035-1042 (2001).

34. Pettersen, E. F. *et al*. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605-1612 (2004).

35. Congreve, M., Murray, C. W. & Blundell, T. L. Keynote review: Structural biology and drug discovery. *Drug Discovery Today* **10**, 895-907 (2005).

36. Shoichet, B. K. & Kobilka, B. K. Structure-based drug screening for G-protein-coupled receptors. *Trends in Pharmacological Sciences* **33**, 268-272 (2012).

37. Weiss, D. R. *et al*. Conformation Guides Molecular Efficacy in Docking Screens of Activated β-2 Adrenergic G Protein Coupled Receptor. *ACS Chemical Biology* **8**, 1018-1026 (2013).

38. Kolb, P. & Klebe, G. The Golden Age of GPCR Structural Biology: Any Impact on Drug Design? *Angewandte Chemie International Edition* **50**, 11573-11575 (2011).

39. Jensen, N. H. *et al*. N-desalkylquetiapine, a potent norepinephrine reuptake inhibitor and partial 5-HT1A agonist, as a putative mediator of quetiapine's antidepressant activity. *Neuropsychopharmacology* **33**, 2303-2312 (2007).

40. Chen, X. *et al*. Structure–Functional Selectivity Relationship Studies of β-Arrestin-Biased Dopamine D2 Receptor Agonists. *J. Med. Chem.* **55**, 7141-7153 (2012).

Gloss to Chapter 3

Not everything in graduate school becomes a raging success. It's what I was warned from before I started graduate school but the work in this chapter was when I faced tough setbacks in this thesis. The project encompassed in chapter 3 was a major aspect of the first "crazy goals" that Brian and I laid out for my thesis project. The major idea was to combine SEA[1] and DOCK[2-4] together. In my lab meeting from June 1, 2012, I coined the title of my presentation as "Blind date at the PIER: when SEA and DOCK meet" and decided to call this project PIER.

Although there were not many major successes with this project, the core ideas and principles are still worth further investigation and tuning of methods. We first used the database of useful decoys – extended (DUD-E) [5] set of targets to dock the combination of all known ligands from ChEMBL database[6] for these targets as a proof of concept set. The scope of the project would be expanded to, in theory, the entire proteome as long as there was a "dockable" x-ray crystal structure.

Before going full force into the project, we first wanted to calculate the concordance between using docking hit lists as surrogates for ligands and using actual annotated ligands to relate proteins with SEA. At first we were optimistic about the project since the concordance was relatively high with only one pair of proteins that SEA on annotated ligands deemed it significant but there was no combination of docking hit lists for the same pair that was significant. What was exciting for us were the pairs of proteins that were brought together by the docking hit lists but were not significant when annotated ligands were used. These were

opportunities that we could exploit with the thinking that docking enriched different and perhaps new chemotypes of ligands not previously reported in the literature for the targets. We aimed to interrogate a few of these pairs of proteins to find molecules to test against new targets.

Despite our best efforts, only one ligand, Tenidap, was active out of the 10 pairs of targets we tested. Tenidap is a known cyclooxygenase 2 (COX2) inhibitor that we predicted for binding to peroxisome proliferator-activated receptor γ (PPARG) with a SEA E-value of 2.1×10^{-38}. Tenidap showed dose response activity against PPARG with IC_{50} of 15 µM, which we considered a success. Although we only show one example, we expect there to be more ligands that can be predicted from docking since the broader idea is that different binding sites can bind the same ligand and that the similarity between docked molecules can pick up the similarities between the binding sites using SEA. Additional implications and uses of this method are discussed.

I. References

1. Keiser, M. J. et al. Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. 25, 197-206 (2007).

2. Lorber, D. M. & Shoichet, B. K. Hierarchical docking of databases of multiple ligand conformations. Current topics in medicinal chemistry 5, 739 (2005).

3. Irwin, J. J. et al. Automated docking screens: a feasibility study. J. Med. Chem. 52, 5712-5720 (2009).

4. Mysinger, M. M. & Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. Journal of chemical information and modeling 50, 1561-1573 (2010).

5. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J. Med. Chem. 55, 6582-6594 (2012).

6. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100-D1107 (2012).

Chapter 3: “Blind date at the PIER: when SEA and DOCK meet”

3.1 Introduction

Structure based virtual screening is now a fairly established method in terms of finding small molecule ligands for protein targets with either solved or modeled structures.[1-4] Even for proteins structures in which no ligand was co-crystalized in the binding site (APO state) or even with homology models[5, 6], molecules have still been docked and successfully tested in prospective experiments.[7-9] More and more recently approved drugs have attributed at least some part of the drug discovery effort of the project to structure based virtual screening or modeling of ligands to the active site.[2, 10] It has provided a platform to discover new ligands and generate ideas on what analogs to synthesize to increase potency of a ligand to the protein target.

Certain classes of protein targets, such as kinases[11], proteases[12, 13] and other enzymes, lend themselves well to structure based ligand discovery since structures are more readily available to crystallography. There is still a lot of discussion about the overall utility of structure for drug discovery, however, since there has not been a dramatic increase of drug approvals per year since structure based drug discovery has been used.[2, 10] Although structure based drug discovery has not lived up to what it had originally promised, in terms for drug discovery, there have been still many interesting and novel molecules for targets that have been discovered using this technique.[1, 6, 14, 15] For G protein-coupled receptors (GPCRs), the field of structure based ligand discovery is still in its early stages since the first pharmaceutically relevant structure was solved only in 2007.[16, 17] Some of the reasons structure based drug discovery has not been as successful are perhaps independent of virtual screening or protein-ligand

interactions, such as off target side effects and efficacy issues in regards to the biology of the target.[1, 2, 10] Nevertheless, virtual screening methods can and are being refined and intensely researched so as to better model the biophysics of protein-ligand interactions.[18-20]

Evaluating docking hit-lists have in itself become a field of intense study with many different criteria and parameters to measure docking "success".[4, 21, 22] The introduction of DUD[19] and subsequently DUD-E[18] has helped confine the field to a standardized benchmark in which docking performance can be evaluated. One of the common, standard metrics to evaluate docking performance suggested from the DUD papers is to calculate the adjusted log area under the curve (log AUC) of enrichment of known ligands over matched decoys. The higher the adjusted log AUC, the more known ligands are ranked higher than their matched decoys. Retrospectively evaluating docking hit lists provides a useful gauge on how well docking can separate ligands from decoys and can correlate to prospective docking performance. For example, with recent GPCR screens, enrichment has tended to be fairly high and prospective docking has led to hit rates of over 20% while discovering ligands with nanomolar binding constants.[6, 8]

What sprang up next after structure-based virtual screening techniques like docking, was a focus less on the protein side, but on the ligand side. These methods are generally classified as ligand-based virtual screening (LBVS).[23-28] An advantage of LBVS is the lack of requiring a protein structure, whether it be solved or modeled. The protein itself is defined by the set of ligands that bound to the protein and not its amino acid sequence, structural fold or function. A major disadvantage is that LBVS methods all require known ligands to model or

learn after whereas docking can predict ligands for a protein that has no known ligands. LBVS generally falls under two categories: pharmacophore or fingerprint based. The Similarity Ensemble Approach (SEA)[25] is a fingerprint based method that defines a protein by the 2D fingerprints of the ligands known to.

The initial successes with SEA were with predicting new targets of drugs.[25, 29] In essence, SEA is used as a screening tool to predict a new protein target for drugs. Subsequently, SEA was used for target vs. target comparisons in relating GPCRs to each other and GPCRs to non-GPCRs.[30] SEA was used to relate proteins together in an orthogonal way to standard bioinformatics methods of sequence, structure or function, by using the ligands themselves. Despite its success at relating proteins from different families together by new shared ligands, it is still a knowledge-based approach and required many known ligands for each protein to be able to link unrelated proteins together.

With an idea that goes as far back as 2004 in the Shoichet lab, docking hit lists themselves can potentially contain enough information about the protein, in terms of its function, to be useful in relating to the docking hit lists of a different, unrelated protein. Naturally, SEA would be the engine that compares the docking hit lists to find enough similarity between the top scoring molecules to statistically relate two proteins together. In one of my first lab meeting presentations on this project, I gave it the title “Blind date at the PIER: When DOCK and SEA meet”.

The hypothesis for the project is that docking hit lists contain enough information content to describe a protein’s pharmacology and ligand chemical space despite the

inaccuracies of dock. The description of a protein's pharmacology, although not complete, is informative when taken in aggregate with a set of reasonable biogenic molecules, such as the ChEMBL database.[31] Comparing docking hit-lists would expand the range of chemotypes associated with a target, and allow us to consider targets that are "dockable" but for which ligands are unknown. Using SEA on docking hit lists would free us from our current shackles of known ligands/chemotypes on known targets. If we could rely on docking hit lists, we would vastly increase the number of targets that could be substrates for SEA. Docking's high false-positive rate is definitely a concern but even when docking hits do not actually bind to the protein, they are rarely ridiculous. In a highly enriching docking hit list, the same chemotypes are repeated among the top ranking molecules when there are enough similar analogs making similar interactions with the protein. As a result of DOCK, two related targets, in that they bind similar molecules, can have hit lists composed of mostly non-binding molecules but might still resemble each other, according to SEA.

Applying SEA to docking hit lists has the potential to predict a lot of novel protein links that might not have been related with just known ligands and SEA. I applied this method to a proof-of-concept dataset of sorts, the DUD-E set of targets and their respective ligands.[18] The DUD-E set contains over 100 protein targets comprised from over 7 different protein families with at least 40 known ligands for each target. This rich dataset has been docked, evaluated and well-studied both in terms of docking and biologically since so many ligands have been annotated for each protein.[20] Although the biological function of each of the proteins are generally well established, the links and relationships between the proteins have not been as studied. I aimed to dock the ChEMBL database of ligands to each of these targets and use their

hit lists as input into SEA to relate the proteins pharmacologically together. The protein links would strictly be based on docking hits and not annotated ligands.

As a first pass, I compare the SEA E-values using the annotated ligands for each protein pair to the SEA E-values using the docking hit lists for each pair. The results are surprising: almost all the pairs of protein that had significant SEA E-values using annotated ligands also had significant SEA E-values using the docking hit lists. Conversely, many of the pairs that do not have significant SEA E-values using annotated ligands also do not have significant SEA E-values using docking hit lists. As expected, there are many pairs of proteins that SEA related together using docking hit lists that do not have significant E-values using annotated ligands. These SEA predictions provide opportunities for us to pursue finding shared ligands for targets that, to the best of our knowledge, do not have any shared ligands. The shared ligands will be predicted from docked ligands and not expected by just SEA on annotated ligands alone. We evaluate several of the predictions and experimentally test 17 compounds for 9 targets to identify new links between two DUD-E proteins. With the few number of protein links tested, a lot of potential for discovering new protein links in the DUD-E set remains.

3.2 Results

I. Running SEA on docking hit lists of DUD-E set

The ligand sets of the DUD-E[18] targets from ChEMBL[31] annotations were compiled together, over 200,000 ligands, and docked to each DUD-E target. Targets and ligands were prepared for docking as in DUD-E paper. Docked ligands were extracted and filtered to generate SEA sets for each protein target. The criteria for filtering were by top N, molecular weight and ligand efficiency. In total, 48 sets of molecules were compiled for each protein target and SEA was run pairwise for 101 of the DUD-E targets resulting in 489,648 SEA set calculations, including self-self.

II. Evaluating SEA results on docking hit lists of DUD-E set

Over half of the SEA set calculations had an E-value less than 1×10^{-5}. For each protein target, the self-self comparison always had significant E-values for each of the 48 combinations of docking hit lists, suggesting the self-consistency and similarity of the molecules that are enriched by docking. With SEA on annotated ligand sets alone, 312 of the possible 5,151 pairs had significant E-values lower than 1×10^{-5}. Using the docking hit lists, SEA was able to recapture 311 of the 312 significant pair associations with at least one of the 48 combinations per pair. In addition to the 311 significant pair associations, 591 new predictions were significant from the docking hit lists that would not have been predicted by using the annotated ligand sets alone. (**Figure 3.1**) Assuming the relationships picked up by SEA on annotated ligand sets are considered "True", the "sensitivity" for SEA on docking hit lists was 0.997 and the

"specificity" was 0.878. (**Table 3.1**) The new predictions from docking hit lists were considered significant if at least half of the 48 combinations had E-values lower than 1 x 10^{-5} and a known shared ligand must not have been known already in ChEMBL.

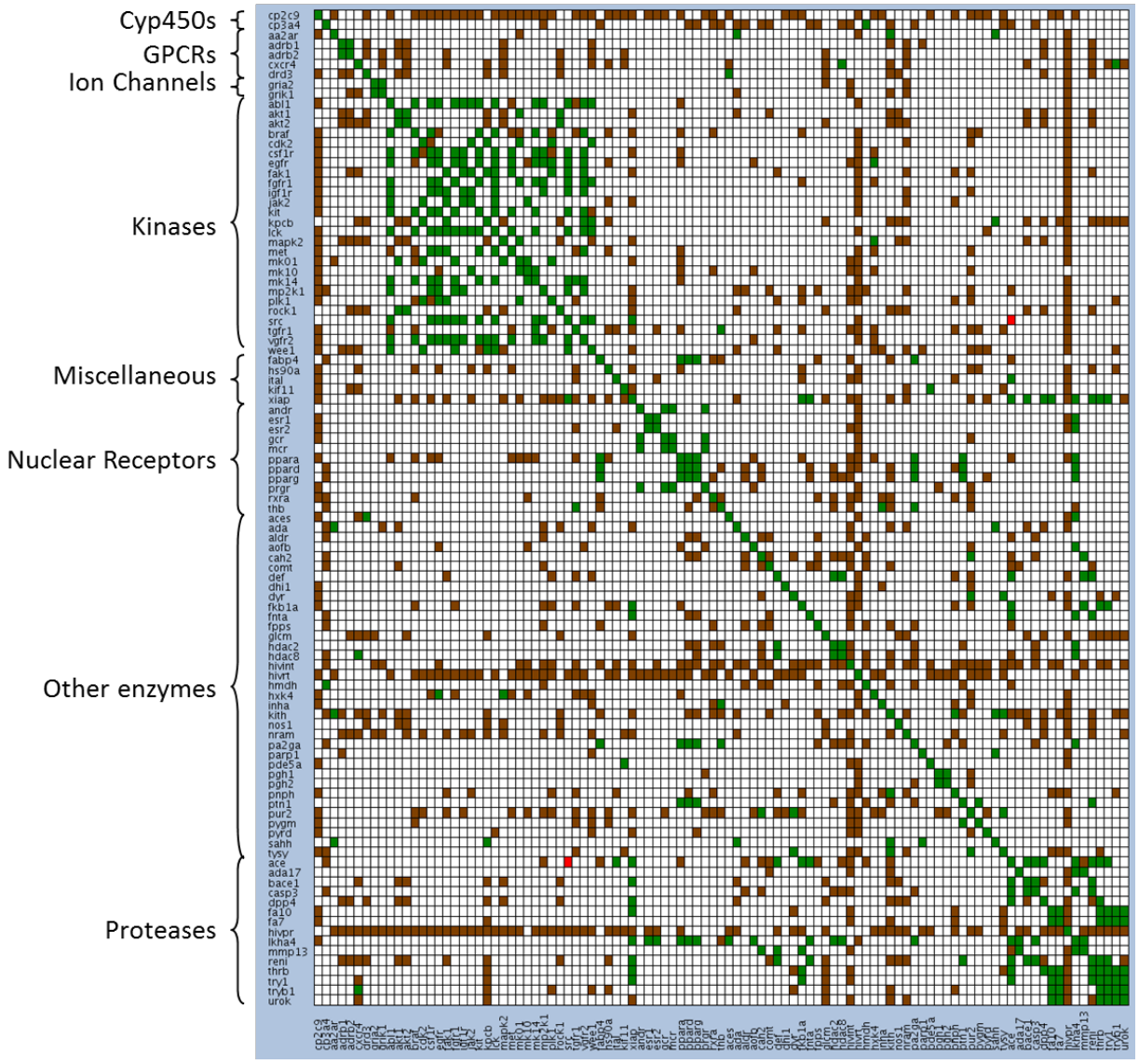


Figure 3.1
Heatmap view of protein pairs.
Significant by docking hit list and annotated ligands (green), only by annotated ligands (red), only by docking hit lists (brown), or insignificant by both (white).

Number of protein pairs		Significant from annotated ligands	
		Yes	No
Significant by DOCK	Yes	311	591
	No	1	4248

Table 3.1 "Truth" table for concordance between relating proteins using docking hit lists and annotated ligands with SEA.
Nearly all the significant pairs using annotated ligands were recaptured by docking hit lists. There were still 591 pairs of proteins that docking hit lists predict there to be a potential shared ligand. These were the pairs that we investigated prospectively for experimental testing.

III. Testing of predicted links from docking hit lists

Of the 591 predictions, we evaluated and ranked the predictions by the number of known ligands for each target appearing in the docking hit lists and the number of known ligands for one target that also was ranked highly in the docking hit list of the other paired target. We also only considered pairs of proteins that belong to different protein families (e.g. a GPCR and a nuclear hormone receptor). The hit lists for the pairs were then manually inspected for pose and surface complementarity with the respective binding sites. Ten pairs of protein targets were selected for prospective testing with one to four compounds per target pair.

One pair of interest was the link between Cyclooxygenase 2 (COX2) and peroxisome proliferator-activated receptor γ (PPARG) with a SEA E-value of 2.1×10^{-38}. In the top hit list of PPARG there are 115 known COX2 ligands. Conversely, 11 known PPARG ligands are in the top hit list of COX2. We purchased 4 COX2 ligands for testing at PPARG for binding. One compound, Tenidap, showed dose response activity against PPARG with IC_{50} of 15 µM. (**Figure 3.2**) The docking poses for Tenidap are reasonable where nearly all polar atoms are coordinated by polar residues in PPARG. (**Figure 3.3**) The other three compounds showed less than 50% inhibition at 20 µM.

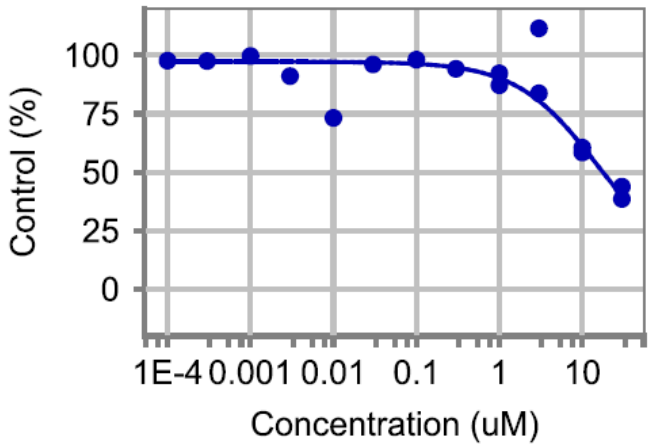


Figure 3.2 Dose-response curve for Tenidap against PPARG in antagonist mode.

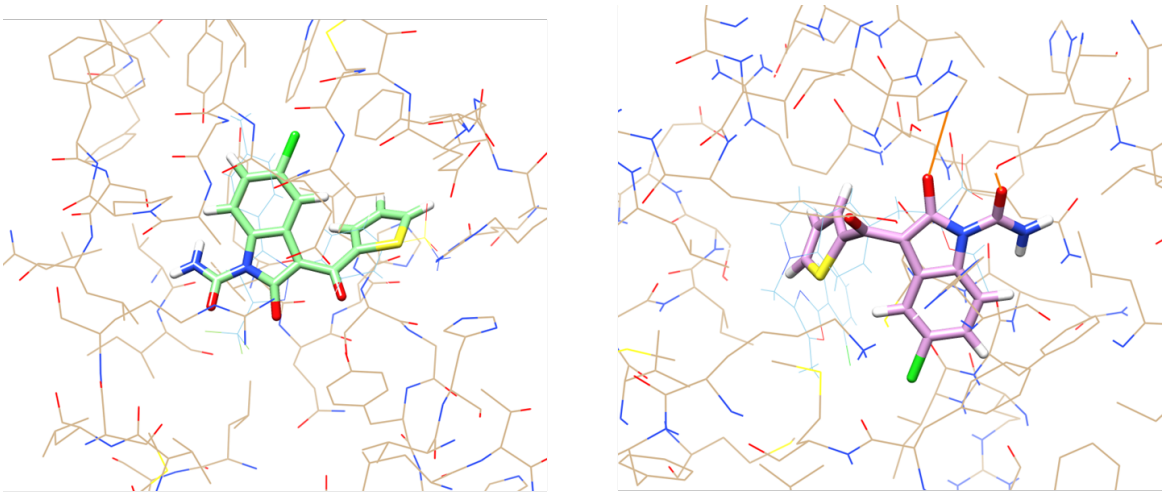


Figure 3.3 Docking poses of Tenidap at previously known target (COX2) and newly determined target (PPARG)
Tenidap is predicted to have its carbonyl oxygens coordinated by histidine and tyrosine whereas in COX2, there is no predicted polar interactions from the docking poses.

The other predictions we tested did not show enough activity to pursue dose response at reasonable concentrations. We still have a few more promising predictions yet to be tested such as the link between Adenosine A2A and Progesterone receptors and Farnesyl pyrophaspate synthase with Angiotensin converting enzyme and Aldose reductase.

3.3 Discussion

Docking and SEA have thus far been independent from each other. We aimed to combine the two methods by applying SEA[25] to docking hit lists. ChEMBL[31] ligands of DUD-E[18] targets were docked and the hit lists were divided into different sets based on molecular weight, ligand efficiency and top N. The individual hit list sets were then compared using SEA for significance in the similarity between the compounds of the hit lists. The results using docking hit lists were compared those using annotated ligands. Surprisingly, there was very high concordance between the pairs that were considered significant using annotated ligands and docking hit lists. There was only one pair of targets that their annotated ligands were considered significantly similar by SEA but not similar by all 48 docking hit lists between the same targets.

One important factor that allowed us to achieve such high concordance was the splitting of the docking hit lists into different sets. There was not a universal hit list cutoff in terms of number of top scoring molecules or molecular weight. Although docking has been most successfully used by our lab with lead-like compounds rather than larger, peptidic or drug like molecules, some DUD-E targets have larger cavities and binding sites that accommodate larger molecules than others. For example, C-X-C chemokine receptor type 4's ligands have an average molecular weight over 400 for known ligands. Conversely, Adenosine deaminase has a smaller cavity and generally binds small ring systems. Due to the diversity of binding sites, it became necessary to use different molecular weight cutoffs for the hit lists. Having too many molecules that do not dock or score well in the hit list sets simply because of their size adds

more noise to the data for SEA comparisons since all compounds are equally weighted in the SEA set.

Another factor in consideration was the size of the binding site. Depending on the size, a different number of molecules may be reasonably docked. The ChEMBL ligands that were docked are not equally distributed in chemical space so there exists a library bias for certain targets over others. For certain targets that generally have lower enrichment for docking, fewer molecules may dock and score well so including poor scoring molecules in docking hit lists can also influence SEA similarity predictions. On the other hand, some targets have very good docking enrichment and having more compounds in the docking hit list can better describe the available chemical space that the target can accommodate. For these targets, larger docking hit list sets tended to better represent the ligand chemical space.

The range of molecular weight cutoffs were under 350, 400, 450, 500, 550 and 600 along with ligand efficiency and the top N molecules range was 100, 250, 500, 1000, 2500 and 5000. For each target's docking hit list, we were able to divide it up into 48 different sets using the ranges. Although more combinations or ways of pruning the hit lists for each target are possible, we were comfortable with the 48 combinations we used that covered a diverse enough range for size and number of molecules. Ultimately, we were driven by the concordance with the SEA E-values using annotated ligands. We were, however, left with over 500 predicted target-target links from the docking hit lists that were not predicted by the annotated ligands.

With the predicted links, we were cautiously hopeful since they presented opportunities in finding new, shared ligands for targets that did not have shared ligands before but knew there would be a lot of noise in the results. Not only did we have to consider the errors typically associated with docking,[3, 18, 32-34] but we also had to take into consideration the deficiencies of SEA. Many of the predicted target pairs were interesting since they did not share a known ligand, to the best of our knowledge, and the individual targets belonged to different protein families and had different folds and structure. SEA was still able to relate these disparate proteins with docking hit lists rather than known ligands. What was encouraging was the number of target pairs that had their own known ligands in the hit list and also in the hit list of opposing target. We began exploring these pairs since there were ligands known to bind to one of the two proteins already and only one experiment would be needed for the other target.

A downside to docking annotated ligand databases, such as ChEMBL, is the fact that most of the compounds are not commercially available or require custom synthesis. Although the ligands are useful for docking and relating targets together with SEA, prospective testing was more difficult due to the large barrier of sourcing chemical matter for experiments. As a result, several top docking molecule predictions were not testable. Nonetheless, there were still many reasonable pairs of targets worth investigating and testing.

Unfortunately, only one of the tested hypothesis was active with Tenidap weakly antagonizing PPARG with IC_{50} = 10 µM. The docking pose suggests that Tenidap likely maintains a similar confirmation when interacting with PPARG and COX2 but the local chemical environments of the two binding sites are very different.

A broader idea that the one example suggests is that different binding sites can bind the same ligand and that the similarity between docked molecules can pick up the similarities between the binding sites using SEA. We,[25, 29, 30, 35] and others,[28, 36] have previously found many new targets for ligands, whether they are wanted or unwanted effects. Ligand selectivity, in terms of just binding to one protein, has been difficult to achieve for commercial drugs. What we aimed to do in this project was to take advantage of the fact that ligands do not just bind to one protein and use that idea to relate proteins together by the ligands themselves.[30, 37] One way to describe the protein is by the ligands that bind it. What docking can do is take it to a more descriptive level and limit it to just one binding site and the ligands that may best bind to that one site. Docking a larger, broader set of molecules into a protein binding site can help discover new, previously uncharacterized ligands for the target. Often times, docking explores new chemical space not associated with a particular target since docking campaigns can test a much larger, over 3 million compounds,[3, 38] and diverse chemical space even more so than large pharmaceutical high-throughput screens that screen roughly one million molecules.[39] Docking has also previously been shown to help reveal the function of proteins by identifying substrates for enzymes with unknown function.[40, 41]

Combining the two approaches with docking mentioned could have broader applications beyond relating proteins that already have many known ligands and in turn, relate a protein without known ligands or function to a protein with many defined ligands and a defined function. Docking to a protein with unknown function and using SEA to relate the hit list to other proteins would be a potential use case that can be generalized and applied on a large scale. With recent improvements to DOCK, in terms of speed and accuracy, high

throughput docking on a proteomic scale is becoming increasingly realistic but also could potentially generate too much data to process manually.[3, 42] SEA would be able to filter down which proteins with unknown function are more similar, in terms of ligand pharmacology, to proteins with known function in an automated and quick manner.

Although the future of this project is unclear, the hypotheses and methods are still highly encouraging. With just one example of an active ligand linked by docking hit lists, it remains to be seen if the method is useful or if it was just the wrong examples tested. With each pair of proteins, only a few ligands were tested. The results may have been different if a more standard 20 ligands are tested per docking campaign, but comes at a much greater cost. Either way, there are more methods that can be developed and it remains a difficult problem worth further exploration.

3.4 Methods

I. Molecular docking screens

The DUD-E set of proteins was downloaded from the DUD-E webpage (http://dude.docking.org) [18] with each protein prepared using DOCK Blaster (http://blaster.docking.org) [3] as described previously. Ligand poses were scored by summing the receptor-ligand electrostatics and van der Waals interaction energy, corrected for ligand desolvation. ChEMBL 12[31] ligands for each protein were filtered for ligands less than 600 Da, 20 or fewer rotatable bonds and better than 10 µM affinity (IC_{50}, EC_{50}, K_i, K_d, and log variants thereof). The entire set of all the ligands was docked into the receptor using DOCK3.6 (http://dock.compbio.ucsf.edu/).[33, 34, 42]

II. Generation of top hit-lists from DOCK for SEA

Each target's docking output was divided into 48 different sets using molecular weight cutoffs or ligand efficiency and top N molecules based on docking energy score. The range of molecular weight cutoffs were under 350, 400, 450, 500, 550 and 600 along with ligand efficiency and the top N molecules range was 100, 250, 500, 1000, 2500 and 5000. Molecular weights of ligands were extracted from the ZINC database (http://zinc.docking.org). [38] Ligand efficiency was calculated by taking the docking energy divided by the number of heavy atoms.

III. SEA calculations on top hit lists from DOCK and ChEMBL annotated ligands

SEA background was generated using the ChEMBL ligands that were docked. For each slice of a protein's docking hit list, it was compared with the same slice for each of the other proteins resulting in 489,648 SEA set calculations, including self-self. The ChEMBL ligands for each target were also compared using SEA pair-wise to every other target in DUD-E for a total of 5050 SEA set calculations, including self-self. SEA E-values were considered significant if the value was $< 1 \times 10^{-5}$.

IV. Comparing SEA E-values from docking hit lists to annotated ChEMBL ligand sets

For each significant SEA E-value from annotated ChEMBL ligand sets, the 48 docking hit list SEA set comparisons for the same pair of targets was checked for significance. These pairs were considered matched to "true positives". Conversely, for each non-significant SEA E-value from annotated ChEMBL ligand sets, the 48 docking hit list SEA set comparisons for the same pair of targets were further evaluated. If fewer than 24 of the 48 docking hit list SEA set comparisons for each target pair was considered non-significant, the target pair was considered matched to "true negatives". If greater than 24 of the 48 docking hit list SEA set comparisions for each target pair was considered significant, the pair was considered a "false positive" but manually investigated by viewing top docked molecules in Chimera. Only the molecules that were known to bind to one of the targets in the pair were considered to minimize the number of docking poses to consider and experimentally test.

3.5 References

1. Cheng, T., Li, Q., Zhou, Z., Wang, Y. & Bryant, S. H. Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS journal* **14**, 133-141 (2012).

2. Congreve, M., Murray, C. W. & Blundell, T. L. Keynote review: Structural biology and drug discovery. *Drug Discov. Today* **10**, 895-907 (2005).

3. Irwin, J. J. *et al*. Automated docking screens: a feasibility study. *J. Med. Chem.* **52**, 5712-5720 (2009).

4. Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **7**, 1047-1055 (2002).

5. Fan, H. *et al*. Molecular docking screens using comparative models of proteins. *Journal of chemical information and modeling* **49**, 2512-2527 (2009).

6. Carlsson, J. *et al*. Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nature chemical biology* **7**, 769-778 (2011).

7. Fan, H., Irwin, J. J. & Sali, A. Virtual ligand screening against comparative protein structure models. *Computational Drug Discovery and Design* 105-126 (Springer, 2012).

8. Kolb, P. *et al*. Structure-based discovery of β2-adrenergic receptor ligands. *Proceedings of the National Academy of Sciences* **106**, 6843-6848 (2009).

9. Schlessinger, A. *et al*. Structure-based discovery of prescription drugs that interact with the norepinephrine transporter, NET. *Proceedings of the National Academy of Sciences* **108**, 15810-15815 (2011).

10. Shoichet, B. K. & Kobilka, B. K. Structure-based drug screening for G-protein-coupled receptors. *Trends Pharmacol. Sci.* **33**, 268-272 (2012).

11. Cui, J. J. *et al*. Structure based drug design of crizotinib (PF-02341066), a potent and selective dual inhibitor of mesenchymal–epithelial transition factor (c-MET) kinase and anaplastic lymphoma kinase (ALK). *J. Med. Chem.* **54**, 6342-6363 (2011).

12. Lin, C., Kwong, A. & Perni, R. Discovery and development of VX-950, a novel, covalent, and reversible inhibitor of hepatitis C virus NS3. 4A serine protease. *Infectious disorders-drug targets* **6**, 3-16 (2006).

13. Rahuel, J. *et al*. Structure-based drug design: the discovery of novel nonpeptide orally active inhibitors of human renin. *Chem. Biol.* **7**, 493-504 (2000).

14. Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **8**, 365-370 (2004).

15. Powers, R. A., Morandi, F. & Shoichet, B. K. Structure-based discovery of a novel, noncovalent inhibitor of AmpC β-lactamase. *Structure* **10**, 1013-1023 (2002).

16. Rasmussen, S. G. *et al*. Crystal structure of the human β2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383-387 (2007).

17. Cherezov, V. *et al*. High-resolution crystal structure of an engineered human β2-adrenergic G protein–coupled receptor. *Science* **318**, 1258-1265 (2007).

18. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582-6594 (2012).

19. Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **49**, 6789-6801 (2006).

20. Novikov, F. N. *et al*. Lead Finder docking and virtual screening evaluation with Astex and DUD test sets. *J. Comput. Aided Mol. Des.* **26**, 725-735 (2012).

21. Wallach, I. & Lilien, R. Virtual decoy sets for molecular docking benchmarks. *Journal of chemical information and modeling* **51**, 196-202 (2011).

22. Spitzer, R. & Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput. Aided Mol. Des.* **26**, 687-699 (2012).

23. Hert, J. *et al*. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *Journal of chemical information and modeling* **46**, 462-470 (2006).

24. Jain, A. N. QMOD: physically meaningful QSAR. *J. Comput. Aided Mol. Des.* **24**, 865-878 (2010).

25. Keiser, M. J. *et al*. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197-206 (2007).

26. Koeppen, H., Kriegl, J., Lessel, U., Tautermann, C. S. & Wellenzohn, B. Ligand-Based Virtual Screening. *Virtual Screening: Principles, Challenges, and Practical Guidelines*, 61-85 (2011).

27. Wolber, G. & Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *Journal of chemical information and modeling* **45**, 160-169 (2005).

28. Yera, E. R., Cleves, A. E. & Jain, A. N. Chemical structural novelty: on-targets and off-targets. *J. Med. Chem.* **54**, 6771-6785 (2011).

29. Keiser, M. J. *et al*. Predicting new molecular targets for known drugs. *Nature* **462**, 175-181 (2009).

30. Lin, H., Sassano, M. F., Roth, B. L. & Shoichet, B. K. A pharmacological organization of G protein-coupled receptors. *Nature methods* **10**, 140-146 (2013).

31. Gaulton, A. *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100-D1107 (2012).

32. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery* **3**, 935-949 (2004).

33. Mysinger, M. M. & Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *Journal of chemical information and modeling* **50**, 1561-1573 (2010).

34. Lorber, D. M. & Shoichet, B. K. Hierarchical docking of databases of multiple ligand conformations. *Current topics in medicinal chemistry* **5**, 739 (2005).

35. Lounkine, E. *et al*. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361-367 (2012).

36. Varela, R., Walters, W. P., Goldman, B. B. & Jain, A. N. Iterative Refinement of a Binding Pocket Model: Active Computational Steering of Lead Optimization. *J. Med. Chem.* **55**, 8926-8942 (2012).

37. Knight, Z. A., Lin, H. & Shokat, K. M. Targeting the cancer kinome through polypharmacology. *Nature Reviews Cancer* **10**, 130-137 (2010).

38. Irwin, J. J. & Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling* **45**, 177-182 (2005).

39. Austin, C. P., Brady, L. S., Insel, T. R. & Collins, F. S. NIH molecular libraries initiative. *Science* **306**, 1138-1139 (2004).

40. Hermann, J. C. *et al*. Predicting substrates by docking high-energy intermediates to enzyme structures. *J. Am. Chem. Soc.* **128**, 15882-15891 (2006).

41. Hermann, J. C. *et al*. Structure-based activity prediction for an enzyme of unknown function. *Nature* **448**, 775-779 (2007).

42. Coleman, R. G., Carchia, M., Sterling, T., Irwin, J. J. & Shoichet, B. K. Ligand Pose and Orientational Sampling in Molecular Docking. *PloS one* **8**, e75992 (2013).

Gloss to Chapter 4

In the chapter that follows, the follow-up to the paper in chapter 1 is discussed. We previously showed how a pharmacological organization of G protein-coupled receptors (GPCRs) can be dramatically different than a sequence based organization[1]. Furthermore, GPCRs can be related to *non*-GPCRs also by the similarity of their ligand sets. We were challenged by an editor of *Nature Methods* to show biological relevancy and meaning to these new protein-protein links. We hypothesize that proteins undergo chemical pressures to not only react, but respond and evolve to new functions and regulation of signaling pathways because there are a limited number of endogenous signaling molecules[2]. These molecules are difficult to synthesize, degrade and control, demanding lots of proteins. They are not by themselves direct substrates for evolution and so once evolved, they become frozen (due to too many dependencies to change) and signaling/life begins to evolve around them. Thus, proteins that although are from vastly different families with different structure and sequence, can evolve to bind and function through the same molecules. Moreover, these proteins can also share similar function or phenotype despite being so different from a bioinformatics perspective.

In this chapter, we show a number of examples of proteins that have similar phenotypes or functions and have ligands that are similar by SEA[3] but not identical ligands. We test a number of these predictions by showing dose response activity for new ligands against the predicted target either through direct binding or functionally. In one example, we relate vesicular monoamine transporter 2 (VMAT2) to dopamine receptors (DRDs) due to the shared phenotypes of bradykinesia, hypoactivity, impaired coordination, abnormal dopaminergic

neurons and others. VMAT2 transports dopamine and is one of the main controllers of dopamine encapsulation and its extracellular release. DRDs, on the other hand, react to extracellular dopamine and transmit signals intracellularly upon dopamine binding. Here the endogenous interaction molecule is the same, namely dopamine, and thus provides an opportunity to tune the circuit with another ligand. Lobeline, a natural alkaloid, was previously known to interact with VMAT2 and reduce methamphetamine binding[4], thus reducing extracellular concentrations of dopamine so the reward system stimulated by dopamine through the DRDs will not be as activated. By the pharmacological and phenotypic link we calculated with the method presented, lobeline was chosen for testing and interestingly it also directly modulates dopamine receptors and acts as an antagonist.

An area of active research is natural language processing and text mining for relating medical, functional, and phenotypic keywords and terms into an ontology or relationship map. The annotations from OMIM[5] and other databases[6, 7] are often redundant or reworded slightly differently for the same indication, phenotype, function, etc… When we aimed to relate proteins by these annotations, we realized there was not as much overlap simply due to word semantics and wording of the annotation. SEA, by itself, over predicts far too many little networks of unrelated proteins and we needed something else to prioritize what we should examine first. An orthogonal way of judging whether two unrelated targets in fact share a function and wanted functional annotations to be that tool. We thus devised an approach to calculate annotation similarity with principles borrowed from SEA and chemoinformatics. Briefly, the targets for each phenotype are used as bits in a “fingerprint”. For two phenotypes, a similarity value is calculated by the overlap of shared bits, or targets in this case, over the

number of total bits with a Tanimoto coefficient (Tc) [8]. This idea was borrowed from how we compare ligands in SEA and enabled us to quickly filter down and relate phenotypes together.

Although unfinished, the ideas and examples presented here provide a framework to move forward with the project and set the basis for a paper to expand on the ideas from Chapter 1 and take it a step further by phenotypically relating disparate proteins with new shared ligands.

I. References

1. Lin, H., Sassano, M. F., Roth, B. L. & Shoichet, B. K. A pharmacological organization of G protein-coupled receptors. *Nature methods* **10**, 140-146 (2013).

2. Tomkins, G. M. The metabolic code. *Science* **189**, 760-763 (1975).

3. Keiser, M. J. *et al*. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197-206 (2007).

4. Eyerman, D. J. & Yamamoto, B. K. Lobeline attenuates methamphetamine-induced changes in vesicular monoamine transporter 2 immunoreactivity and monoamine depletions in the striatum. *J. Pharmacol. Exp. Ther.* **312**, 160-169 (2005).

5. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-D517 (2005).

6. McGary, K. L. *et al*. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences* **107**, 6544-6549 (2010).

7. Woods, J. O., Singh-Blom, U. M., Laurent, J. M., McGary, K. L. & Marcotte, E. M. Prediction of gene-phenotype associations in humans, mice, and plants using phenologs. *BMC Bioinformatics* **14**, 203 (2013).

8. Rogers, D. J. & Tanimoto, T. T. A computer program for classifying plants. *Science* **132**, 1115-1118 (1960).

Chapter 4: Relating proteins pharmacologicaly and phenotypically

4.1 Introduction

Protein families are often organized by the similarity of their amino acid sequences or by their functions.[1-4] Another method is to organize them pharmacologically by using the ligands themselves and relating them together.[5, 6] At least with family A GPCRs, the receptors are organized much differently using ligand similarity compared to sequence. Beyond intra-family relationships, ligands often cross protein family boundaries and bind or modulate proteins with completely different folds, function and family.[7, 8] A hypothesis is that proteins evolve partially by the chemical pressures which they face since there are a limited number of primary signaling molecules in an organism but a much more vast space to mutate proteins to signal through these molecules and metabolites.[5, 9, 10]

We previously suggested that primary signaling molecules are constrained by cells responding to signals in multiple time domains and the mechanistic life cycle of the signaling molecule.[5] Examples to support these claims are serotonin, which signal through GPCRs, an ion channel, and a transporter, glutamate and GABA, which signal through ion channels and GPCRs and many others. We aim to take this idea further by identifying new targets for metabolites that signal in different time domains than their known targets. Although not definitive, the direction is to illustrate more examples of how interconnected cell signaling and regulation is with primary metabolites through interactions with different proteins.

Thus far, our examples have primarily focused on drugs or synthetic small molecules and not natural ligands present in humans.[5, 11, 12] Polypharmacology is not only limited to synthetic small molecules, but with endogenous small molecules as well.[9, 13] Although primary signaling

molecules and metabolites can be considered privileged molecules since they are naturally occurring, they still abide by the same biophysical protein-ligand interactions and principles. In healthy individuals, the expression, distribution and function of these ligands are highly regulated by proteins in the cell. However, in non-healthy individuals, over production of growth hormone or cortisol, for example, lead to diseases such as acromegaly and cushing's syndrome. Many neurological diseases are also a result of mis-regulation of neurotransmitters such as dopamine and serotonin. Diseases such as schizophrenia and depression are a result of systemic effects that result from abnormal levels of many neurotransmitters and various neuroreceptors and transporters. All of these examples of neurotransmitters all target multiple proteins in different time domains. The complexity of these biological circuits led us to hypothesize that there must be many more undiscovered targets for metabolites and signaling molecules similar to how almost all drugs and synthetic small molecules have some type of off-target and polypharmacology.

Another approach is to focus on how the proteins are related functionally and even phenotypically in organisms. Whereas function is often related within a family, the emphasis is to relate proteins from different families. Linking proteins by new shared ligands can be a "just so" story [14] or imply deeper biological connections that are not just a result of unexpected biophysical protein-ligand interactions. Pharmacological relationships provide a start in relating proteins together with share ligands. Linking those same proteins that are brought together by pharmacology with the addition of phenotypic and functional information strengthens the protein link by suggesting a deeper biological relationship driven by the ligands.

Taken from the flip side, we suggest that proteins that have similar function, phenotype or are in the same pathway may have common shared ligands. There are many examples of proteins in a biological pathway that are regulated by a primary signaling molecule and then degrade the signaling molecule further down the pathway. An example are the endocannabinoids that signal through the cannabinoid receptors, GPCRs, and are degraded by downstream enzymes such as Fatty acid amide hydrolase (FAAH).[15] Another downstream enzyme in the pathway is Epoxide hydrolase 2 (HYES) which we previously found a shared ligand with upstream cannabinoid receptors.[5] Interestingly, the shared ligand is not chemically similar to the endogenous endocannabinoid yet still binds to both proteins.

To suggest a deep-rooted biological link between distant proteins that share ligands, there should be a functional or phenotypic link. Proteins with very different sequence, structure and family can potentially co-evolve towards each other to bind same ligands and signaling molecules. This implies that proteins with related function or same phenotypes can bind the same ligands so that they can be regulated by the same messengers but act in different time domains so as to not conflict in signaling and cellular function. The proteins can also be expressed at different levels depending on tissue type so different tissues or cells use different proteins to perform the same function, but use the same primary signaling molecules to regulate the proteins.

Phenotypes are generally identified for proteins through gene knockouts in whole organisms. The Online Mendelian Inheritance in Man (OMIM)[16] database catalogues all known diseases that are implicated by genetics to genes and proteins. For model organisms such as

mice, other databases have been compiled from the literature with genetic knockdown or knockout or mutational experiments such as Mouse Genome Informatics (MGI).[17] The compilation of unique gene-phenotype associations has exploded in recent years as the genetic experiments become easier and larger scale for model organisms more so than for humans and the OMIM database. The idea of using phenologs has helped to infer the phenotypic relationships that are discovered for model organisms to gene-phenotype relationships for humans.[18, 19] It is a powerful tool that statistically enriches datasets from model organisms to suggest human phenotype-gene links that otherwise would have or have been missed. The phenologs database provides gene-phenotype links using orthologous phenotypes from different species and potential candidate disease targets. From a systems pharmacology perspective, the database can be applied with SEA to further build networks of proteins together, linked by their function, phenotype or disease, to better refine predicted networks of targets related using the pharmacology of their ligands with SEA.[5, 6]

Adding phenotype, function and disease information to pharmacologically derived protein networks provides a biological relationship between proteins on top of the link from ligand chemistry. By combining the two sets of data, we ask the following questions: Do the proteins linked pharmacologically also have cellular relationships by signaling through the same pathways? Or are implicated biologically in the same disease or phenotype? Can we link proteins that have similar or related phenotypes and do not yet have known shared ligands? We aim to address these questions by identifying new ligands for targets that are predicted to be related pharmacologically and phenotypically.

The ChEMBL[20] database of protein-ligand associations provides a starting point to pharmacologically relate proteins together using SEA. The phenologs database of phenotype, function, or disease annotations to proteins can be mapped on to the SEA associated proteins to filter pharmacologically related proteins more biologically. We develop a new method to relate protein annotations together to filter pairs of proteins that are predicted to be related pharmacologically and functionally. With these new predictions, we experimentally test some pairs by identifying ligands that bind to both proteins. In some cases, we test drugs that potentially explain observed effects of patients with proteins that were previously not implicated with these drugs. Potential biological implications of relating proteins both phenotypically and pharmacologically are discussed.

4.2 Results

I. Relating proteins pharmacologically

The ChEMBL 14[20] database of protein-ligand activity data was used to create sets of ligands for each protein target and compared using SEA to obtain E-values for pharmacological similarity. Over 205,142 pairs had significant E-values less than 1×10^{-5}, which 117,216 pairs already had known shared ligands. The associated pairs were further filtered to only those pairs that were from different protein families to 8,269 pairs. Unfortunately, not all pairs of proteins had phenotypes that we could match onto them and we were left with 2,002 pairs to further investigate.

II. Phenotype annotation similarity metric

The annotations from the phenologs database were largely text based descriptions or general classification terms for each protein. For many proteins there were several seemingly different phenotype annotations. Since phenotype annotations text cannot always easily be compared directly, we came up with a similarity metric relating any two phenotype terms. The similarity metric was based on how many overlapping targets and non-overlapping targets between each pair of phenotypes. It was essentially a Tanimoto coefficient (Tc)[21] for phenotypes and results in a range of 0 to 1 for phenotype similarity. There were 598 pairs of proteins that had at least one phenotype with a $Tc > 0.2$ that we manually inspected for experimental testing.

III. Testing of phenotypic and pharmacologically related protein pairs

Thus far, we have tested 14 compounds for 9 protein pairs. The protein pairs all had one of them as a GPCR and the other targets were either ion channels, enzymes, transporters or nuclear hormone receptors. The shared or related phenotypes ranged from general and systemic terms such as inflammation response or metabolism/homeostasis phenotype to specific indications or effects such as bradykinesia or hypoactivity. For 6 of the compounds, there was either dose response activity for predicted targets or had functional activity in a cellular assay.

A strong link we first observed was between Estrogen Receptor α (ERα) and Histamine Receptors, particularly with Histamine H3 Receptor (HRH3). Estrogen receptors are most commonly associated with breast cancer and reproductive system phenotypes, but also have a homeostasis/metabolism and nervous system phenotypes. HRH3 has phenotypes related to inflammation response and decreased interleukin secretion. Additionally, HRH3 was annotated with homeostasis/metabolism and nervous system phenotypes. (**Figure 4.1**) With the same phenotype annotation and a SEA E-value between the two targets of 9.93×10^{-36}, we looked for potential shared ligands between the two targets. Amongst the most similar ligands between the two sets were two ERα drugs, tamoxifen and bazedoxifene.[22] (**Figure 4.2**) They are first and third generation, respectively, selective estrogen receptor modulators (SERMs) both with clinical side effects similar to increased systemic levels of histamine.[23] The primary signaling hormone estrogen actives mast cells to release histamine but it was unclear how the two receptors are directly linked. We wondered if these SERMs directly modulated Histamine

Receptors. In whole tissue experiments, it was previously suggested that tamoxifen inhibits histamine receptors or a histamine-like receptor through an antiestrogen binding site [24, 25] but was not tested through direct binding of cloned receptor. Experimental testing of cloned histamine receptors showed that hydroxy-tamoxifen, the bioactive metabolite of tamoxifene, inhibited HRH1 and HRH4 with K_i of 2.2 and 2.0 µM, respectively. (**Figure 4.3**) Hydroxy-tamoxifen also inhibited HRH3 91.3% @ 10 µM. Bazedoxifene had a K_i of 2.1 µM at HRH4 and > 10 µM at HRH1. (**Figure 4.3**) It also inhibited HRH3 42.2% @ 10 µM. Unfortunately, both compounds were inactive at HRH1 agonist and antagonist functional assays and HRH2 agonist functional assay. The key HRH3 dose response curve and functional assays have yet to be done.

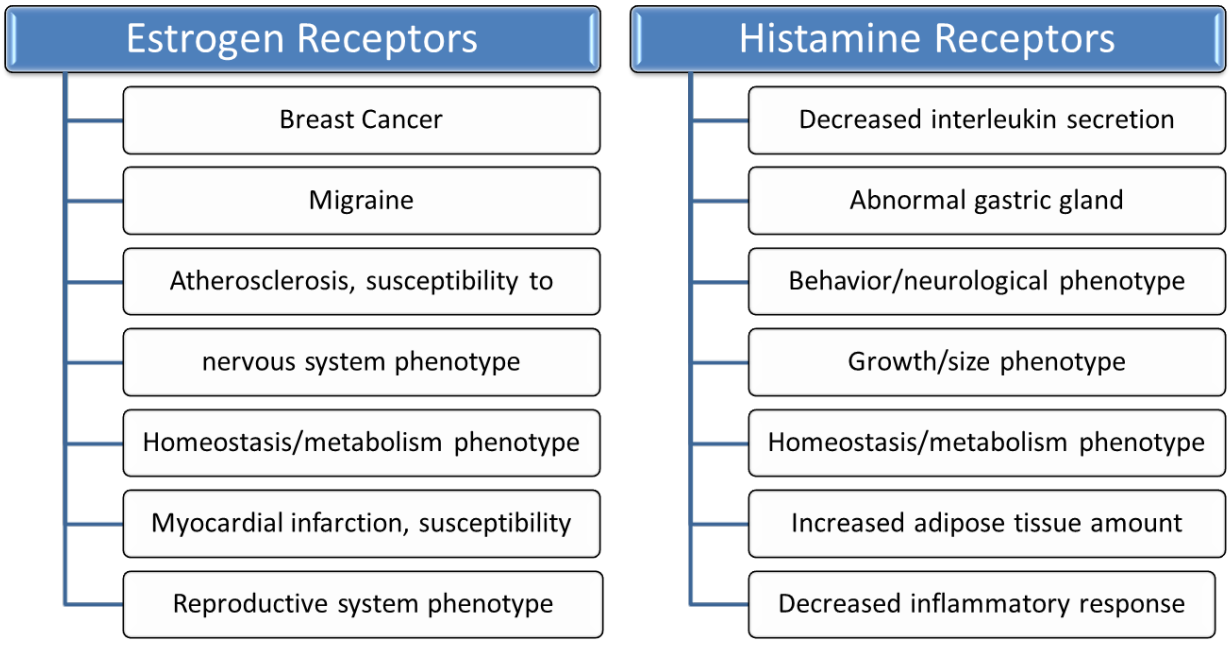


Figure 4.1 Phenotypes for estrogen receptors and histamine receptors
The similar phenotypes between the two receptor families led us to find new ligands to test to link these two target families together.

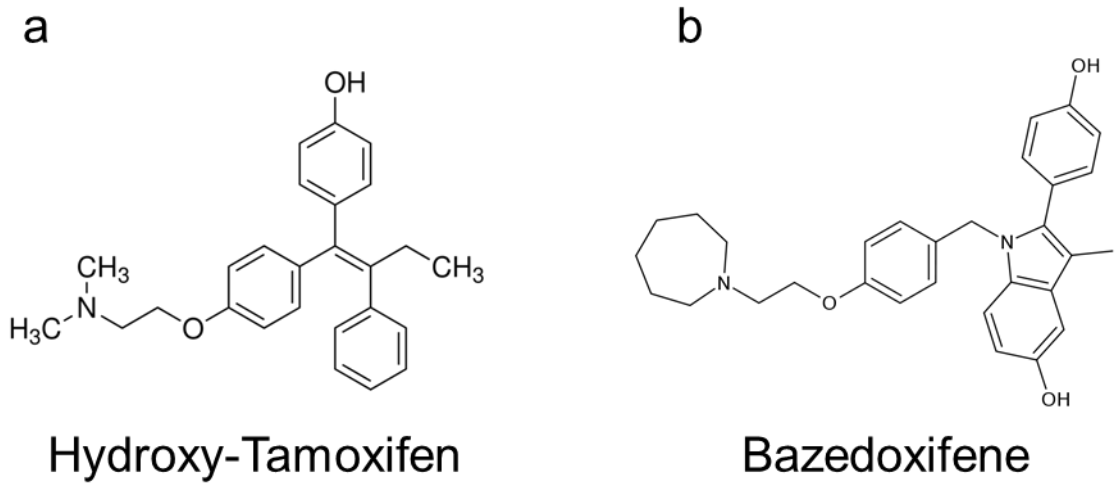


Figure 4.2 Chemical structures of ERα drugs, (a) hydroxy-Tamoxifen and (b) Bazedoxifene.

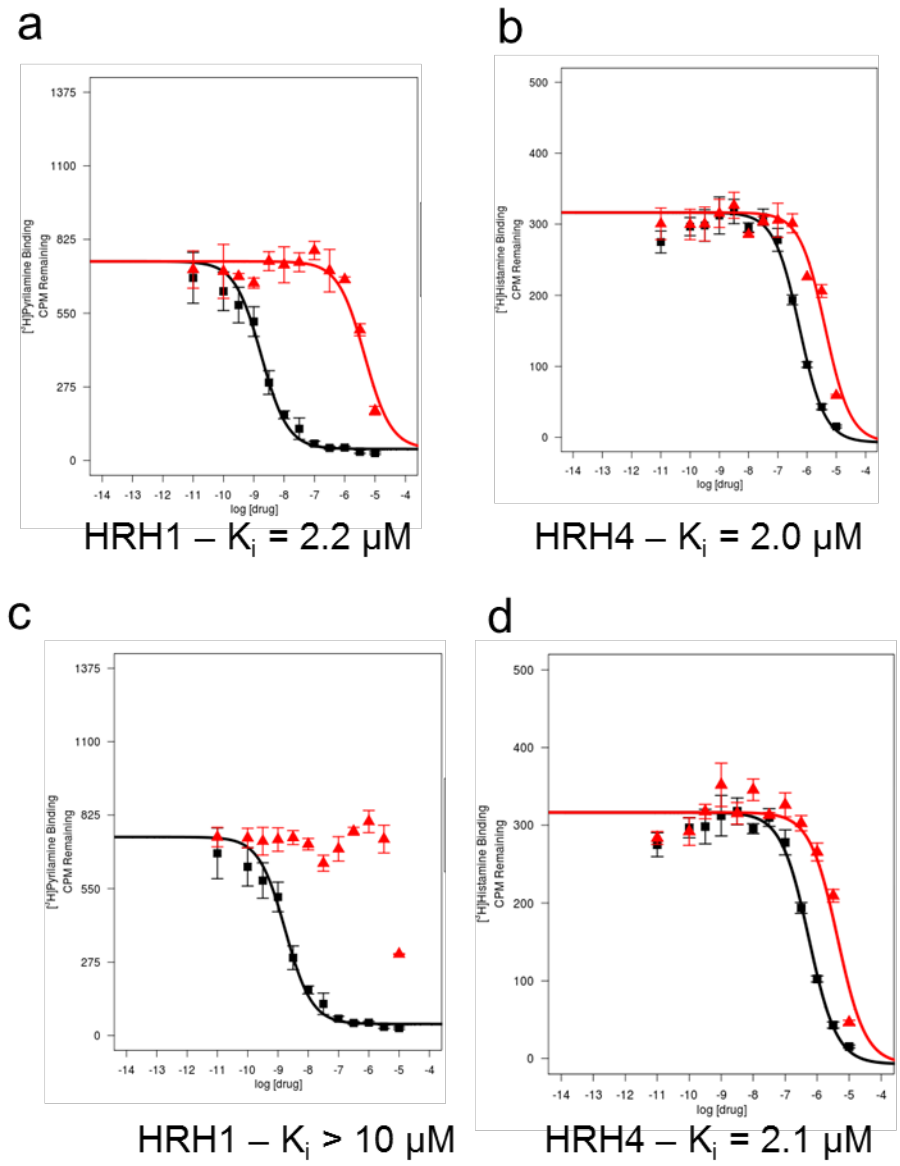


Figure 4.3 Dose-response curves for hydroxy-Tamoxifene and Bazedoxifene against Histamine Receptors

Dose-response curves of Hydroxy-Tamoxifene and Bazedoxifene. (a–d) Radioligand competition binding assay: Hydroxy-Tamoxifene at HRH1 (a) and HRH4 (b), Bazedoxifene at HRH1 (c) and HRH4 (d). Data represent mean values ± s.e.m, performed on triplicate experiments.

Vesicular monoamine transporter 2 (VMAT2) shared many phenotype annotations with Dopamine D2 receptor (DRD2) such as bradykinesia, growth, hypoactivity, impaired

coordination and abnormal dopaminergic neuron morphology. VMAT2 is a transporter of monoamine neurotransmitters, including dopamine, from the cytosol to synaptic vesicles. VMAT2 function is closely tied to Dopamine receptor function since VMAT2 is a regulator of dopamine concentration in the synapses for Dopamine receptors to activate. Inhibition of VMAT2 causes accumulation and storage of monoamine neurotransmitters and disrupts the function of these neurotransmitters. The SEA E-value between VMAT2 and DRD2 was 4.81×10^{-19} and led us to find shared ligands between the two targets. Other than the primary neurotransmitter dopamine, which is transported by VMAT2 and agonizes DRD2, we were unable to find any non-endogenous ligands. Upon analysis of the sets of ligands, we predicted the natural product lobeline would bind to dopamine receptors. Lobeline was already known to inhibit VMAT2, dopamine transporter (DAT), serotonin transporter (SERT), nicotinic acetylcholine receptors (nAchRs) and μ-opioid receptor (MOR). It was an attractive drug to test for direct dopamine receptor binding since it was already known to change dopamine concentrations in synapses, cytosol, vesicles and extracellularly. Lobeline was in clinical trials as a smoking cessation aid[26] and most recently in helping with patients addicted or taking methamphetamines (METH).[27, 28] We tested Lobeline against all five dopamine receptors and measured K_i's of 7,328, 161 and 4,599 nM against DRD1, DRD4 and DRD5, respectively. (**Figure 4.4**) The K_i's for DRD2 and DRD3 were, surprisingly, greater than 10 μM. In the DRD2 functional assay in antagonist mode, Lobeline had an EC_{50} of 1.2 μM with a 69.5 % max efficacy relative to Haloperidol, suggesting its effect as a selective DRD4 antagonist.

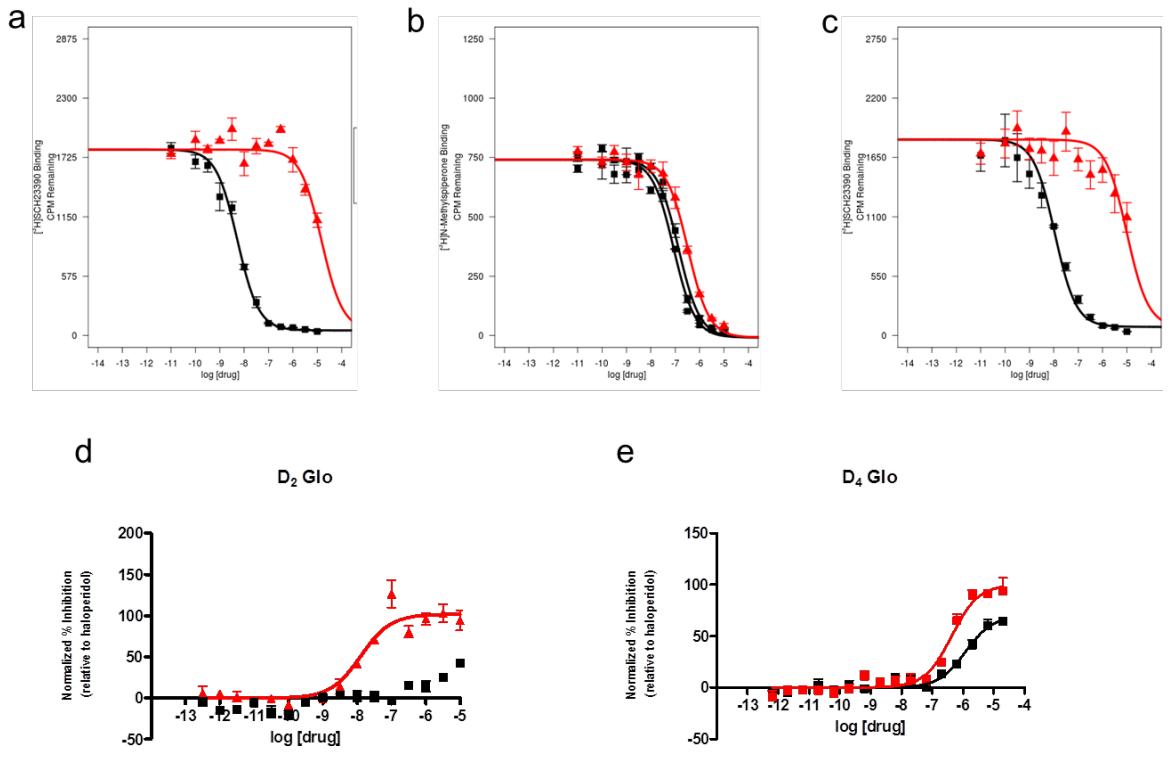


Figure 4.4 Dose-response curves for Lobeline against Dopamine Receptors (DRDs) Dose-response curves of Lobeline. (a–c) Radioligand competition binding assay: Lobeline at DRD1 (a), DRD4 (b) and DRD5 (c). (d-e) Functional antagonist activity represents inhibition of the response to haloperidol at DRD2 (d) and DRD4 (e). Data represent mean values ± s.e.m, performed on triplicate experiments.

Another pair of targets that were brought together phenotypically and pharmacologically was the histone deacetylases (HDACs) and vasopressin receptors (VRs). The phenotypes associated with these proteins were decreased body weight/size and postnatal lethality. Defects in either protein generally lead to early death for model organisms when the genes are deleted. Endogenously, the proteins bind peptides. ChEMBL has annotated both cyclic peptides and small molecules to these targets but there was no shared ligand. The SEA E-values between the two protein subfamilies ranged from 7.38×10^{-80} to 9.81×10^{-15}. We

selected three molecules (AR-42, CI-994 and Entinostat)[29-31] that were known HDAC family inhibitors and tested them against V1AR, V1BR, V2R and Oxytocin receptors. All three molecules were inactive in functional assays with V1AR, V2R and Oxytocin receptors. AR-42 was active in antagonist mode functional assay at V1BR with EC_{50} of 761 nM with 43.2% max inhibition relative to SR49059 (a potent vasopressin receptor antagonist). (**Figure 4.5**)

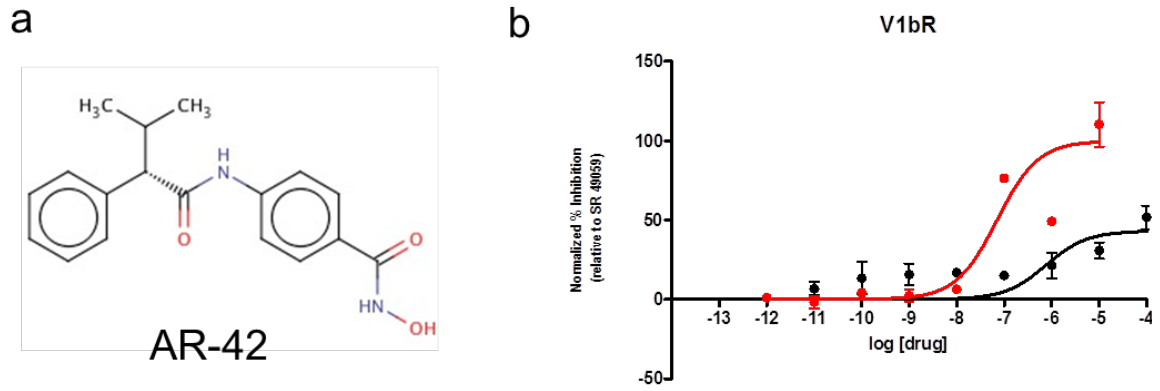


Figure 4.5 AR-42, a potent HDAC inhibitor predicted to inhibit Vasopressin receptors. Chemical structure of AR-42 (a) with dose response curve at Vasopressin 1b receptor (V1bR). AR-42 has slight peptidic character but still a relatively small molecule. Functional antagonist activity represents inhibition of the response to SR49059 with EC_{50} of 761 nM with 43.2% max inhibition. Data represent mean values ± s.e.m, performed on triplicate experiments.

The SK ion channels are small conductance Ca^{2+} activated K^+ channels (KCNNs) that control K^+ ion flow across the cell membrane based on Ca^{2+} concentration. They are largely expressed in the central nervous system but more recently have been shown to be present in cardio myocytes. We were drawn to these ion channels because of the phenotype link with β-adrenergic receptors. Both target subfamilies share prenatal lethality, behavioral/neurological and cardiovascular system phenotypes. Pharmacologically, the two sets of proteins both have

ligands with methylated-catecholamines. Three KCNN inhibitors we studied were laudanosine, NS8593 and bicuculline.[32-34] We tested these compounds in functional assays against all three β-adrenergic receptors. Laudanosine and bicuculline were weak agonists at β_2-adrenergic receptor with EC_{50} of 260 nM with max inhibition of 48.8% and 62 nM with max inhibition of 49.6%, respectively, and inactive against other β-adrenergic receptors. NS8593, a potent and selective KCNN blocker, shows functional agonism at all three β-adrenergic receptors with EC_{50} of 384 nM with max inhibition of 104.4%, 5 nM with max inhibition of 82% and 704 nM with max inhibition of 111.1%, respectively. NS8593 showed higher potency against β_2-adrenergic receptor than the targets the ligand was synthesized for, the KCNNs, and over 75 fold potency over β_1-adrenergic receptor. (**Table 4.1**)

Molecule	Receptor	High cc	Low cc	Top	Bottom	Hill	EC50(nM)
Laudanosine	Beta1	1.0E-04	1.0E-12	-1.0	0	0.0	0
NS8593	Beta1	1.0E-04	1.0E-12	104.4	0	1.0	384
Bicuculline	Beta1	1.0E-04	1.0E-12	-0.5	0	0.0	0

Molecule	Receptor	High cc	Low cc	Top	Bottom	Hill	EC50(nM)
Laudanosine	Beta2	1.0E-04	1.0E-12	48.79	0	1.0	260
NS8593	Beta2	1.0E-04	1.0E-12	82	0	1.0	5
Bicuculline	Beta2	1.0E-04	1.0E-12	49.55	0	1.0	62

Molecule	Receptor	High cc	Low cc	Top	Bottom	Hill	EC50(nM)
Laudanosine	Beta 3	1.0E-04	1.0E-12	1.0	0	0.0	0
NS8593	Beta 3	1.0E-04	1.0E-12	111.1	0	1.0	704
Bicuculline	Beta 3	1.0E-04	1.0E-12	1.1	0	0.0	0

Table 4.1 Agonist dose-response activity data for 3 KCNN inhibitors against β-adrenergic receptors.
Top refers to the maximum percent efficacy of stimulus relative to norepinephrine. NS8593 is highly potent at β2-adrenergic receptor relative to the other subtypes with over 75 fold potency over β1 and over 140 fold potency over β3.

A common phenotype that was shared by many proteins was hypoactivity. Many neurological targets whether they are ion channels, GPCRs, enzymes or transporters had this phenotype annotated for the targets. For example, serotonin transporter, histamine receptors, glutamate receptors, nitric-oxide synthase, GABA receptors, monoamine transporters, dopamine receptors and adenosine receptors all have the hypoactivity phenotype annotated for them. Two proteins that also have this phenotype but also were pharmacologically similar according to SEA were anandamide amidohydrolase 1 (FAAH) enzyme and the somatostatin receptors (SSRs). FAAH degrades bioactive fatty acids such as endogenous cannabinoid, anadamide and oleamide. On the other hand, SSRs bind 14 or 28 amino acid long peptides called somatostatins. Their endogenous molecules are completely different classes of molecules in terms of their chemistry but share the same phenotype. One of the top molecules, in terms of ligand similarity, was the investigational drug Amperozide.[35] Amperozide is a known FAAH inhibitor and binds to serotonin receptors, dopamine receptors, histamine receptor and α-adrenergic receptors. It was investigated for the treatment of schizophrenia and acts through inhibiting neurological targets by decreasing aggression and stress. Interestingly, SSRs and FAAH are both highly expressed in the brain and have hypoactivity as one of their phenotypes so we believed Amperozide to be a suitable ligand to test for SSR activity. In functional SSR agonist assays, Amperozide showed dose response activity against all five SSRs with EC_{50} values of 3,303, 353, 282, 657 and 470 nM, respectively, with max efficacies over 100% for all five receptors. Unfortunately, Amperozide may be toxic to cells at higher concentrations since it caused a decrease in signal similar to an agonist at G_i since it activated all SSRs similarly in the mid nanomolar EC_{50} range. (data not shown)

We tested several other compounds against different pairs of targets but they were inactive either in functional or binding assays. For most pairs we only tested between one to three compounds. There are also several other pairs of compounds that we have predicted to bind to new targets but will soon be tested and moved forward by other members in the Shoichet lab, namely Matthew O'Meara.

IV. Identifying new targets for metabolites and endogenous signaling molecules

The Human Metabolite Database (HMDB)[36] of small molecule metabolites found in the human body was compared to the ChEMBL database of targets using SEA to predict new targets for metabolites and endogenous signaling molecules. The HMDB molecules were first filtered for molecular weight, rotatable bonds and peptide characteristic. This left 6,130 metabolites to predict new targets for. There were 92,872 predictions with SEA E-value better than 1×10^{-5}, of which, 91,009 did not have a max Tc of 1.0. Predictions were manually inspected and further filtered for purchasability and availability for assay.

Thus far, we have 12 metabolites that we have decided to follow up on predictions of new targets. Literature searching was done for each prediction for novelty and to identify potential biological effects or links of predicted target with known function of metabolite. For example, melatonin was predicted to inhibit Sodium channel protein type 5 subunit alpha (SCN5A) by ligand similarity but the literature also suggests that melatonin concentration affects Na^+ currents in cells.[37] The affect in cells can potentially be attributed to inhibition of sodium channels directly. The other metabolites will soon be tested and moved forward by other members in the Shoichet lab, namely Matthew O'Meara.

4.3 Discussion

The relationships between pharmacologically related protein targets do not just stop at having shared ligands, but can also have shared phenotypes. Many drugs bind to and modulate many off-targets that are not structurally or functionally related to the on-target and thus the drugs themselves are not constrained by bioinformatics barriers we easily place on the targets they are intended to bind to.[8, 11, 12, 38] The polypharmacologic properties of drugs extend to many biologically active small molecules as well, without constraint by or within any protein family.[5, 7] By now, these observations should not be surprising since natural and endogenous small molecules and metabolites themselves modulate and bind to many different protein targets and families. Based on these observations, the pharmacologically related proteins may have deeper biological roots and perhaps evolved towards each other by the pressure from signaling molecules and ligands. Furthermore, the evolution of proteins towards binding the same ligands also can co-evolve with function. Proteins that share phenotypes or function may share ligands as well. Complex diseases such as depression or cancer implicate many different classes of proteins from GPCRs, kinases, ion channels, and other enzymes, just to name a few, but are regulated or triggered by a smaller subset of signaling molecules in the body. Identifying related proteins that share the same phenotype was one of the major goals of this study. This was accomplished by finding new ligands to bind to proteins that previously had no shared ligands but had shared or similar phenotypes.

The targets that we related with new shared ligands were from different protein families, were pharmacologically similar by SEA and had shared or similar phenotype

annotations. The relationships were drawn in a predictive manner using various databases of ligand and phenotype annotations and ligands were chosen to be tested experimentally. The estrogen receptors and histamine receptors were brought together by the loose phenotype association of homeostasis/metabolism and nervous system phenotypes. The SERM drugs, tamoxifen and bazedoxifene, were similar enough structurally to histamine ligands and had side effects and patient phenotypes associated with histamine response. The primary signaling hormone, estrogen, also activates mast cells to release histamine.[39] With the supporting information, we were strongly led to directly test SERM drug binding to histamine receptors as a possible explanation for the side effects. Also, the result illustrates how two seemingly unrelated targets that have an inherently linked cellular co-response and regulation could also share phenotypes and ligands.

For proteins that are largely necessary for survival and where deletions and malfunctions lead to death, such as the HDACs and VRs, but by very different mechanisms, it was even more unexpected for the targets to share a ligand. HDACs act on histones and regulate general DNA transcription where in specific tissues or organs, gene deletions lead to developmental disorders and embryonic lethality.[40] On the other hand, mutations in VRs lead to premature death by diabetes insipidus when water reabsorption in the kidneys is not properly regulated.[41] In these cases of such dissimilar mechanisms and functions, it is unclear if finding a shared ligand represents deep rooted biological meaning or is a "just-so"[14] story where peptide binding proteins also share a small molecule ligand.

The link between VMAT2 and dopamine receptors illustrates an interesting example of proteins that act on the same signaling molecules, namely dopamine, and in turn are linked functionally. Lobeline acts not only on VMAT2 but also several neurological GPCRs and transporters and now against dopamine receptors. Although not a synthetic medicinal chemistry compound, the polypharmacological properties of Lobeline are synergistic and add to the efficacy of this drug in that not only is fewer dopamine released by vesicles, the dopamine receptors themselves are antagonized and the reward system is not as active for patients addicted to methamphetamines and other illicit drugs. The ability to predict and target certain combinations of proteins that ligands can bind can have useful implications for drug discovery. The broader implication is that proteins that signal through the same messengers can potentially be regulated or opposed by other molecules and change the outcome of the shared phenotype.

The targets that are annotated for any phenotype can themselves be used to relate phenotypes together. Controlled vocabulary and ontologies are not always conclusive or definitive and having a naïve, biologically unaware method can still result in a signal and quantified metric for phenotype similarity.[42] Similar to how ligands can be used to describe a protein, the set of proteins annotated to a phenotype can also describe it. Taking advantage of the Tanimoto similarity metric for phenotype similarity quickly quantifies an expected similarity between the phenotypes and allowed us to filter the predictions.

As with pharmacological similarity and SEA, phenotype similarity has the same limitations of needing prior information and data. Predictions are inference based and

phenotypes without any target annotations are not applicable for this method. The method is heavily influenced by the amount of data available and completeness of target annotations. The target annotations from OMIM and other databases can be vague or only loosely associated as they are not explicit if they are results from animal models, mutations, deletions, knock-downs or knock-outs. For the purposes of this study, we used as much information as possible to draw protein relationships linked by pharmacology and phenotype. A heavy component of this method still requires manual selection and curation through literature searching for filtering and deciding which targets to pursue.

The limitations presented should not diminish the results presented. Proteins that are from different families but have similar phenotypes can be linked pharmacologically by shared ligands. The examples presented in this study are only a snapshot of the potential unexplored target links possible. The methods presented can be applied and extended when more data is available and experiments can be performed on new predictions. In general, they provide a systematic technique and filter for predictions of target links that are testable. The new small-molecule ligands can also be used as probes to better understand the association and relationships between proteins and help illuminate potential endogenous signaling molecules that may already link the targets but are yet to be discovered.

4.4 Methods

I. Relating proteins pharmacologically with SEA

The ChEMBL 14[20] database of protein-ligand activity data was used to create sets of ligands for each protein target If there were at least 5 compounds annotated with better than 10 μM affinity (i.e. EC_{50}, IC_{50}, K_i, $K_d \leq$ 10 μM). Each target set was then pair wise compared using SEA to obtain E-values for pharmacological similarity. This calculation resulted in over 205,142 pairs with significant E-values less than 1×10^{-5}, which 117,216 pairs already had known shared ligands and were discarded for this analysis. The remaining associated pairs were further filtered to only those pairs that were from different protein families to 8,269 pairs.

II. Filtering SEA results with phenotypes

For each protein target, phenotypes, functions and diseases were mapped using the database of annotations from the phenologs.org website.[18] From the significant pairs of proteins from SEA (8,269 pairs), not both proteins had annotations mapped and we were left with 2,002 pairs to further investigate.

III. Similarity metric to relate phenotypes, functions and diseases

The database of phenotypes, functions and diseases from the phenologs.org website[18] was downloaded and filtered for only those proteins that are in ChEMBL 14 database.[20] For each annotation, the list of proteins associated with that annotation was collected and used as fingerprint bits. Each annotation was then pair wise compared using the proteins as the bits to

calculate a similarity score between each annotation pair using the Tanimoto coefficient (Tc).[21] The more overlap, or shared targets, between two annotations, the more related the annotations are to each other in our calculation. We used this as a text mining and natural language processing method to relate annotations and overcome some of the semantic text issues. Only those pairs of proteins with annotations that had at least a 0.20 Tc did we first consider for relating and experimental testing.

IV. Predicting new targets for metabolites

The Human Metabolite Database (HMDB)[36] of small molecule metabolites found in the human body was filtered for molecular weight < 600, number of rotatable bonds less than 12 and number of oxygens and nitrogens less than 20 to filter for peptide characteristics. The remaining 6,130 metabolites were then compared to the ChEMBL 16 database[20] of targets using SEA. There were 92,872 predictions with SEA E-value better than 1×10^{-5}, of which, 91,009 did not have a max Tc of 1.0. Predictions were manually inspected and further filtered for purchasability and availability for assay.

4.5 References

1. Bateman, A. *et al*. The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-D141 (2004).

2. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25-29 (2000).

3. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).

4. Orengo, C. A. *et al*. CATH–a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1109 (1997).

5. Lin, H., Sassano, M. F., Roth, B. L. & Shoichet, B. K. A pharmacological organization of G protein-coupled receptors. *Nature methods* **10**, 140-146 (2013).

6. Keiser, M. J. *et al*. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197-206 (2007).

7. Hopkins, A. L., Mason, J. S. & Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **16**, 127-136 (2006).

8. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Reviews Drug Discovery* **3**, 353-359 (2004).

9. Tomkins, G. M. The metabolic code. *Science* **189**, 760-763 (1975).

10. Shoichet, B. K. & Kobilka, B. K. Structure-based drug screening for G-protein-coupled receptors. *Trends Pharmacol. Sci.* **33**, 268-272 (2012).

11. Lounkine, E. *et al*. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361-367 (2012).

12. Keiser, M. J. *et al*. Predicting new molecular targets for known drugs. *Nature* **462**, 175-181 (2009).

13. Keiser, M. J., Irwin, J. J. & Shoichet, B. K. The chemical basis of pharmacology. *Biochemistry (N. Y.)* **49**, 10267-10276 (2010).

14. Kipling, R. in *Just so stories for little children* (Oxford University Press, 1998).

15. Di Marzo, V., Melck, D., Bisogno, T. & De Petrocellis, L. Endocannabinoids: endogenous cannabinoid receptor ligands with neuromodulatory action. *Trends Neurosci.* **21**, 521-528 (1998).

16. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-D517 (2005).

17. Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E. & Blake, J. A. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.* **36**, D724-D728 (2008).

18. McGary, K. L. *et al*. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences* **107**, 6544-6549 (2010).

19. Woods, J. O., Singh-Blom, U. M., Laurent, J. M., McGary, K. L. & Marcotte, E. M. Prediction of gene¿ phenotype associations in humans, mice, and plants using phenologs. *BMC Bioinformatics* **14**, 203 (2013).

20. Gaulton, A. *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100-D1107 (2012).

21. Rogers, D. J. & Tanimoto, T. T. A computer program for classifying plants. *Science* **132**, 1115-1118 (1960).

22. Komm, B. S. *et al*. Bazedoxifene acetate: a selective estrogen receptor modulator with improved selectivity. *Endocrinology* **146**, 3999-4008 (2005).

23. Duffy, S. M., Lawley, W. J., Kaur, D., Yang, W. & Bradding, P. Inhibition of human mast cell proliferation and survival by tamoxifen in association with ion channel modulation. *J. Allergy Clin. Immunol.* **112**, 965-972 (2003).

24. Kroeger, E. A. & Brandes, L. J. Evidence that tamoxifen is a histamine antagonist. *Biochem. Biophys. Res. Commun.* **131**, 750-755 (1985).

25. Brandes, L. J. & Bogdanovic, R. P. New evidence that the antiestrogen binding site may be a novel growth-promoting histamine receptor (?H_3) which mediates the antiestrogenic and antiproliferative effects of tamoxifen. *Biochem. Biophys. Res. Commun.* **134**, 601-608 (1986).

26. Stead, L. & Hughes, J. Lobeline for smoking cessation. *The Cochrane Database of Systematic Reviews* **3** (1997).

27. Eyerman, D. J. & Yamamoto, B. K. Lobeline attenuates methamphetamine-induced changes in vesicular monoamine transporter 2 immunoreactivity and monoamine depletions in the striatum. *J. Pharmacol. Exp. Ther.* **312**, 160-169 (2005).

28. Wilhelm, C. J., Johnson, R. A., Lysko, P. G., Eshleman, A. J. & Janowsky, A. Effects of methamphetamine and lobeline on vesicular monoamine and dopamine transporter-mediated dopamine release in a cotransfected model system. *J. Pharmacol. Exp. Ther.* **310**, 1142-1151 (2004).

29. Lin, T. *et al*. AR-42, a novel HDAC inhibitor, exhibits biologic activity against malignant mast cell lines via down-regulation of constitutively activated Kit. *Blood* **115**, 4217-4225 (2010).

30. Loprevite, M. *et al*. In vitro study of CI-994, a histone deacetylase inhibitor, in non-small cell lung cancer cell lines. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* **15**, 39-48 (2005).

31. Fandy, T. E. *et al*. Early epigenetic changes and DNA damage do not predict clinical response in an overlapping schedule of 5-azacytidine and entinostat in patients with myeloid malignancies. *Blood* **114**, 2764-2773 (2009).

32. Scuvee-Moreau, J., Liegeois, J., Massotte, L. & Seutin, V. Methyl-laudanosine: A new pharmacological tool to investigate the function of small-conductance Ca^{2+} -activated K^{+} channels. *J. Pharmacol. Exp. Ther.* **302**, 1176-1183 (2002).

33. Mulholland, P. J. $K_{Ca}2$ channels: Novel therapeutic targets for treating alcohol withdrawal and escalation of alcohol consumption. *Alcohol* **46**, 309-315 (2012).

34. Jenkins, D. P. *et al*. Negative gating modulation by (R)-N-(benzimidazol-2-yl)-1, 2, 3, 4-tetrahydro-1-naphthylamine (NS8593) depends on residues in the inner pore vestibule: pharmacological evidence of deep-pore gating of KCa2 channels. *Mol. Pharmacol.* **79**, 899-909 (2011).

35. Vincent, F., Nguyen, M. T., Emerling, D. E., Kelly, M. G. & Duncton, M. A. Mining biologically-active molecules for inhibitors of fatty acid amide hydrolase (FAAH): Identification of phenmedipham and amperozide as FAAH inhibitors. *Bioorg. Med. Chem. Lett.* **19**, 6793-6796 (2009).

36. Wishart, D. S. *et al*. HMDB: the human metabolome database. *Nucleic Acids Res.* **35**, D521-D526 (2007).

37. Vanecek, J. & Klein, D. Sodium-dependent effects of melatonin on membrane potential of neonatal rat pituitary cells. *Endocrinology* **131**, 939-946 (1992).

38. Gregori-Puigjané, E. *et al*. Identifying mechanism-of-action targets for drugs and probes. *Proceedings of the National Academy of Sciences* **109**, 11178-11183 (2012).

39. Cocchiara, R. *et al*. Modulation of rat peritoneal mast cell and human basophil histamine release by estrogens. *Int. Arch. Allergy Immunol.* **93**, 192-197 (1990).

40. Haberland, M., Montgomery, R. L. & Olson, E. N. The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nature Reviews Genetics* **10**, 32-42 (2009).

41. Lolait, S. J. *et al*. Cloning and characterization of a vasopressin V2 receptor and possible link to nephrogenic diabetes insipidus. (1992).

42. Bard, J. B. & Rhee, S. Y. Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics* **5**, 213-222 (2004).

Conclusions and Future directions

"So what's the key experiment?" Brian would often ask. In an effort to always keep us focused on moving the project forward, Brian would pose this question to us at meetings after we present large amounts of data while still not drawing any conclusions. I was constantly challenged to take a step back and look at the data, predictions and initial reactions as a whole to come up with experiments that would help drive a convincing narrative. In my thesis presented here, I leave a lot of learnings and initial results but still am perplexed and challenged to delve deeper into the fundamental questions we tried to address which were: What biological implications or meaning is behind pharmacologically related proteins? Are the chemical pressures placed on proteins to evolve due to the limited number of primary signaling molecules that proteins can use for cellular function and signaling? There seems to be an endless number of directions to take to try to answer these questions but are there true key experiments to answer these questions? Maybe, maybe not, but alas here I present future directions that attempt to identify the "key experiments".

I. Phenotypic links *in vivo*

Thus far, the proteins linked by new shared ligands are only performed in in vitro binding assays or functional assays in cells. Although examples such as tamoxifen binding to histamine receptors can be explained by observed effects from patients taking the medication, it still does not prove direct causation of tamoxifen binding. A potential experiment would be to test the examples in an *in vivo* study with a model organism. The experimental model would be for the phenotype linked with relevant controls. A compound that is selective for just estrogen

receptors and a compound that is selective for just histamine receptors can be tested in an inflammation response mouse model. These would be the controls of basal response in the animal model. Tamoxifen and bazedoxifene would then be dosed in another mouse study and the results can then be compared with the selective estrogen and histamine receptor ligands. A stronger or higher effect in producing an inflammation response would support the hypothesis that inhibiting both receptor families causes a synergistic effect. These experiments can help illustrate the underlying phenotypic link between the receptors and the effects of modulating them. Unfortunately it would still not be proof of co-evolution driven by pharmacology or their signaling molecules.

II. Identifying new targets for metabolites and endogenous molecules

Previously, we have screened individual drugs against targets from databases such as MDDR, WOMBAT or ChEMBL. This was the first study where we screened metabolites, signaling molecules and endogenous molecules against ligand sets for targets. We have made several new target predictions for these molecules and are just beginning to test them in *in vitro* assays. Discovering new targets for ligands naturally occurring in the body would illuminate biological and cellular activities not previously known. These new chemical probes can be used to discover new function for targets or better understand how the proteins are regulated. If the interactions were only weak, the molecules may have no physiological effect or would only reach high enough concentrations to be active in certain tissues or cell types. Nonetheless, a more complete endogenous molecule/metabolite “interactome” could be mapped out that organizes the proteome as spokes and the signaling molecules as hubs. The results of this map

may illustrate how central ligands are to cellular function and imply that proteins must be an evolving factor since such disparate proteins can bind the same ligands.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature  Date 1/6/2014