Title
Relating protein pharmacology by ligand chemistry

Permalink
https://escholarship.org/uc/item/5vp5h8g4

Author
Keiser, Michael James

Publication Date
2009

Peer reviewed|Thesis/dissertation

Relating protein pharmacology by ligand chemistry

by

Michael James Keiser

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological & Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

To my family

Acknowledgements

Eswar Narayanan put BLAST papers into my hands and the Extreme Value Distribution into my head precisely when each was most necessary. Paul Valiant gave much mathematical guidance; he then informed me of all the ways in which my statistics were wrong, intimated at solutions, and left the remainder as an exercise to the reader. Michael Mysinger implemented libraries during his rotation used by SEA today, and has since suggested many improvements to SEA's random-background routines and to its code. Likewise, Jérôme Hert weather-tested nearly every design decision I made in SEA by implementing its opposite and is, I suspect, the reason I now find myself an author on a paper that proclaims creationism a 'laughable canard.'

Brian Feng has provided sage advice and ideas over many a dinner, and Yu Chen superb scientific conversation and tea. Veena Thomas made sure I went into Orals with a plan and came out with a pass. Sarah Boyce, Kerim Babaoglu, Kristin Coan, and Denise Teotico I thank for their friendship and perspective. Christian Laggner and Henry Lin have taken up the SEA mantle; may they carry it with aplomb. Matt Merski strove mightily to make a chemist out of a computer scientist, and I thank Peter Kolb and Christian for their contributions here too. Pascal Wassam proved to be the most curious and capable systems administrator I have had the pleasure to meet, while Julia Molla and Rebecca Brown have kept the world spinning and all administrative requirements met. I thank Johannes Hermann, Kaushik Raha, Alan Graves, Oliv Eidam, Rafaela Ferreira, Allison Doak, Gabe Rocklin, and Jens Carlsson for many ideas, and also the lab, which has been a most enjoyable place to spend these last five years.

Finally, I thank my parents, Drs. Judy and Wayne, and my sisters, Elizabeth and Jenn, for more than could fit on these pages. This work owes much in its motivation to countless biomedical conversations with my dad over the years (as per bowling ball theory), and my mom's Ph.D. in science education may also be rather relevant to the topic at hand.

I dedicate this thesis to my family.

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

>Keiser, M.J. et al. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* **25**(2), 197-206 (2007).

It appears here with permission from the authors. The supplementary material from this paper has been included as Appendix A.1.

The text of Chapter 2 is a reprint of the material as will appear in:

>Keiser, M.J.*, Setola, V.*, et al. Predicting new molecular targets for known drugs. *Nature* (2009); accepted. *Co-first authors.

It appears here with permission from the authors. The supplementary material from this paper has been included as Appendix A.2.

The text of Chapter 3 is a reprint of the material as it appears in:

>Adams, J.C.*, Keiser, M.J.*, et al. A mapping of drug space from the viewpoint of small molecule metabolism. *PLoS Comput Biol* **5**(8), (2009). *Co-first authors.

It appears here with permission from the authors. The supplementary material from this paper has been included as Appendix A.3.

The text of Appendix B is a reprint of the material as it appears in:

>Keiser M.J., Hert J. Off-target networks derived from ligand set similarity. *Methods Mol Biol* 575, 195-205 (2009).

It appears here with permission from the authors.

Abstract

Relating protein pharmacology by ligand chemistry

Michael James Keiser

The identification of protein function based on biological information is an area of intense research. Here we consider a complementary technique that quantitatively groups and relates proteins based on the chemical similarity of their ligands. We began with 65,000 ligands annotated into sets for hundreds of drug targets. The similarity score between each set was calculated using ligand topology. A statistical model was developed to rank the significance of the resulting similarity scores, which were expressed as networks to map the sets together. Although these networks were connected solely by chemical similarity, biologically sensible clusters nevertheless emerged.

When we used this "Similarity Ensemble Approach" to compare drugs to target sets, unexpected links emerged. Methadone, Emetine, and Imodium were predicted and experimentally found to antagonize muscarinic M_3, α_2 adrenergic, and neurokinin NK2 receptors, respectively. Whereas drugs are intended to be selective, at least some bind to several physiologic targets, explaining their side effects and efficacy. We thereby sought further unexpected links by comparing a collection of 3,665 FDA-approved and investigational drugs against hundreds of targets. Chemical similarities between drugs and ligand sets predicted thousands of unanticipated associations. Thirty were tested experimentally, including the antagonism of the β_1 receptor by the transporter inhibitor Prozac, the inhibition of the 5-HT transporter by the ion channel drug Vadilex, and the antagonism of the histamine H_4 receptor by

the enzyme inhibitor Rescriptor. Overall, 23 additional novel drug-target associations were confirmed, five of which were potent (< 100 nM). The physiological relevance of one, the drug DMT on serotonergic receptors, was confirmed in a knock-out mouse. This Similarity Ensemble Approach is systematic and comprehensive, and may suggest side-effects and new indications for many drugs.

Small molecule drugs also target many core metabolic enzymes in humans and pathogens. We therefore grouped and compared drugs and metabolites by their associated targets and enzymes, mapping these associations onto existing metabolic networks. This revealed what novel territory remains for metabolic drug discovery. We calculated these networks for 385 model organisms and pathogens. Chemical similarity links between drugs and metabolites may suggest drug toxicity, routes of metabolism, and polypharmacology.

Table of Contents

List of Tables

List of Figures

Introduction

Half a decade ago, we set out to build a protein function identification engine. Its proof-of-concept was to span a three-month rotation project—and instead became this thesis. The engine itself remains incomplete, but then, the branching paths of its manufacture were those that led to its most intriguing results. In this introduction, I delineate the motivations for this project, describe its development, and provide a guide to the chapters that follow.

But first, what should we care for protein function? One answer derives from the combination of medicine with Francis Crick's central dogma:[1]

> The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that information cannot be transferred back from protein to either protein or nucleic acid.

If this dogma holds, a protein is the last step in the deterministic flow of sequence information through biopolymers. Unlike DNA or RNA, whose functions are entangled with information flow, a protein's function stands alone. This function comprises actions such as catalysis, trans-membrane signaling, scaffolding, or transport, and we can observe its therapeutic effects. Proteins are the major actors on the biological stage and agents that modulate their actions can directly modulate therapeutic outcomes.

Many proteins interact with small-molecule ligands that modulate their function, and these interactions are not arbitrary. Furthermore, some ligands bind to multiple proteins in blatant defiance of the fact that these proteins have unrelated sequences and structures. For instance, serotonin and serotonergic molecules bind to serotonin receptor subtypes 1, 2, and 4-7

($5\text{-HT}_{1,2,4\text{-}7}$), which are G-protein coupled receptors, but also to an ion channel, the 5-HT_3 receptor.[2] Similarly, the opioid methadone binds not only to the μ-opioid receptor, a GPCR, but also to the NMDA receptor, an ion channel, and both are thought to be involved in the drug's biological activity.[3, 4] Existing bioinformatics approaches that are based on sequence or structure would miss these relationships.

In the days before modern molecular biology, pharmacology and drug discovery efforts operated without knowledge of molecular protein targets. Successful drugs were those that produced desirable phenotypes in disease models and acceptable safety profiles—as measured at the organism level. Shockingly, this approach appears to have been no less productive than our newer target-based efforts.[5] Drug discovery in the absence of targets follows an older logic of drug action, articulated by the structures of the drugs themselves. In this thesis, I seek to combine the strength of both approaches, by applying this older logic *directly to* our understanding of protein targets.

The key technique we borrow from pre-target drug discovery is that of chemical similarity. This follows from the "similarity principle" of chemoinformatics, which states that molecules with similar structures are likely to have similar physicochemical properties and biological activities.[6] Whereas this principle may be violated in specific cases, chemical similarity is often a good guide to the biological action of an organic molecule.[7] With this principle at its core, the protein function identification engine is a *similarity engine*. This entails great limitation but also great power. Just as the requirement of similarity shackles us to the past—which is a haphazard admixture of serendipity, incremental advance, and the sweeping insight of others— so too does this similarity leverage the data accreted over drug discovery's full history, as encoded into the hundreds of thousands of known drug-like molecules. Only recently have these data become readily available.

With the data and the engine at hand, we sought to find and formalize patterns among ligands that would in turn reflect the functional relationships among their respective proteins.

I. Chemical backgrounds

Given the many possible interpretations of protein function, we narrowed our scope to a protein's "pharmacological" function, e.g., that which we can perturb and test by small molecule agents. We worked with ligands derived from patent literature and annotated by their protein target or therapeutic function. To compare these ligands, we turned to "fingerprints," a chemoinformatics tool that computationally encodes small molecules as collections of structural patterns. The identification engine's parts were all in place; one detail remained—the design to bring them together. How were we to actually "find and formalize" these presumed patterns among the ligands, in their teeming thousands?

In the early days, we sought out chemical similarity patterns visually. To do so, we encoded each target-vs.-target comparison as its own histogram, designed to summarize the distribution of ligand structure similarity between two targets (**Figure 1a-b**). This was an extension of rotation-project work done by Morris Feldman. In each histogram, the horizontal axis ran from zero to one hundred percent similarity—any given pair of ligands between the two targets must necessarily fall somewhere within this range. The vertical axis denoted frequency of scores at each bin. It immediately became apparent that most targets did not share even *one* pair of ligands with better than 40% similarity to each other (**Figure 1a**, **Figure 1c**). This seems sensible enough, as one certainly wouldn't expect an androgen receptor agonist to look much like an antifolate—after all; the one target is a nuclear hormone receptor and the other an

enzyme.[1] But how similar would we expect ligands to be for two enzymes that do operate on folate, dihydrofolate reductase (DHFR) and glycinamide ribonucleotide formyltransferase (GART)? These enzymes had ligand pairs between them that scored well, some in excess of 80% (**Figure 1c**).

Most striking, however, was the observation that targets compared to their most perfect match—that is, to themselves—still showed the highest-scoring peaks far down in the 20% similarity region (**Figure 1b**). This could mean several things: Perhaps we weren't representing the molecules well. Perhaps many inhibitors were based on wildly different scaffolds, even among those intended for a single target. Or perhaps it was wrong to think that all ligands for a particular target should look that much alike after all. What if only some did?

---

[1] *Caveat lector*—Chapters 1 and 2 challenge the assumption that ligands of structurally different proteins should themselves always have different chemical structures.

a Protein target vs. unrelated protein target

b Protein target vs. itself

c Excerpt from target-vs-target matrix
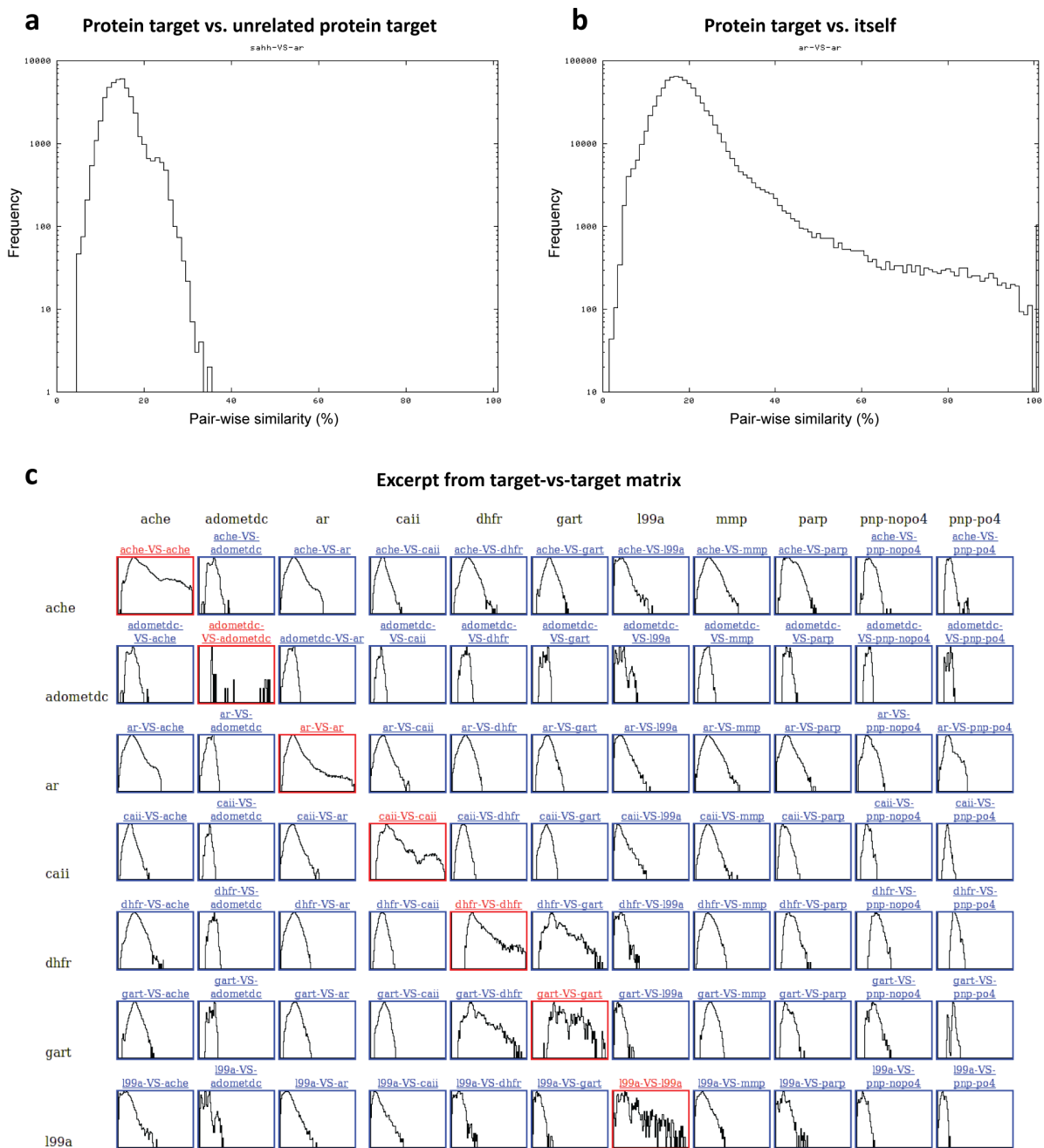
Figure i.1 Early similarity-histogram analyses

Early attempts to find patterns among sets of protein target ligands were qualitative. The histogram in (a) plots the pair-wise similarity scores between each ligand of S-adenosyl-L-homocysteine-hydrolase (SAHH) paired with a ligand of the adenosine receptor (AR); no pair of ligands across sets has better than 38% similarity to each other. In contrast, many ligand pairs in (b) score better than 50% similar, and this is because this histogram compares the set

of AR ligands against themselves. (c) The matrix excerpt here is from an exhaustive comparison of fourteen drug targets. The red histograms on the diagonal are comparisons of the target vs. itself, such as in (b); most of the off-target histograms resemble panel (a).

While intuitive, this approach remained cumbersome on the broad scale; comparing even 14 targets against themselves required significant human review and interpretation (**Figure 1c**). We needed to simplify these results if we were to have any hope of focusing on the most "interesting" target-vs.-target relationships. Later, we realized that by "interesting," we had actually meant "significant;" this was to be the basic motivation for the statistical model we would eventually develop. The first attempt at simplification, however, was a minor one: We automatically removed from the matrix any off-diagonal histogram that lacked at least one pair-wise ligand-ligand score above 50%. In **Figure 1c**, this would remove the majority of the blue histograms shown, while still retaining some of the most interesting ones such as DHFR vs. GART (a target pair that I address in Chapter 1).

The second significant development during this period arose from frequent discussions with Brian Shoichet and John Irwin on the concept of a "chemical background" for these histograms. Over and over again, we saw that the majority of target-vs.-target comparisons yielded uninteresting histograms – ones with not a single ligand-pair scoring over 50% similar. But the shapes of these "uninteresting" histograms gave us pause. They were far from ideal normal distributions, but were more normal than the on-diagonal symmetric-target histograms (colored red in **Figure 1c**). What's more, why did they all tend to fall off by about the 40% similarity mark? What kind of similarity among ligands would we *expect* by random chance, anyhow? We didn't yet know.

My first foray in this direction was wrong. Motivated by the shapes and the scales of the histograms, I built machinery to calculate a random background for each histogram bin position

6

and to normalize the actual histograms on a bin-by-bin basis (**Figure 2**). After doing so, we saw

that the bin-by-bin z-scores were weakest in the 4-20% similarity range, as most random

molecule pairs fell into this region. While this was consistent with our earlier observation that

even nonrandom ligand pairs preferentially scored near the 20% mark, it did little to help us

discriminate the strength of overall similarity among targets that did share highly similar ligands.

Whereas this approach was statistically invalid, it nonetheless brought us closer to a better one.
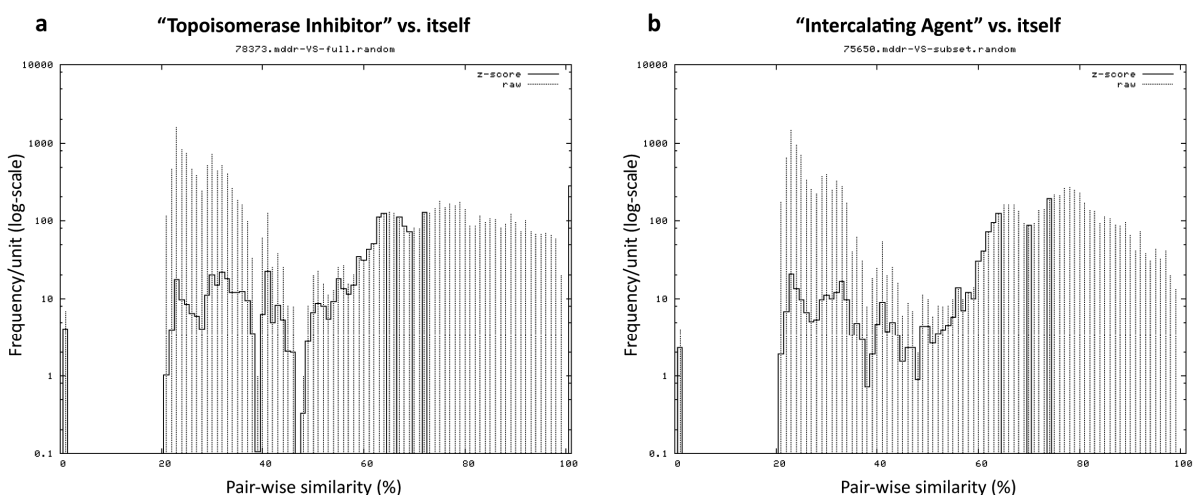
Figure i.2 Initial but incorrect histogram normalization procedure

Panels (a) and (b) present two examples of incorrectly normalized similarity histograms for random chemical backgrounds. I first calculated the mean and standard deviation of completely random molecule-molecule comparisons, evaluated at each pair-wise similarity (x-axis) bin. Then, for each real histogram such as those show here, I expressed each bin's raw (unmodified) count as a z-score (a z-score is the number of standard deviations above the expected mean that a particular raw score achieves). This led to interesting similarity-distribution patterns, but is invalid because it assumed independence of the similarity bins, which is not the case.

II. taniBLAST and SEA

Two people in particular helped enable the transition to the first true version of "taniBLAST," which was the precursor to the Similarity Ensemble Approach: Paul Valiant and Eswar Narayanan. Paul immediately saw that bin-by-bin normalization was wrong and suggested condensing the *entire histogram* into a single score instead. Eswar, a postdoctoral research assistant in Andrej Sali's lab, directed me to empirical BLAST theory for ideas on random background models, and also to papers that described the Extreme Value Distribution underlying the BLAST statistics.[8-10] Michael Mysinger also deserves mention here; during his rotation project, he wrote code for rapid ligand-ligand comparisons using bitwise logic that decreased calculation time by several orders of magnitude and made large-scale random chemical backgrounds practical.

Even so, it was not immediately apparent how, or even that, these parts would fit together—time makes stories out of journeys, and polishes the fits and starts away. Over the next few months, I embarked on several false directions, filled my lab book with forty-five scribbled pages of incremental advances and their failures, and discussed countless ideas with Brian and John. I began to see an analogy to early BLAST empirical models, wherein our comparison of protein targets was similar to BLAST's alignment of protein sequences, our ligand-ligand pairs corresponded to BLAST's heuristic "word" pairs, and in both cases threshold-based "raw" scores with statistical correction could be tuned to fit extreme value distributions. The analogy was a strained one and our ligands were unordered, unlike the residues of a protein sequence—but it was enough to get started.

The only problem was that it didn't work. That is to say, it didn't work until Thursday the 11th of August 2005, lab book #1 page 46, when I finally stumbled upon and slew the last of

the major conceptual bugs in the background generation code.[†] Then, shockingly, it did work. And we called it the Similarity Ensemble Approach, or SEA.

III. Guide to the chapters

This thesis comprises three major chapters based on published first-author or co-first-author papers. Lest the reader wonder at the ties that bind them together, I preface each chapter with a short "gloss" that is both a summary and an attempt to provide research context.

The first chapter introduces the Similarity Ensemble Approach and the global mapping of pharmacological space that it enables; it was published in *Nature Biotechnology* two years ago. In the course of this work, we began to use SEA to predict the protein targets of commercial drugs. We pick up and expand on this theme in the second chapter, scouring all commercial drugs for unexpected target associations predicted by SEA. Chapter 2 also presents a new bipartite view of pharmacological space, where we link drug targets only by the presence of commercial drugs that are known—or predicted—to bind them. In some cases, we find that drugs thought to operate only within a particular class of proteins, such as ion channels, also had unreported activity in a new class, such as GPCRs. This work was recently accepted at *Nature*. In Chapter 3, we explore the use of SEA beyond another boundary; we ask how similar are drug targets, represented by their drugs, to metabolic reactions, represented by their substrates, cofactors, and products.

The fourth chapter discusses future directions. Appendix A provides the full supplementary information for the first three chapters, except where this was not feasible in the case of large datasets. Appendix B is a technical reference for the interested reader, where we

---

[†] Raw scores = 0 were overrepresented in the random background, shifting the distribution to a Gaussian, when in truth they should have been discarded because they represented "no score."

have detailed the steps necessary to generate the pharmacological SEA networks of Chapter 1. It was published contemporaneously with this thesis, as a chapter of *Chemogenomics: Concepts and applications of a new design and screening paradigm*, by Humana Press.

IV. References

1. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

2. Kroeze, W.K., Kristiansen, K. & Roth, B.L. Molecular biology of serotonin receptors structure and function at the molecular level. *Curr Top Med Chem* **2**, 507-528 (2002).

3. Ebert, B., Andersen, S. & Krogsgaard-Larsen, P. Ketobemidone, methadone and pethidine are non-competitive N-methyl-D-aspartate (NMDA) antagonists in the rat cortex and spinal cord. *Neurosci Lett* **187**, 165-168 (1995).

4. Callahan, R.J., Au, J.D., Paul, M., Liu, C. & Yost, C.S. Functional inhibition by methadone of N-methyl-D-aspartate receptors expressed in Xenopus oocytes: stereospecific and subunit effects. *Anesth Analg* **98**, 653-659, table of contents (2004).

5. Cohen, F.J. Macro trends in pharmaceutical innovation. *Nat Rev Drug Discov* **4**, 78-84 (2005).

6. Johnson, M.A. & Maggiora, G.M. Concepts and applications of molecular similarity. (Wiley, New York; 1990).

7. Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J Med Chem* **40**, 1219-1229 (1997).

8. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

9. Karlin, S. & Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* **87**, 2264-2268 (1990).

10. Pearson, W.R. Empirical statistical estimates for sequence similarity searches. *J Mol Biol* **276**, 71-84 (1998).

Gloss to Chapter 1

The chapter that follows describes the development and testing of the Similarity Ensemble Approach (SEA) in the world of soi-disant perfect data. It considers SEA's first applications to protein-protein relationships and its first extension to drug-target predictions. In doing so we ask, how related are drug targets? To what extent does chemical similarity reflect the biological and pharmacological relationships present among targets? Conversely, does a chemical-centric method like SEA actually tell us anything new? And what does it mean when we get results that we didn't expect—are they indeed new, or merely wrong?

To address these questions, we built networks of target-target similarity and also compared these SEA networks to BLAST alignments of the target sequences. We considered many approaches to the target networks and ultimately built maps from minimum spanning trees[1] to simplify the tangle of inter-target relationships, as was particularly apparent among neurological receptors. To address the question of unexpected results, however, we needed predictions amenable to experimental testing. But this was problematic; although we had many anecdotal successes using SEA to rediscover relationships among drug targets that were otherwise hidden in the literature, it was not clear how we could validate a truly new relationship between two targets. Would two "related" targets bind some of the same ligands? If so, to which ligand, when established drug targets often have tens or hundreds to choose from? Methadone emerged as a serendipitous solution, both providing the foundation for our experimental results in Chapter 1, and by example setting the course for the entire research direction of Chapter 2.

Methadone is a synthetic opioid agonist that also antagonizes the *N*-methyl-D-aspartic acid (NMDA) receptor and is thus a prime example of "polypharmacology" – the phenomenon

wherein a single drug binds multiple protein targets, often in a way that contributes to its function. Bryan Roth coined the term "magic shotguns" for drugs that hit multiple targets, in contrast to the traditional view of a drug as a "magic bullet" that should hit just one.[2] Methadone was an apt example because the μ-opioid receptor is metabotropic (a G-protein coupled receptor, or GPCR) whereas the NMDA receptor is ionotropic (an ion channel). It is thus a single drug that binds two proteins of wildly different structure and, presumably, evolutionary history. This example illustrates how a chemo-centric view of drug-to-target relationships may find links that a sequence- or structure-centric view cannot. Of course, it would only illustrate this if SEA could actually predict it. But we had developed SEA as a means of comparing drug *targets* against each other—could we use it to compare single drugs against targets?

Well, yes. If the set of known ligands for a particular target was large enough, the statistics would support a very small set—even one of "size 1," which is to say, a single drug—on the other side of the equation. This was, however, a substantial divergence from that way we had originally envisioned SEA, and thus its first application to a single drug was actually to the set of methadone plus its 20 closest neighbors[‡] from the ZINC database. Reassuringly, SEA recapitulated methadone's known binding to the opioid and NMDA receptors as I had expected. Less reassuringly, the expectation values (E-values) for these predictions were weak—especially when compared to the much stronger SEA prediction of methadone's extraordinary similarity to the known muscarinic M_3 antagonists. This was initially disheartening. On review, however, Brian and John confirmed that methadone's chemical structure really *did* look like that of several of the antimuscarinics. On the recommendation of Mark von Zastrow, we contacted Bryan Roth

---

[‡] The field is littered with cases of "a single methyl" whose lack throws off chemical-similarity predictions or, conversely, the self-same predictions not taking into account the catastrophic effect of an extra such methyl. In any case, for methadone's SEA predictions, I later verified that it did not make a substantial difference whether we included the additional 20 derivatives or not.

at the Psychoactive Drug Screening Center, who ran the experimental tests that ultimately confirmed methadone's antimuscarinic binding and functional activity. It was exciting; we'd found unreported and therapeutically-relevant off-target binding for a drug that has been in use for over seventy years.

Could we do it again? We noticed that SEA predicted unexpected protein targets for two more drugs, emetine and loperamide (tradename Imodium). SEA found emetine had high structural similarity to α_2 adrenergic receptor ligands, and that loperamide had high structural similarity to neurokinin NK2 receptor ligands. As emetine is associated with congestive heart failure and loperamide was thought to modulate only the NK3 receptor (albeit indirectly, via opioid receptors[3]), these targets seemed feasible. In collaboration with Bryan Roth, we confirmed both experimentally.

I. References

1. Kruskal, J. On the shortest spanning subtree and the traveling salesman problem. *Proceedings of the American Mathematical Society* **7**, 48-50 (1956).

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

3. Kojima, S., Ikeda, M. & Kamikawa, Y. Loperamide inhibits tachykinin NK3-receptor-triggered serotonin release without affecting NK2-receptor-triggered serotonin release from guinea pig colonic mucosa. *J Pharmacol Sci* **98**, 175-180 (2005).

Chapter 1:

Relating protein pharmacology by ligand chemistry

Michael J. Keiser,[1,2] Bryan L. Roth,[3,4] Blaine N. Armbruster,[4] Paul Ernsberger,[3] John J. Irwin,[1]*

and Brian K. Shoichet[1]*

1   Department of Pharmaceutical Chemistry, University of California San Francisco, 1700 4th St, San Francisco CA 94143-2550

2   Biological and Medical Informatics, University of California San Francisco, 1700 4th St, San Francisco CA 94143-2550

3   Departments of Biochemistry (BLR) and Nutrition (PE) and NIMH Psychoactive Drug Screening Program Case Western Reserve University Medical School, 2109 Adelbert Road, Cleveland, OH 44106

4   Department of Pharmacology (BLR, BNA) and Division of Medicinal Chemistry and Natural Products (BLR), The University of North Carolina Chapel Hill Medical School, Chapel Hill, NC 27705

* Corresponding authors (BKS and JJI)

1.1 Abstract

The identification of protein function based on biological information is an area of intense research. Here we consider a complementary technique that quantitatively groups and relates proteins based on the chemical similarity of their ligands. We begin with 65,000 ligands annotated into sets for hundreds of drug targets. The similarity score between each set is calculated using ligand topology. A statistical model was developed to rank the significance of the resulting similarity scores, which are expressed as a minimum spanning tree to map the sets together. Although these maps are connected solely by chemical similarity, biologically sensible clusters nevertheless emerged. Links among unexpected targets also emerged, among them that methadone, emetine, and loperamide may antagonize muscarinic M3, α2 adrenergic, and neurokinin NK2 receptors, respectively. These predictions were subsequently confirmed experimentally. Relating receptors by ligand chemistry organizes biology to reveal unexpected relationships that may be tested directly by the ligands themselves.

1.2 Introduction

It is a curious pharmacological fact that related drugs and biological messengers can bind to receptors that appear unrelated by many bioinformatics metrics. For instance, serotonin and serotonergic drugs bind to G-protein coupled receptors (GPCRs) such as the 5-hydroxytryptamine subtypes 1, 2, and 4-7 ($5\text{-}HT_{1,2,4\text{-}7}$), but also to an ion channel, the $5\text{-}HT_{3A}$ receptor.[1, 2] Ionotropic and metabotropic 5-HT receptors are unrelated by sequence and structure, yet both are involved in the pharmacological effects of serotonergic drugs. Similarly, the well-known opioid methadone binds not only to the μ-opioid receptor, a GPCR, but also to the NMDA receptor,[3] an ion channel, and both are thought to be involved in the drug's

biological activity.[4] Benzodiazepines affect mitochondrial proteins in addition to their primary

therapeutic actions on ion channels.[5] The enzymes thymidylate synthase (TS), dihydrofolate

reductase (DHFR), and glycinamide ribonucleotide formyltransferase (GART) all recognize folic

acid derivatives and are inhibited by antifolate drugs. Despite this, the three enzymes have no

substantial sequence identity and are structurally unrelated. This disregard for typical biological

categories on the part of small molecules can lead to infamous side effects—although cisapride

stimulates 5-HT$_4$ receptors and astemizole inhibits histamine H$_1$ receptors, both also inhibit the

hERG ion channel, leading to unexpected cardiac pathologies.[6] The ability of chemically similar

drugs to bind proteins without obvious sequence or structural similarity can confound a purely

biological logic to understanding and categorizing their action.

A chemo-centric approach to this problem is to compare not the biological targets

themselves but rather the chemistry of their ligands.[7] The motivating hypothesis is that two

similar molecules are likely to have similar properties,[8] and will bind to the same group of

proteins. Whereas this hypothesis may be violated in specific cases—a small change in chemical

structure can dramatically change binding affinity—chemical similarity is often a good guide to

the biological action of an organic molecule.[9] Indeed, chemical similarity is a central principle in

ligand design,[10] and an extensive chemoinformatic literature explores many methods to compare

pairs of ligands for such similarity.[11] Recently, Hopkins and colleagues found that using the

simplest form of chemical similarity, full chemical identity among ligands shared by two or more

receptors, linkage maps can be calculated to relate targets.[12] Vieth and colleagues, using a

different approach, have used dendrograms of inhibitors to organize the selectivity relationships

among kinases.[13] Izrailev and Farnum have also linked ligand sets by focusing on the most

similar molecules between them.[14] These and recent efforts in predicting pharmacologic

profiles[15-19] have led to the development of probabilistic models to predict polypharmacology and assess the "druggability" of protein targets.

Here we investigate techniques to relate receptors quantitatively based on the chemical similarity among their ligands. In this method, which we call the Similarity Ensemble Approach (SEA), two sets of ligands are often judged similar even though no single identical ligand is shared between them. We use a collection of about 65,000 ligands annotated for drug targets, where most annotations contain hundreds of ligands. To compare sets without size or chemical composition bias, we introduce a technique that corrects for the chemical similarity we might expect between ligand sets at random, using a model resembling that of BLAST.[20-22] This technique enables us to link hundreds of ligand sets—and correspondingly the protein targets— together in minimal spanning trees. Whereas these trees are calculated by chemical similarity, recognizable clusters of biologically related proteins emerge from them. We consider the origins and possible significance of both the recognized and unexpected relationships, and their use for uncovering side effects and polypharmacology of individual chemical agents. We test several such unexpected relationships in biochemical and cell-based assays.

1.3 Results

I. Similarity scores between ligand sets

We used a 246-receptor subset of the MDL Drug Data Report (MDDR), which annotates ligands according to the receptor whose function they modulate. Each ligand in each set was compared to each ligand in every other set. Overall, 246 versus 246 set comparisons were made, involving 65,241 unique ligands and 5.07×10^9 total ligand pairs. Tanimoto coefficients (Tc) of chemical similarity were calculated for each pair of ligands. For most ligand pairs the Tc was low,

in the 0.2 to 0.3 range, which is typically considered insubstantial similarity. This was true even when comparing a set to itself. For instance, when comparing the 216 ligands of the antifolate enzyme dihydrofolate reductase to themselves, 80.4% of the pairs had a Tc in the 0.1 to 0.4 range, with only 4.7% having more substantial scores in the 0.6-1.0 range and only 0.5% having Tc of 1.0 (only 216 ligands are, after all, identical) (Figure 1). This pattern was also observed comparing the 253 ligands of the antifolate enzyme thymidylate synthase to the DHFR ligands. Here only 0.06% of ligand pairs were identical (Tc of 1.0), 1.6% of pairs had Tc values of 0.6 to 1.0, and 85.5% had Tc values between 0.1 and 0.4. When the set of 1226 ligands for the protease thrombin was compared to that of DHFR, a peak containing 97.1% of all pairs was observed between Tc values of 0.1 to 0.4, but no identical pairs were observed nor were there any ligand pairs that had Tc values greater that 0.5. The raw similarity score, which is the sum of ligand pair Tc's over all pairs with Tc≥0.57 (see Methods), between the DHFR and thrombin ligand sets was therefore 0; the raw score between DHFR and TS ligand sets was 772.25, whereas that of the DHFR set against itself was 1931.60. This is consistent with the lack of similarity between the ligand sets of thrombin and DHFR, but the considerable similarity between the sets of TS and DHFR, both of which contain related antifolate drugs and their analogs.

DHFR Set Similarity Distributions
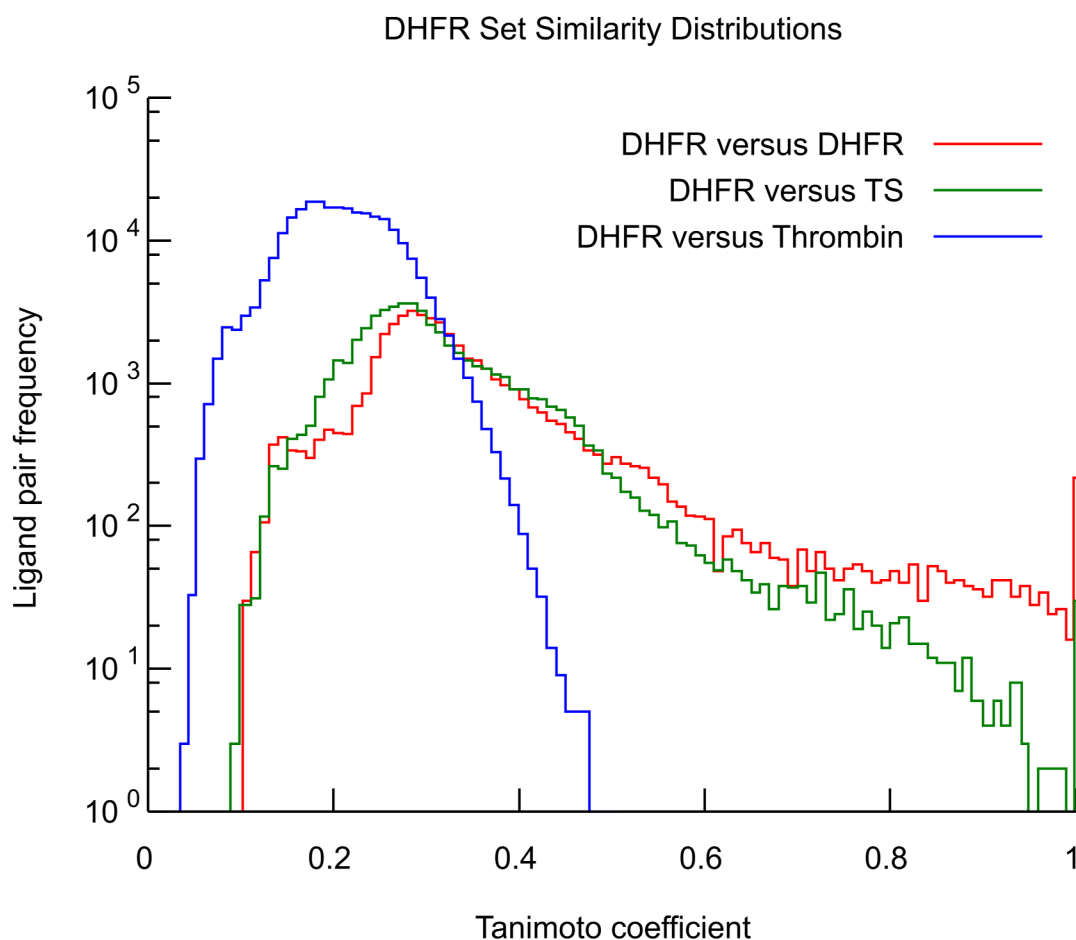
Figure 1.1 Comparing similar and dissimilar ligand sets to that of DHFR

Log-scale distributions of ligand-ligand similarity for different ligand sets: dihydrofolate reductase (DHFR) ligands compared to themselves (red), DHFR ligands compared to the related thymidylate synthase (TS) ligands (green), and DHFR ligands compared to the unrelated thrombin ligands (blue). The Tc ranges from 0 (complete dissimilarity) to 1 (identity). The ligand sets were derived from MDDR annotations.
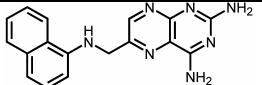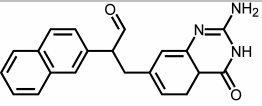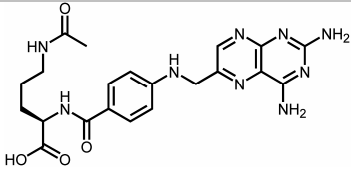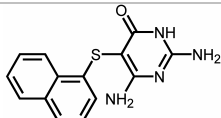
II. Patterns of similarity

Most pairs of ligand sets resembled the TS vs. thrombin comparison and had no raw score similarity. Of the 60,516 set pairs, 70.8% had raw scores of 0. As the size of the sets grew,

however, the likelihood that two would have pairs of ligands with Tc≥0.57 also grew. Indeed, there was a linear relation between the raw score and the number of ligands in the sets being compared (see **Supplementary Figure 1**). To compare the significance of the set similarity raw scores across sets of different sizes, we developed a statistical model of the similarity we would expect at random for sets drawn from the same large but finite database of ligands (see **Methods**). This allowed us to calculate Z-scores and expectation values for any raw score for ligand sets of any size, such that the background fit an extreme value distribution (see **Supplementary Figure 1c**). As far as we know, a statistical model for random set similarity has not been previously used in chemoinformatics (although Z-scores have been used for comparisons of individual compounds[23, 24]). As in sequence comparisons, the expectation values that such a model allows are critical for unbiased and quantitative comparison of multiple ligand sets. As would be expected, 95.2% of set-to-set comparisons had expectation values greater than one. The similarity of the overwhelming majority of ligand sets was thus no greater than what one would expect at random. Returning to the comparison of DHFR, TS, and thrombin, the DHFR set vs. itself had a Z-score of 333.4 and an expectation value of 7.07×10^{-182} (**Table 1**), suggesting very high similarity, whereas DHFR vs. TS had a Z-score of 117.6 and an E-value of 1.11×10^{-61}. As DHFR vs. thrombin did not yield a raw score >0, no Z-score was calculated and the comparison was unranked.

Table 1.1 MDDR activity classes resembling MDDR "Dihydrofolate Reductase Inhibitor"

Rank	Activity Class	E-value	Example Molecule
1	Dihydrofolate Reductase Inhibitor	7.07×10^{-182}	
2	Glycinamide Ribonucleotide Formyltransferase Inhibitor	3.97×10^{-100}	
3	Folylpolyglutamate Synthetase Inhibitor	4.59×10^{-62}	
4	Thymidylate Synthase Inhibitor	1.11×10^{-61}	

With a model of random similarity we could compare statistically weighted versions of the raw scores for all pairs of sets. Even fewer sets had statistically significant similarity after correction for random expectation. On average, any given receptor was similar to only 5.8 other receptors with an expectation value better than 10^{-10}. Further down the rank-ordered list, the expectation values among targets fell off steeply, and within a few targets the similarity typically fell to insignificance. For example, the set of α-amino-5-hydroxy-3-methyl-4-isoxazole propionic acid (AMPA) receptor antagonists was highly similar to two other ligand sets: Kainic acid antagonists and *N*-methyl-D-aspartic acid (NMDA) antagonists, with E-values of 5.28×10^{-80} and 3.08×10^{-63}, respectively. The third most significant ligand set was the anaphylatoxin receptor antagonists, with an E-value of 3.81×10^{-4}, and by the sixth ranked target the similarity was insignificant (E-value 1.00×10^{-1}, **Table 2**; for more detail see **Supplementary Table 1**). Correspondingly, few targets were unrelated to any others; only 18 such orphans were found (see **Supplementary Table 2**). A few targets were relatively promiscuous, with 14 being related to more than 10 other targets with expectation values better than 10^{-50}.

Table 1.2 MDDR activity classes resembling five example MDDR activity classes

Query	Rank	Size	Similar Activity Classes	E-value	TC1.0	Max TC
AMPA Receptor Antagonist	1	569	AMPA Receptor Antagonist	2.45×10^{-219}	577	1.00
	2	75	Kainic Acid Receptor Antagonist	5.28×10^{-80}	74	1.00
	3	1485	NMDA Receptor Antagonist	3.08×10^{-63}	181	1.00
	4	22	Anaphylatoxin Receptor Antagonist	3.81×10^{-4}	0	0.70
	5	130	μ Agonist	1.69×10^{-3}	0	0.83
	6	99	Ribonucleotide Reductase Inhibitor	1.00×10^{-1}	0	0.73
Carbacephem	1	98	Carbacephem	0*	106	1.00
	2	1614	Cephalosporin	1.11×10^{-222}	14	1.00
	3	35	Isocephem	2.30×10^{-17}	0	0.64
	4	257	Penem	2.43×10^{-4}	0	0.68
	5	13	Oxacephem	8.38×10^{-3}	0	0.69
	6	39	Lactam (β) Antibiotic	2.62×10^{-2}	0	0.62
	7	223	Lactamase (β) Inhibitor	6.58×10^{-1}	1	1.00
	8	116	Monocyclic β-Lactam	3.18×10^{2}	0	0.61
Androgen	1	50	Androgen	0*	138	1.00
	2	577	Aromatase Inhibitor	6.87×10^{-307}	0	0.88
	3	43	Antiglucocorticoid	2.30×10^{-102}	0	0.89
	4	6	Cytochrome P450 Oxidase Inhibitor	4.01×10^{-93}	0	0.92
	5	179	Estrogen	9.97×10^{-89}	0	0.91
	6	86	Antiestrogen	2.18×10^{-76}	0	0.84
	7	936	Steroid (5α) Reductase Inhibitor	1.58×10^{-72}	0	0.80
	8	103	Antiandrogen	1.14×10^{-70}	0	0.99
	9	86	17α-Hydroxylase/C17-20 Lyase Inhibitor	7.88×10^{-66}	0	0.76
	10	164	Progesterone Antagonist	3.26×10^{-44}	0	0.89
	11	62	Prostaglandin	1.93×10^{-38}	0	0.75
5 HT1F Agonist	1	111	5 HT1F Agonist	6.72×10^{-187}	113	1.00
	2	621	5 HT1D Agonist	8.08×10^{-38}	0	0.95
	3	51	5 HT1B Agonist	2.96×10^{-10}	0	0.95
	4	65	5 HT1 Agonist	3.03×10^{-8}	0	0.81
	5	670	Dopamine (D4) Antagonist	1.90×10^{-6}	0	0.79
	6	565	5 HT1A Antagonist	8.64×10^{-1}	0	0.71
	7	33	5 HT2 Antagonist	8.78×10^{-1}	0	0.65
	8	705	5 HT2A Antagonist	1.47	0	0.73
Adrenergic (β1) Agonist	1	8	Adrenergic (β1) Agonist	3.85×10^{-241}	10	1.00
	2	305	Adrenergic (β) Agonist	9.50×10^{-34}	0	0.81
	3	67	Adrenergic (β1) Blocker	4.99×10^{-32}	0	0.64
	4	563	Adrenoceptor (β3) Agonist	2.98×10^{-24}	0	0.72
	5	212	Adrenergic (β) Blocker	3.96×10^{-13}	0	0.78
	6	13	Adrenergic, Ophthalmic	2.77×10^{-7}	0	0.70

Query	Rank	Size	Similar Activity Classes	E-value	TC1.0	Max TC
	7	518	Adrenergic (α1) Blocker	6.84×10^{-5}	0	0.73
	8	124	Melatonin Agonist	1.04×10^{-1}	0	0.63
	9	76	Dopamine (D1) Agonist	2.18×10^{-1}	0	0.71
	10	102	Adrenergic (α2) Agonist	4.72×10^{-1}	0	0.66

* E-value < 10^{-320}

The similarity of ligand sets to small archipelagos of other ligand sets allowed us to calculate maps connecting almost all sets together through sequential linkage (**Figure 2a**). In this map and in the sparser minimal spanning tree, where we connect only the most similar neighbors (**Figure 2b**), clusters of biologically related targets may be observed as an emergent property, as no explicit biological information, only ligand information, is used to calculate the cross-target similarity. Thus, the glutamate receptors group together (**Figure 2b**), and the steroids localize around androgen and estrogen receptor ligands (**Figure 2b iv**). Likewise, the folate, phosphodiesterase, and β-lactam sets each co-localize and intra-connect (**Figure 2b**). Conversely, whereas the serotonin metabotropic receptors cluster together, and ionotropic ligand receptors do so as well, the two receptor subtypes are distinct (**Figure 2b ii, 2b iii**). Similar clustering may be observed in other regions of the map.
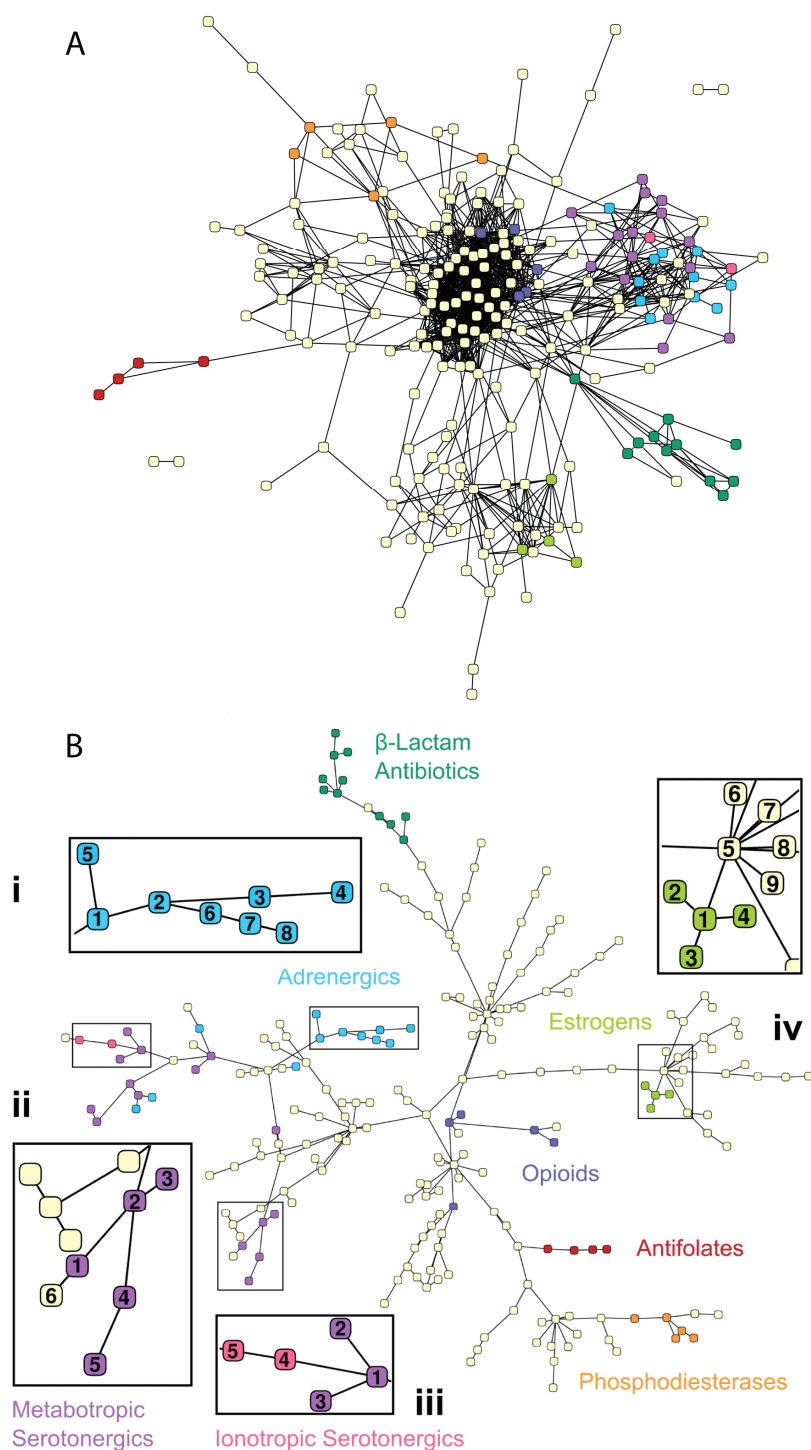
Figure 1.2 Similarity maps for 246 enzymes and receptors

A. Network view of pharmacological space, in which each node represents a particular target in the MDDR. The nodes are colored for several pharmacologically related targets: antifolates (red), phosphodiesterases (orange), opioids (blue), β-lactam antibiotics (dark green),

metabotropic serotonergics (violet), ionotropic serotonergics (pink), adrenergics (cyan), and estrogen modulators (light green). This network is a naïve threshold graph that only includes edges that have expectation values better than one. **B.** A tree view of pharmacological space. This is an alternate view of the same network as in part (A), over which we have calculated a minimal spanning tree. This approach connects all nodes (protein targets) using only the most significant connections. The node coloring is the same as that in part (A). i) Detailed view of adrenergics: β adrenergic agonists (1), β_1 adrenergic agonists (2), β_1 adrenergic blockers (3), β adrenergic blockers (4), β_3 adrenoceptor agonists (5), ophthalmic adrenergics (6), α_2 adrenergic agonists (7), and α_1 adrenoceptor agonists (8). ii) Detailed view of metabotropic serotonergics subset: 5-HT_{1F} agonists (1), 5-HT_{1D} agonists (2), 5-HT_1 agonists (3), 5-HT_{1B} agonists (4), and 5-HT_{1D} antagonists (5). iii) Detailed view of ionotropic (5-HT_3) serotonergics: 5-HT_4 agonists (1), 5-HT_4 antagonists (2), 5-HT_2 antagonists (3), 5-HT_3 antagonists (4), and 5-HT_3 agonists (5). iv) Detailed view of steroids: estrogens (1), antiestrogens (2), estrone sulfatase inhibitors (3), estrogen receptor modulators (4), androgens (5), HMG-CoA reductase β inhibitors (6), antiandrogens (7), aromatase inhibitors (8), and glucocorticoids (9).

For this method to have wide utility, it is important that sets of ligands from different sources – for instance, not just from within the MDDR – can be compared. To test this, we built 23 ligand sets from 1,421 compounds in PubChem Compound (http://pubchem.ncbi.nlm.nih.gov/) that were not in the MDDR, organized by their MeSH Pharmacological Actions. We then queried these sets against our collection of 246 MDDR activity classes and ranked them by ligand-set pharmacological similarity (**Table 3**). Of the 23 PubChem query sets, 17 found a matching MDDR activity class as the top-ranked hit. When repeated using the mean pair-wise similarity (MPS)[14, 25, 26] of the sets instead of the statistically-corrected expectation values, only 9 of the queries found a matching top-ranked hit. On average, a matching MDDR hit was found within the top 1.4 ranks of the PubChem queries' hit lists using pharmacological similarity (SEA), compared to within the top 8.2 ranks when ranked by

MPS (see **Supplementary Table 3**). This attests to the importance of a statistical control for similarities expected at random.

III. Comparison to sequence similarity

The statistical model for ligand set similarity allowed us to directly compare the resulting E-values with those derived from sequence comparison. We mapped 193 MDDR activity classes to their protein target sequences and determined the sequence similarity among them using PSI-BLAST.[27] We then computed a heat map highlighting the differences between pharmacological similarity and sequence similarity among these targets (**Figure 3a**). In this heat map, many ligand sets with enzyme targets were pharmacologically similar but sequence-dissimilar. Examples include folate recognition enzymes and adenosine binding enzymes (**Figure 3b**). By comparison, many neurological receptors had stronger sequence than pharmacological similarity (**Figure 3c**).

Table 1.3 Out-group comparison of 1,421 PubChem compounds organized into 23 MeSH pharmacological actions vs. 246 MDDR activity classes

	Size	MeSH Pharmacological Action	Pharmacological similarity		Mean pair-wise similarity	
			MDDR Activity Class	E-value	MDDR Activity Class	MPS
1	131	Adrenergic α-Antagonists	Adrenergic (α) Blocker	1.18×10^{-22}	Somatostatin Analog	0.287
2	138	Adrenergic β-Agonists	Adrenergic (β1) Agonist	1.54×10^{-203}	Adrenergic (β1) Agonist	0.395
3	132	Adrenergic β-Antagonists	Adrenergic (β1) Blocker	6.65×10^{-77}	Adrenergic (β1) Agonist	0.370
4	30	Androgen Antagonists	Androgen	4.54×10^{-125}	Androgen	0.300
5	21	Androgens	Androgen	0	Androgen	0.551
6	10	Aromatase Inhibitors	Androgen	4.36×10^{-108}	Androgen	0.226
7	29	Carbonic Anhydrase Inhibitors	Carbonic Anhydrase Inhibitor	1.24×10^{-152}	Carbonic Anhydrase Inhibitor	0.269
8	11	Cholinergic Antagonists	Anticholinergic	4.80×10^{-155}	Anticholinergic	0.396
9	91	Cholinesterase Inhibitors	Acetylcholinesterase Inhibitor	1.87×10^{-70}	Melatonin Agonist	0.207
10	98	Cyclooxygenase Inhibitors	Androgen	4.50×10^{-58}	3-Hydroxyanthranilate Oxygenase Inhibitor	0.249
11	111	Dopamine Agonists	Dopamine Agonist	5.50×10^{-120}	Adrenoceptor (α2) Antagonist	0.306
12	52	Estrogen Antagonists	Antiestrogen	3.56×10^{-112}	Antiestrogen	0.281
13	20	Estrogens	Estrogen	0	Estrogen	0.401
14	80	Glucocorticoids	Glucocorticoid	0	Glucocorticoid	0.506
15	34	Histamine H2 Antagonists	H2 Antagonist	1.47×10^{-53}	H2 Antagonist	0.248
16	20	HIV Protease Inhibitors	HIV-1 Protease Inhibitor	8.41×10^{-108}	Somatostatin Analog	0.378
17	28	Lipoxygenase Inhibitors	Lipoxygenase Inhibitor	2.05×10^{-16}	Melatonin Agonist	0.245
18	106	Muscarinic Antagonists	Anticholinergic	2.67×10^{-151}	Anticholinergic	0.343
19	22	Nicotinic Agonists	Nicotinic Agonist	3.00×10^{-22}	Anaphylatoxin Receptor Antagonist	0.297
20	94	Phosphodiesterase Inhibitors	Phosphodiesterase I Inhibitor	8.33×10^{-25}	Anticholinergic, Ophthalmic	0.227
21	86	Protease Inhibitors	Renin Inhibitor	2.25×10^{-78}	Anaphylatoxin Receptor Antagonist	0.334
22	65	Reverse Transcriptase Inhibitors	Thymidine Kinase Inhibitor	1.63×10^{-145}	Thymidine Kinase Inhibitor	0.333
23	12	Trypsin Inhibitors	Trypsin Inhibitor	3.14×10^{-19}	3-Hydroxyanthranilate Oxygenase Inhibitor	0.346
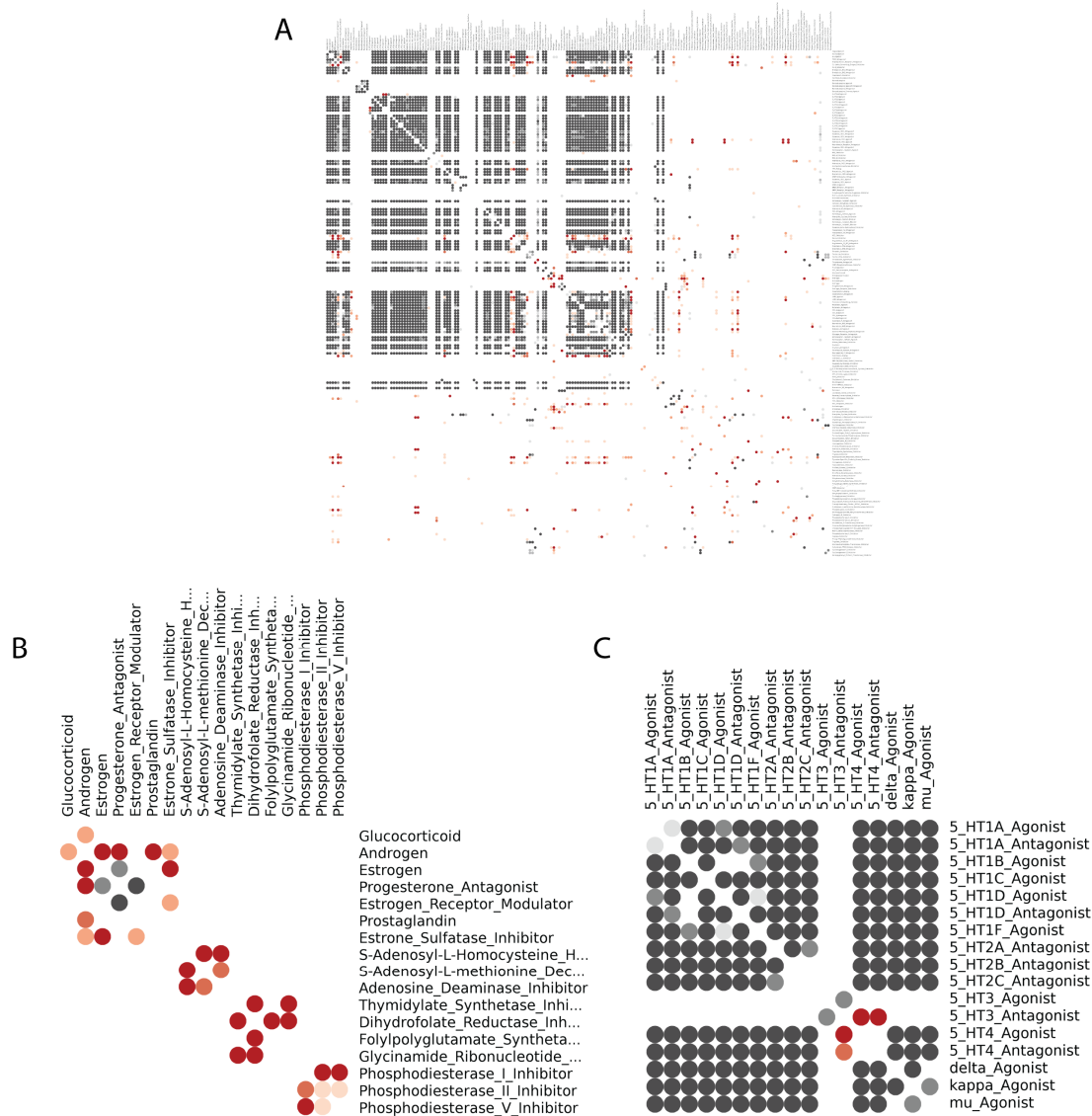
Figure 1.3 Comparison of sequence and ligand-based protein similarity

In this difference heat map, red ellipses mark activity class pairs with strong ligand-set similarity but weaker sequence similarity. Activity classes that map to EC numbers often fall into this category. Dark gray regions mark target pairs with strong sequence similarity but comparatively lower ligand-set similarity. This region includes many GPCRs, ion channels, and nuclear hormone receptors; such receptors may share evolutionary history but have often diverged in terms of pharmacological function. The white regions mark cases where pharmacological and sequence similarity approaches agree. This heat map was calculated by taking the difference of the two log-space heat maps available in the supplementary materials (see **Supplementary Figure 6** and **Supplementary Figure 7**).

IV. Predicting and testing drug promiscuity

We were interested in exploring the behavior of single agents that were known to have either promiscuous or off-target actions. An example of the latter was methadone, known to have dual specificity for NMDA and μ-opioid receptors. Methadone is an unusual chemotype for μ-opioid agonists, one that is not represented in the MDDR, although it and several congeners can be found in PubChem. Because of this, when the methadone ligand set was queried against all 246 MDDR targets, the μ-opioid ligands were only found as the third-ranking hit. Unexpectedly, the set of methadone and its analogs was found by this method to be far more similar to the antimuscarinics activity class, particularly the M3 receptor antagonists (**Table 4**). This attests to the MDDR's known false-negative problem,[28] but more provocative was the predicted M3 antagonism, as methadone is not known to have muscarinic activity. To test this possibility experimentally, the affinity and activity of methadone on M3 muscarinic receptors was measured by direct binding and a cell-based functional assay. Methadone was observed to have a K_i of 1.0 μM (**Figure 4a**) and to antagonize activation of M3 receptors, consistent with the prediction (**Figure 4b**).
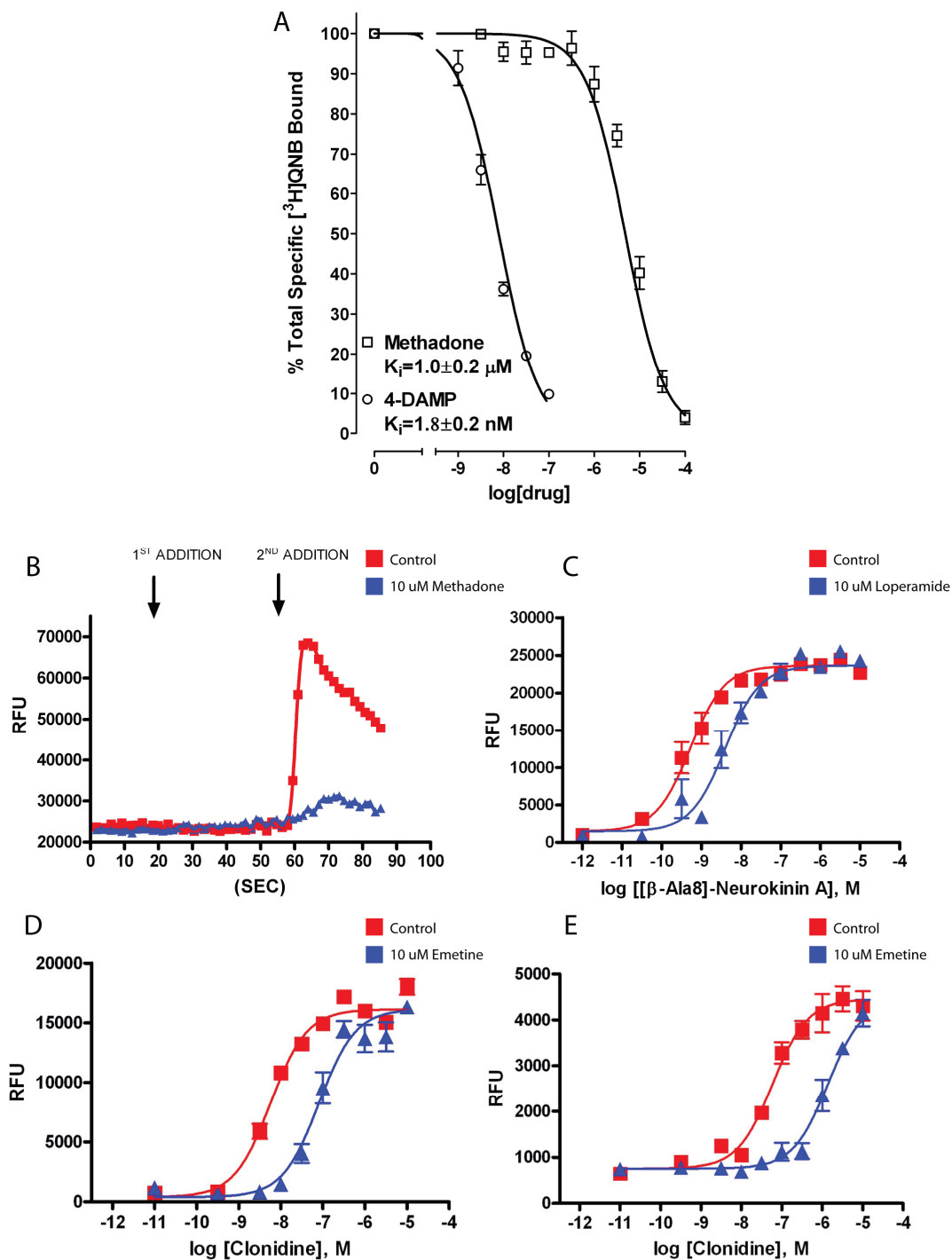
Figure 1.4 Testing the off-target activities of Methadone, Loperamide, and Emetine
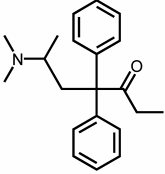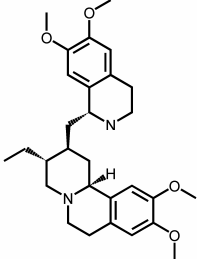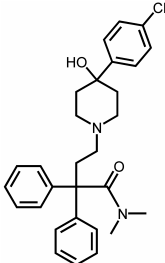
A. Antagonism of M3 muscarinic receptors by the μ opioid agonist methadone in a direct binding assay. Competition binding curves with [³H]quinuclidinyl benzilate in membrane fractions from CHO cells stably transfected with the human M3 muscarinic receptor. Each

data point represents the mean and standard error of 4 conducted in duplicate or quadruplicate. Competition curves represent the best fit to a single-component logistic equation (GraphPad Prism 4.0, San Diego, CA). Two-site models did not yield a better fit. Membranes were incubated for 60 min at 25°C with 0.5 nM [^{3}H]quinuclidinyl benzilate and increasing concentrations of competing drug. Incubations were terminated by rapid vacuum filtration. Nonspecific binding was defined in the presence of 1.0 μM atropine and represented less than 10% of total binding. **B.** Methadone antagonism of M3 muscarinic receptors by functional assay. Either methadone (10 μM final concentration) or vehicle was added at T=20 sec (1st addition), and then at T=50 sec (2nd addition) 1 μM carbachol was added to CHO-M3 cells and intracellular Ca++ mobilization was measured,

as previously described.[34] Dose-response curves (not shown) indicated that methadone was a competitive antagonist at M3-muscarinic receptors. **C.** Loperamide antagonism of Neurokinin NK2 receptors. Dose responses of CHO cells expressing Neurokinin NK2 receptors treated with [β-Ala8]-Neurokinin A were measured following administration of either DMSO vehicle or 10 μM loperamide. **D and E.** Emetine antagonism of adrenergic receptors. Dose response of clonidine treatment of MDCK cells expressing either **D)** alpha 2a adrenergic or **E)** alpha 2c adrenergic receptors after incubation with DMSO vehicle or 10 μM emetine. Shown are representative curves, mean values ± SEM, of intracellular calcium release experiments performed in quadruplicate for each drug concentration per pre-treatment condition as described in **Methods**.

Table 1.4 Novel target selectivity predictions for three existing drugs

Query	Rank	Size	Activity Class	E-value	Max TC
Methadone*	1	188	Antimuscarinic	4.45×10^{-50}	0.77
	2	266	Muscarinic M3 Antagonist	1.22×10^{-11}	0.67
	3	68	Opioid Agonist	1.84	0.61
	4	1485	NMDA Receptor Antagonist	9.04	0.67
	5	975	Muscarinic (M1) Agonist	61.9	0.60
	6	717	Cyclooxygenase Inhibitor	12.1	0.61
Emetine	1	277	Adrenergic (α2) Blocker	4.34×10^{-118}	0.85
	2	564	Dipeptidyl Aminopeptidase IV Inhibitor	6.50×10^{-17}	0.94
	3	180	Dopamine (D1) Antagonist	1.23×10^{-10}	0.74
	4	1820	Substance P Antagonist	25.8	0.64
	5	288	Dopamine (D3) Antagonist	179	0.61
	6	212	Neurokinin NK3 Antagonist	2.76×10^{4}	0.60
Loperamide	1	462	Neurokinin NK2 Antagonist	1.55×10^{-20}	0.75
	2	1820	Substance P Antagonist	2.12×10^{-15}	0.75
	3	212	Neurokinin NK3 Antagonist	2.63×10^{-14}	0.66
	4	518	Adrenergic (α1) Blocker	1.64×10^{-10}	0.72
	5	583	Protein Kinase C Inhibitor	1.45×10^{-1}	0.63
	6	266	Muscarinic M3 Antagonist	2.42	0.59

Note: No query compound was already present in the reference 246 MDDR activity classes, and thus the TC1.0 (identity) column is omitted.

* While methadone was compared as a set of analogs (see **Methods**), only the structure for methadone itself is displayed for clarity.

Emboldened by this result, we looked for other single compounds with novel off-target effects. To increase the chance of novel action, we screened PubChem compounds—many of which are not in the MDDR database—against 246 MDDR targets. Over 12,000 PubChem compounds with annotated activities were compared to the MDDR ligand sets, using an automated procedure, looking for those where the target annotated in PubChem differed from

that of the highest scoring MDDR set, using SEA. For the vast majority of the resulting 6,000 high-scoring hits, the annotations differed only trivially and could be rapidly excluded by post-filtering (e.g., "androgen antagonist" is formally different from "steroid antagonist", but not in an interesting way). There were, however, 30 PubChem compounds that had very low (good) expectation values against genuinely unrelated MDDR categories. Two stood out by visual examination of their structures and by our ability to actually acquire and test them in the appropriate assay. These were the drugs emetine and loperamide, which were predicted to antagonize adrenergic α2 and neurokinin NK2 receptors, respectively, based on set similarities (**Table 4**) (see also Methods). Both predictions were tested by functional assay: 10 µM emetine was observed to induce 10.6- and 27.5-fold increases in the EC_{50} of the α2-agonist clonidine for α2a and α2c adrenergic receptors, respectively, and 10 µM loperamide induced a 7.5-fold increase in the EC_{50} of the NK2 agonist [β-Ala8]-Neurokinin (**Figure 4c,d,e**, see **Supplementary Table 4**). Assuming competitive binding, these results put the affinity of emetine for the adrenergic receptors in the 400 nM to 1 µM range, and the affinity of loperamide for NK2 receptors in the 1-2 µM range.

1.4 Discussion

Protein targets may be quantitatively related by their ligands—a Similarity Ensemble Approach reveals both expected and unexpected similarities that may be tested by the 'off-target' activities of the ligands themselves. Three aspects of these similarities merit particular emphasis. First, most ligand sets are highly related to only a few others; the vast majority of ligand sets are unrelated. Second, there are nevertheless enough connections among them to link almost all sets together, through sequential linkages, in coherent maps of pharmacologically interesting chemical space. Third, biologically related targets cluster in these maps. No biological

information was used to make these connections, only ligand chemistry, and such clustering is an emergent property of this technique. It is also an imperfect property, in that the clusters of targets can differ from those expected from biological information alone. Both the expected and unexpected connections among the ligand sets have implications for understanding the effects of bioactive molecules, and lead to testable hypotheses.

The similarity of the ligand sets to only a few others owes to the intrinsic chemical differences between most sets and to the statistical model's discrimination between significant (e.g., E-value < 1×10^{-10}) and insignificant (e.g., E-value > 1.0) similarity. In the case of dihydrofolate reductase inhibitors, for instance, the three most related target sets are the folate recognition enzymes glycinamide ribonucleotide formyltransferase, folylpolyglutamate synthetase (FPGS), and thymidylate synthase, with expectation values ranging from 3.97×10^{-100} to 1.11×10^{-61}; i.e., highly significant. The next most related set had no measurable similarity and the other 241 are even less related (**Table 1**). Likewise, AMPA receptor antagonists score strongly against both kainic acid receptor and NMDA receptor antagonists (**Table 2**); all three are all ionotropic glutamate receptors traditionally subdivided into NMDA and non-NMDA types.[29] A key point is that many related targets would be missed if ligand identity was substituted for chemical similarity between sets, i.e., if we only related sets that shared common ligands (the flip side of this is that many large ligand sets would be related artifactually if we did not control for similarity expected at random). For instance, the antiglucocorticoids, estrogen agonists, estrogen antagonists, progesterone antagonists, and prostaglandins all rank as highly similar to the androgen agonists, as is sensible (**Table 2, Figure 2b iv**). Yet not one of these sets shares a single ligand with the androgens (**Table 2**). Correspondingly, serotonergic 1F agonists closely resemble serotonergic 1B, 1D, and $5-HT_1$ agonists and D_4-dopamine receptor antagonists without sharing a single ligand in common (**Figure 2b ii, Table 2**); the same is true for the

relationship of β_1 adrenergic receptor agonists to other β-receptor agonists and antagonists (**Figure 2b i**).

Related by chemical similarity, almost all of the 246 receptors may be mapped, through intermediate receptors, to all others. We found it convenient to interrogate this map interactively: clicking on any node displays a table of all nearest ligand set neighbors, including the molecules that comprise any given set (available at http://sea.docking.org). Thus, different classes of β-lactam antibiotics cluster together in this map, as do the several classes of phosphodiesterase inhibitors (**Figure 2**). The serotonergics form their own branch of the tree, with the ionotropic (5-HT$_3$) agents isolated (**Figure 2b iii**), just as the androgens and estrogens group closely but separately (**Figure 2b iv**).

Another way to view such clustering is through a heat map that compares ligand-set with sequence similarities between the same targets (**Figure 3a**). When the ligand-set and sequence similarities agree, as with μ receptor agonists vs. δ receptor agonists (**Figure 3c**) and neurokinin NK2 antagonists vs. NK3 antagonists, the matrix element in the heat map is white (it will also be white when there is neither sequence nor ligand-set similarity). Such correspondences are comforting, but more interesting are those targets for which the chemoinformatic and bioinformatics techniques disagree. Many target sequences are more similar than their ligand sets (dark gray matrix elements). For instance, the serotonin 5-HT$_{1A-C}$ subtypes are highly related by sequence but less so by ligand sets (**Figure 3c**), although the latter are not dissimilar. However, the serotonergics are also highly similar to the opioids by sequence, yet the ligands are different (**Figure 3c**); much of this similarity arises from non-ligand-binding regions. Conversely, some targets unrelated by sequence are closely related by ligand sets (red matrix elements in **Figure 3**). Thus, the antifolates cluster together even though DHFR, GART, TS, and FPGS are dissimilar by sequence (**Figure 3b**). The differences between the chemoinformatic and bioinformatics

views have several bases, among them that sequence similarity arises from evolutionary history, but chemoinformatic similarity and dissimilarity arise from the state-of-the art of medicinal chemistry. Indeed, designing the specificity necessary to pharmacologically distinguish receptor subtypes, such as 5-HT$_{1A}$, 1B, and 1C, is a longstanding goal of medicinal chemistry, one executed in the teeth of their evolutionary relationships. Both the similarities and dissimilarities between the chemoinformatic and bioinformatics views lead to testable hypotheses.

Perhaps the most compelling result of this study is the experimental testing of three different drugs against targets to which they were not previously known to bind. We looked for candidate drugs based on known polypharmacology or on ligand-set similarities between targets with no clear precedence for cross-reactivity in the literature (see **Methods**). Methadone attracted us because of its well-known polypharmacology, modulating both NMDA and μ opioid receptors. Surprisingly, methadone most resembled the ligand-set of M3 muscarinic receptor antagonists (**Table 4**). Both by direct binding and by functional assay, we find that methadone is a 1 μM antagonist of the M3 receptor, consistent with prediction (**Figure 4a,b**). As far as we know, methadone's action on M3 muscarinic receptors is unprecedented, although a pharmacophore model that may be related to its promiscuity has very recently appeared.[30] Intriguingly, its affinity for the M3 receptor is consistent with some of the side-effects of this drug,[29, 31] which reaches micromolar steady-state concentrations in patients.[32] Emetine and loperamide are further examples of drugs that resemble, by the Similarity Ensemble Approach (SEA), target classes that they are not known to modulate. Emetine is an amebicide that inhibits polypeptide chain elongation in parasites.[33] By SEA, it has striking similarities to the adrenergic α2-blocker ligand-set, with an expectation value of 4.3×10^{-118} (**Table 4**). Consistent with that similarity, we find that emetine antagonizes α2 receptors in the micromolar and possibly sub-micromolar range (**Figure 4d,e**, see **Supplementary Table 4**). Although this activity has not, to

our knowledge, been previously reported, it is consistent with the known side-effects of this drug, which can lead to hypotension, tachycardia, dyspnea, myocarditis, and congestive heart failure. Loperamide (Imodium), is an opioid that is used for relief of diarrhea via action on μ-opioid receptors in the gut[29] (**Table 4**). The drug closely resembles the neurokinin NK2 antagonist ligand-set, when compared by the SEA method (**Table 4**). Consistent with that prediction, we find that loperamide antagonizes NK2 receptors in the micromolar concentration range (**Figure 4c**, see **Supplementary Table 4**). Intriguingly, loperamide has been observed to modulate neurokinin NK3-receptor-triggered serotonin release, though this has been thought to be through its action on opioid receptors.[34] The results of this study suggest that the drug also has a direct effect on neurokinin receptors.

The polypharmacology of drugs and bioactive molecules emerges at the confluence of two channeled currents: medicinal chemistry's elaboration of new molecules, and the molecular evolution of biological function. Fortuitously, this channeled elaboration relates receptors and enzymes frequently enough to link almost all targets together in a single map of chemically relevant biology, and when the background of random possibilities is controlled for, specifically enough to distinguish the significant links from a stochastic sea of possibilities. In the minimum spanning trees that are one result of this analysis, many proteins with related functions cluster together. Thus, ion channels and GPCRs that have no obvious sequence or structure similarity are linked quantitatively based on their bioactive ligands. An advantage of this way of relating biological receptors is that it is articulated through the very agents used to probe biology experimentally—drugs and related reagents. The hypotheses that emerge from this analysis thus may be subjected to experiment, and to this end we have made the relationships and linkage maps among the targets studied here publicly available (http://sea.docking.org). The predictions

and subsequent experimental observations that methadone, emetine, and loperamide act as muscarinic M3, adrenergic α2, and neurokinin NK2 antagonists suggest that at least some of the predicted relationships merit investigation.

1.5 Methods

I. Ligand sets

We extracted ligands from compound databases that annotate molecules by therapeutic or biological category. Multiple ligands in any annotation defined a set of functionally related molecules. As a source of ligands we used the 2006.1 MDL Drug Data Report (MDDR),[35] a compilation of about 169,000 drug-like ligands in 688 activity classes. We focused on a subset of this database, based on an ontology[36] that maps Enzyme Commission (EC)[37] numbers, GPCRs, ion channels, and nuclear receptors to MDDR activity classes. Only sets containing five or more ligands were used. Salts and fragments were filtered, ligand protonation was normalized, and duplicate molecules were removed. Of the 688 targets in the MDDR, 97 were excluded as having too few ligands (<5), and another 345 targets were excluded as being non-molecular targets (e.g., the annotation "Anticancer" was not used). This left 246 targets, made up of a total of 65,241 unique ligands, with a median and mean of 124 and 289 ligands per target. The ligand set for methadone and 14 of its analogs was manually populated by querying "methadone" in PubChem Compound (http://pubchem.ncbi.nlm.nih.gov/). Ligand structures for emetine and loperamide were likewise acquired from PubChem Compound. All ligands were represented as SMILES[38] strings.

II. Quality of ligand set annotations

The activity class annotations available from the MDDR do not include explicit ligand-target affinity values and were primarily derived from the patent literature. Any given set may thus contain compounds with a wide range of affinities to the intended target. While Hopkins and colleagues have recently found it useful to restrict the compounds annotated to a particular

target to a limited affinity range,[12] we have found our methods robust to the number of analogs present and the particular identities of which analogs are used. We address this in two experiments, wherein we (1) pre-filter the MDDR for unique chemotypes at 0.90 and 0.85 Tc distances to test robustness against analog redundancy (see **Supplementary Figure 2**), and (2) delete randomly-chosen subsets of the ligand sets to test robustness against the particular choice of analogs present (see **Supplementary Figure 3**). However, as noted by Sheridan et al., 'false inactives' remain a limitation of patent-based databases such as the MDDR, as any given compound may only be tested for one or two of its potential activities.[28]

III. Set comparisons

All pairs of ligands between any two sets were compared by a pair-wise similarity metric, which consists of a descriptor and a similarity criterion. For the similarity descriptor, we computed standard 2D topological Daylight fingerprints[38] using default settings of 2048-bit array lengths and path lengths of 2-7 atoms. The similarity criterion was the widely-used Tanimoto coefficient.[39-41] For set comparisons, all pair-wise Tanimoto coefficients between elements across sets were calculated (**Figure 1**), and those above a threshold were summed, giving a raw score for the two sets. The threshold was chosen so that the resulting statistics best fit an extreme value distribution (below).

IV. Statistical model

A model for the random chemical similarity of the raw scores, motivated by BLAST[22] theory, was developed and empirically fit. We compared 300,000 pairs of molecule sets, randomly populated from the filtered full MDDR, across logarithmic set size intervals in the range of 10 to 1,000 molecules. This range reflected the set sizes we expected to encounter, though the procedure appears robust over any reasonable range of set sizes.

The raw score for each set comparison was plotted against the total number of ligand pairs in the two sets being compared, and was observed to depend linearly on the product of the number of ligands in the two sets (see **Supplementary Figure 1a**). The standard deviation of the raw scores was fit nonlinearly against this product of the set sizes (see **Supplementary Figure 1b, Supplementary Table 5**). Both fits were determined with the SciPy[42] linear least squares optimizer.

Set comparison Z-scores were calculated as a function of the set raw scores, expected raw scores, and standard deviations. The histogram of Z-scores of the random sets conformed to an extreme value distribution (EVD) (see **Supplementary Figure 1c**). This distribution also underlies BLAST comparisons of protein and DNA sequences.[21, 22] The probability of the score being achieved by random chance alone, given the Z-score, was converted to an expectation value (E-value), as further described in the **Supplementary Methods**. The combination of set comparisons with the described statistical model is referred to as the Similarity ensemble approach (SEA). The ability of SEA E-values to correctly discriminate matching MDDR activity classes was tested against three simpler scoring metrics in **Supplementary Figure 4**.

There is no formal justification for choosing a cutoff for the Tc value between ligands. One criterion that had the virtue of consistency was to insist on a Tc value for which the background Z-scores were best fit by an extreme value distribution as in **Supplementary Figure 1c**. We calculated Z-score distributions for all Tc thresholds in the range 0.00 to 0.99, with step size 0.01. For each such distribution, we plotted the normalized chi-square of their best fit to both normal and EVD distributions (see **Supplementary Figure 5**). This led to a Tc threshold of 0.57 (see **Supplementary Table 5**), which is low compared to accepted cutoffs for comparing individual pairs of ligands, emphasizing our different goal here: comparing ligand sets to inform us on the targets.

V. Similarity maps

All annotations in a given database were exhaustively compared against all others, resulting in a matrix of SEA E-values among the ligand sets. The full matrix is available in the online **Supplementary Data**. This matrix defined a strongly-connected graph. In one approach, we filtered the graph by removing all edges with significance less than an E-value cutoff of 1.0; this is a threshold graph. We also constructed a minimum spanning tree over the original strongly-connected graph via Kruskal's algorithm.[43] We refer to this tree as a similarity map. The final images were rendered with Cytoscape.[44]

VI. Difference heat map

Protein sequences for the targets of 193 of the 246 activity classes were obtained, 77 of which were derived from the MDDR-to-EC number mapping provided by Schuffenhauer et al.[36] The remaining 117 sequences were acquired from PubMed Protein searches. The resulting mapping of MDDR activity class to GI number is available in the online **Supplementary Data**. We computed the sequence comparison matrix with PSI-BLAST,[27] as implemented in the blastpgp binary available from NCBI. The maximum final E-value displayed was 1×10^5, with low-complexity region filtering enabled, and a maximum of 10 iterations computed before convergence. **Supplementary Figure 6** shows a heat map of the 193x193 PSI-BLAST matrix, created with matrix2png.[46]

The unfiltered SEA E-value matrix described in **Similarity maps** is shown as a heat map in **Supplementary Figure 7**. This matrix was compared against the sequence-comparison E-value built above by taking the difference of the natural logarithms of each E-value pair. To avoid math range errors, both E-values were first confined within the range of 1×10^{-50} to 1×10^5. A smaller E-value cap would allow for greater resolution of high-end E-values (e.g., 1×10^{-250} vs.

1×10^{-200}), however this would be at the expense of differentiating from insignificant similarity (e.g., 1×10^{-45} vs. 1×10^{5}). As a cutoff of 1×10^{-50} or better appears necessary for reliable transfer,[45] no larger E-value cap was used.

VII. PubChem out-group analysis

All compounds with annotated MeSH (http://www.nlm.nih.gov/mesh/) "Pharmacological Actions" were downloaded from PubChem and filtered as previously described. Any compound already present in the MDDR was removed, resulting in 10,557 unique non-overlapping structures organized into 352 unique annotated "action sets." Of these, 23 action sets could be specifically mapped to a MDDR "activity class," with mean 62 and median 52 compounds per set. These sets were then ranked by SEA E-values against all 246 MDDR activity classes.

VIII. Choice of compounds for novel selectivity prediction

Methadone and 14 analog structures from PubChem Compound were compared as a set against the MDDR to recapitulate known polypharmacology. Instead, novel selectivity was predicted, deemed plausible, and ultimately tested. Subsequently, an automated system was developed to compare individual PubChem Compound molecules with annotated pharmacological actions against the MDDR. All activity class hits resembling known actions were discarded, leaving 30 PubChem compounds with very low (good) expectation values against genuinely unrelated MDDR categories. Among these molecules, we targeted those that we could acquire and actually test, and that looked like sensible members of the novel target to which they were assigned by SEA (i.e., there was a human filter on the compounds before assays were developed and compounds tested). The drugs emetine and loperamide met both criteria. We note that neither compound was present in the MDDR, nor was any close congener. For emetine this reflects the lack of that family of amebecides in the MDDR, whereas loperamide is a non-classical μ-opioid

antagonist whose chemotype happens to be unrepresented among that MDDR ligand set. Thus neither of the classic targets of either drug was found by SEA, simply because the chemical structures were absent or unannotated or both.

IX. Cell lines and functional calcium assay

Radioligand and functional assays were performed as previously detailed using the resources of the National Institute of Mental Health's Psychoactive Drug Screening Program[47, 48] using cloned, human M3-muscarinic receptors expressed in CHO cells also, as previously described.[49] Neurokinin 2 receptor stably expressed in CHO cells[50] and alpha 2a and alpha 2c adrenergic receptors stably expressed in MDCK II cells[51] were carried in DMEM supplemented with 10% Fetal bovine serum (FBS), 1% penicillin-streptomycin, 1 mM sodium pyruvate and 600 µg/ml G418. Cells were plated onto uncoated or poly-L-lysine coated in 96-well plates in DMEM supplemented with 5% dialyzed FBS and 1% penicillin-streptomycin. The following day, media was replaced with 30 µl/well of Calcium Assay Kit Component A Dye (Molecular Devices) dissolved in 28 ml/bottle of assay buffer (2.5 mM probenecid, 20 mM Hepes, and 1x Hanks' Balanced Salt Solution (Gibco) (138 mM NaCl, 5.3 mM KCl, 1.3 mM $CaCl_2$, 0.49 mM $MgCl_2$, 0.41 mM $MgSO_4$, 0.44 mM KH_2PO_4, 0.34 mM Na_2HPO_4) pH 7.4. Plates were incubated in the dye for 1 hr at 37ºC. Drugs predicted to be antagonists were diluted in assay buffer to a concentration of 30 µM and 30 µl of solutions were added to 96-well plates for ~15 minutes prior to reading. Fluorometric imaging was performed using a FlexStation II plate reader (Molecular Devices) reading the plate at 1.5 second intervals for 1 min. After establishing a fluorescent baseline (excitation at 485 nM and emission at 525 nM, using a 515 nM cutoff), 30 µl of agonist were transferred to assay plates at the 20 second time point with reading for another 40 seconds. Peak relative fluorescence units (RFU) were subtracted from baseline RFUs using

SoftMax® Pro (Molecular Devices) and data were then analyzed by non-linear regression to obtain pEC_{50} values using GraphPad Prism version 4.03 (GraphPad Software). Statistical significance between pEC_{50} values obtained from vehicle and predicted antagonist pre-treatment were analyzed by two-tailed t-test ($P < 0.05$) using GraphPad Prism.

1.6 Acknowledgements

I. Author Contributions

JJI, BKS, & MJK developed the ideas for SEA, MJK wrote the SEA algorithms and undertook the calculations reported here, with some assistance from JJI. BLR & PE performed the methadone assays, BNA performed the emetine and loperamide assays, and BKS & MJK wrote the manuscript with editorial review from JJI and BLR.

1.7 Abbreviations

5-HT$_x$	5-Hydroxytryptamine (serotonin) type x
AMPA	α-amino-5-hydroxy-3-methyl-4-isoxazole propionic acid
DHFR	Dihydrofolate reductase
FPGS	Folylpolyglutamate synthetase
GART	Glycinamide ribonucleotide formyltransferase
NMDA	*N*-methyl-D-aspartic acid
SEA	Similarity Ensemble Approach
Tc	Tanimoto coefficient
TS	Thymidylate synthase

1.8 References

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

2. Kroeze, W.K., Kristiansen, K. & Roth, B.L. Molecular biology of serotonin receptors structure and function at the molecular level. *Curr Top Med Chem* **2**, 507-528 (2002).

3. Ebert, B., Andersen, S. & Krogsgaard-Larsen, P. Ketobemidone, methadone and pethidine are non-competitive N-methyl-D-aspartate (NMDA) antagonists in the rat cortex and spinal cord. *Neurosci Lett* **187**, 165-168 (1995).

4. Callahan, R.J., Au, J.D., Paul, M., Liu, C. & Yost, C.S. Functional inhibition by methadone of N-methyl-D-aspartate receptors expressed in Xenopus oocytes: stereospecific and subunit effects. *Anesth Analg* **98**, 653-659, table of contents (2004).

5. Krueger, K.E. Peripheral-type benzodiazepine receptors: a second site of action for benzodiazepines. *Neuropsychopharmacology* **4**, 237-244 (1991).

6. Finlayson, K., Witchel, H.J., McCulloch, J. & Sharkey, J. Acquired QT interval prolongation and HERG: implications for drug discovery and development. *Eur J Pharmacol* **500**, 129-142 (2004).

7. Schreiber, S.L. Small molecules: the missing link in the central dogma. *Nat Chem Biol* **1**, 64-66 (2005).

8. Johnson, M.A. & Maggiora, G.M. Concepts and applications of molecular similarity. (Wiley, New York; 1990).

9. Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J Med Chem* **40**, 1219-1229 (1997).

10. Whittle, M., Gillet, V.J., Willett, P., Alex, A. & Loesel, J. Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients. *J Chem Inf Comput Sci* **44**, 1840-1848 (2004).

11. Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J Med Chem* **48**, 4183-4199 (2005).

12. Paolini, G.V., Shapland, R.H.B., Hoorn, W.P.v., Mason, J.S. & Hopkins, A.L. Global mapping of pharmacological space. *Nature Biotechnology* **24**, In Press (2006).

13. Vieth, M. et al. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim Biophys Acta* **1697**, 243-257 (2004).

14. Izrailev, S. & Farnum, M.A. Enzyme classification by ligand binding. *Proteins* **57**, 711-724 (2004).

15. Bender, A. et al. "Bayes Affinity Fingerprints" Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When Are Multitarget Drugs a Feasible Concept? *J Chem Inf Model*, Web Release Date: 09-Sep-2006 (2006).

16. Nidhi, Glick, M., Davies, J.W. & Jenkins, J.L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* **46**, 1124-1133 (2006).

17. Steindl, T.M., Schuster, D., Laggner, C. & Langer, T. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J Chem Inf Model* **46**, 2146-2157 (2006).

18. Schuffenhauer, A., Floersheim, P., Acklin, P. & Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comput Sci* **43**, 391-405 (2003).

19. Horvath, D. & Jeandenans, C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces-a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J Chem Inf Comput Sci* **43**, 680-690 (2003).

20. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

21. Karlin, S. & Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* **87**, 2264-2268 (1990).

22. Pearson, W.R. Empirical statistical estimates for sequence similarity searches. *J Mol Biol* **276**, 71-84 (1998).

23. Sheridan, R.P. & Miller, M.D. A Method for Visualizing Recurrent Topological Substructures in Sets of Active Molecules. *J Chem Inf Comput Sci* **38**, 915-924 (1998).

24. Bradshaw, J. & Sayle, R.A. Some thoughts on significant similarity and sufficient diversity. *Presented at the 1997 EuroMUG meeting, 7-8 October in Verona, Italy: http://www.daylight.com/meetings/emug97/Bradshaw/Significant_Similarity.html* (1997).

25. Hert, J. et al. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J Chem Inf Comput Sci* **44**, 1177-1185 (2004).

26. 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. *J Chem Inf Model* **46**, 462-470 (2006).

27. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein

database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

28. Sheridan, R.P. & Kearsley, S.K. Why do we need so many chemical similarity search methods? *Drug Discov Today* **7**, 903-911 (2002).

29. Goodman, L.S., Gilman, A., Brunton, L.L., Lazo, J.S. & Parker, K.L. Goodman & Gilman's the pharmacological basis of therapeutics, Edn. 11th. (McGraw-Hill, New York; 2006).

30. Cleves, A.E. & Jain, A.N. Robust ligand-based modeling of the biological targets of known drugs. *J Med Chem* **49**, 2921-2938 (2006).

31. Methadone in DRUGDEX. (Thomson Micromedex, Greenwood Village, Colorado; 2006).

32. de Vos, J.W., Geerlings, P.J., van den Brink, W., Ufkes, J.G. & van Wilgenburg, H. Pharmacokinetics of methadone and its primary metabolite in 20 opiate addicts. *Eur J Clin Pharmacol* **48**, 361-366 (1995).

33. Emetine in DRUGDEX. (Thomson Micromedex, Greenwood Village, Colorado; 2006).

34. Kojima, S., Ikeda, M. & Kamikawa, Y. Loperamide inhibits tachykinin NK3-receptor-triggered serotonin release without affecting NK2-receptor-triggered serotonin release from guinea pig colonic mucosa. *J Pharmacol Sci* **98**, 175-180 (2005).

35. MDL Drug Data Report. (MDL Information Systems Inc, San Leandro, CA).

36. Schuffenhauer, A. et al. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J Chem Inf Comput Sci* **42**, 947-955 (2002).

37. International Union of Biochemistry and Molecular Biology. Nomenclature Committee. & Webb, E.C. Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. (Published for the International Union of Biochemistry and Molecular Biology by Academic Press, San Diego; 1992).

38. James, C., Weininger, D. & Delany, J. Daylight theory manual. (Daylight Chemical Information Systems Inc, Mission Viejo, CA; 1992-2005).

39. Willett, P. Similarity and clustering in chemical information systems. (Research Studies Press; Wiley, Letchworth, Hertfordshire, England; New York; 1987).

40. Brown, R.D. & Martin, Y.C. Use of structure Activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comp Sci* **36**, 572-584 (1996).

41. Chen, X. & Reynolds, C.H. Performance of similarity measures in 2D fragment-based

similarity searching: comparison of structural descriptors and similarity coefficients. *J Chem Inf Comput Sci* **42**, 1407-1414 (2002).

42. Jones, E., Oliphant, T. & Peterson, P. SciPy: Open Source Scientific Tools for Python. (http://www.scipy.org/, 2001).

43. Kruskal, J. On the shortest spanning subtree and the traveling salesman problem. *Proceedings of the American Mathematical Society* **7**, 48-50 (1956).

44. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).

45. Rost, B. Enzyme function less conserved than anticipated. *J Mol Biol* **318**, 595-608 (2002).

46. Pavlidis, P. & Noble, W.S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics* **19**, 295-296 (2003).

47. Roth, B.L. et al. Salvinorin A: a potent naturally occurring nonnitrogenous kappa opioid selective agonist. *Proc Natl Acad Sci U S A* **99**, 11934-11939 (2002).

48. Davies, M.A., Compton-Toth, B.A., Hufeisen, S.J., Meltzer, H.Y. & Roth, B.L. The highly efficacious actions of N-desmethylclozapine at muscarinic receptors are unique and not a common property of either typical or atypical antipsychotic drugs: is M1 agonism a pre-requisite for mimicking clozapine's actions? *Psychopharmacology (Berl)* **178**, 451-460 (2005).

49. Chelala, J.L., Kilani, A., Miller, M.J., Martin, R.J. & Ernsberger, P. Muscarinic receptor binding sites of the M4 subtype in porcine lung parenchyma. *Pharmacol Toxicol* **83**, 200-207 (1998).

50. Takeda, Y. et al. Ligand binding kinetics of substance P and neurokinin A receptors stably expressed in Chinese hamster ovary cells and evidence for differential stimulation of inositol 1,4,5-trisphosphate and cyclic AMP second messenger responses. *J Neurochem* **59**, 740-745 (1992).

51. Wozniak, M. & Limbird, L.E. The three alpha 2-adrenergic receptor subtypes achieve basolateral localization in Madin-Darby canine kidney II cells via different targeting mechanisms. *J Biol Chem* **271**, 5017-5024 (1996).

Gloss to Chapter 2

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" but "That's funny…"

<div align="right">Isaac Asimov, professor of Biochemistry</div>

A typical drug approved in the United States first undergoes fifteen years of lead discovery, medicinal chemistry optimization, cell-based and animal models, preclinical study, Phase I, Phase II, Phase III, and sometimes Phase IV clinical trials.[1] The drug costs in excess of a billion dollars and requires the resources of an industrial behemoth to drive its development and to guide its path down a gauntlet of FDA regulatory steps.[2] With this effort centered on one small molecule, the result should be a drug that is efficacious, safe within acceptable risk, and well-characterized in its actions. Yet in Chapter 1, we found that methadone, emetine, and Imodium all had uncharacterized activity at new and therapeutically relevant targets. The triumph of the unintended is a recurrent theme in the pages that follow. Either we had been extraordinarily lucky in our molecules of inquiry, or such "polypharmacology" among commercial drugs was unexpectedly prevalent.

Chapter 2 is our attempt to quantify how prevalent such drug polypharmacology may be among the drugs that are currently used in humans. Harnessing SEA to this task, we discovered and experimentally validated 23 new drug-to-target associations among commercial drugs. Retrospectively, we also rediscovered hundreds of drug-target associations unknown to our starting databases (but known in others), and over forty drug-target associations unknown in any drug database (but, again, known from another source – manual searches of the literature). This work revealed a persistent theme; it practically seemed difficult to find a drug that did not bind

more than one protein target with at least mid-micromolar affinity. Lest this seem an argument for hopeless drug promiscuity, it is reassuring that these "off-target" associations were not arbitrary and often explained some aspect of each drug's biological story. Thus Paxil discontinuation raises standing heart rate,[3] and we discovered its affinity for the β_1-adrenergic receptor, which is associated with the heart; Rescriptor use can cause severe skin rashes,[4] and we discovered it binds the histamine H_4 receptor, which has been implicated in atopic dermatitis.[5] For Doralese, we found a new target for which its K_i is a full order of magnitude stronger than that for its canonical target; such examples of "off"-target binding may contribute to drug efficacy.

Others have also considered the question of drug polypharmacology;[6-9] Bryan Roth has argued for its role in the efficacy of neurological and psychoactive drugs,[10] and it is clear that among kinase inhibitors such as Gleevec, polypharmacology is far less the exception than the rule.[11] A goal of this chapter, however, was to examine and take the first steps toward quantifying the role of polypharmacology across all of drug-target space. Many of our prospective results concern drugs that bind aminergic GPCRs, but they extend also to ligands of ion channels, transporters, nuclear hormone receptors, and enzymes. Furthermore, we have used SEA to find individual drugs that cross these protein-class boundaries, such as Vadilex, an ion channel inhibitor that was not previously known to bind either the serotonin reuptake transporter or the μ-opioid G-protein coupled receptor.

In the following chapter, we consider several cases of newly-discovered off-target associations and present a global view of such associations. The chapter is somewhat briefer than the others, due to the page requirements of the journal at which it was accepted.

I. References

1. DiMasi, J.A., Hansen, R.W. & Grabowski, H.G. The price of innovation: new estimates of drug development costs. *J Health Econ* **22**, 151-185 (2003).

2. Chong, C.R. & Sullivan, D.J., Jr. New uses for old drugs. *Nature* **448**, 645-646 (2007).

3. Michelson, D. et al. Interruption of selective serotonin reuptake inhibitor treatment. Double-blind, placebo-controlled trial. *Br J Psychiatry* **176**, 363-368 (2000).

4. Scott, L.J. & Perry, C.M. Delavirdine: a review of its use in HIV infection. *Drugs* **60**, 1411-1444 (2000).

5. Dijkstra, D. et al. Human inflammatory dendritic epidermal cells express a functional histamine H4 receptor. *J Invest Dermatol* **128**, 1696-1703 (2008).

6. Hopkins, A.L. Network pharmacology. *Nat Biotechnol* **25**, 1110-1111 (2007).

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

8. Bajorath, J. Computational analysis of ligand relationships within target families. *Curr Opin Chem Biol* **12**, 352-358 (2008).

9. Oprea, T.I., Tropsha, A., Faulon, J.L. & Rintoul, M.D. Systems chemical biology. *Nat Chem Biol* **3**, 447-450 (2007).

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

11. Rix, U. et al. Chemical proteomic profiles of the BCR-ABL inhibitors imatinib, nilotinib, and dasatinib reveal novel kinase and nonkinase targets. *Blood* **110**, 4055-4063 (2007).

Chapter 2:

Predicting new molecular targets for known drugs

Michael J. Keiser[1,2†], Vincent Setola[3†], John J. Irwin[1], Christian Laggner[1], Atheir Abbas[4], Sandra J. Hufeisen[4], Niels H. Jensen[4], Michael B. Kuijer[3], Roberto C. Matos[3], Thuy B. Tran[3], Ryan Whaley[3], Richard A. Glennon[5], Jérôme Hert[1], Kelan L.H. Thomas[1,6], Douglas D. Edwards[1], Brian K. Shoichet[1*], Bryan L. Roth[3,4*]

1 Department of Pharmaceutical Chemistry, University of California San Francisco, 1700 4th St, San Francisco California 94143-2550, USA.

2 Graduate Group in Bioinformatics, University of California San Francisco, 1700 4th St., San Francisco, California 94143-2550, USA.

3 NIMH Psychoactive Drug Screening Program, Department of Pharmacology, University of North Carolina Chapel Hill School of Medicine, Chapel Hill, North Carolina 27759, USA.

4 Department of Pharmacology and Division of Medicinal Chemistry and Natural Products, The University of North Carolina Chapel Hill School of Medicine, Chapel Hill, North Carolina 27759, USA.

5 Department of Medicinal Chemistry, School of Pharmacy, Medical College of Virginia Campus, Virginia Commonwealth University, 410 North 12th St., P. O. Box 980540, Richmond, VA 23298-0540, USA.

6 University of Michigan Health System, 1500 E. Medical Center Drive Ann Arbor, MI 48109, USA.

† Co-first authors

* Corresponding authors (BLR. and BKS)

2.1 Summary

Whereas drugs are intended to be selective, at least some bind to several physiologic targets, explaining both side effects and efficacy. As many drug-target combinations exist, it would be useful to explore possible interactions computationally. Here, we compared 3,665 FDA-approved and investigational drugs against hundreds of targets, defining each target by its ligands. Chemical similarities between drugs and ligand sets predicted thousands of unanticipated associations. Thirty were tested experimentally, including the antagonism of the β_1 receptor by the transporter inhibitor Prozac, the inhibition of the 5-HT transporter by the ion channel drug Vadilex, and antagonism of the histamine H_4 receptor by the enzyme inhibitor Rescriptor. Overall, 23 new drug-target associations were confirmed, five of which were potent (< 100 nM). The physiological relevance of one such, the drug DMT on serotonergic receptors, was confirmed in a knock-out mouse. The chemical similarity approach is systematic and comprehensive, and may suggest side-effects and new indications for many drugs.

2.2 Results and Discussion

The creation of target-specific "magic bullets" has been a therapeutic goal since Ehrlich[1] and a pragmatic criterion in drug design for 30 years. Still, several lines of evidence suggest that drugs may have multiple physiologic targets.[2-5] Psychiatric medications, for instance, notoriously act through multiple molecular targets and this "polypharmacology" is likely therapeutically essential.[6] Recent kinase inhibitors, such as Gleevec and Sutent, though perhaps designed for specificity, modulate multiple targets at physiologic concentrations and, here too, these "off-target" activities may be essential for efficacy.[7,8] Conversely, anti-Parkinsonian drugs such as

Permax and Dostinex activate not only dopamine receptors but also $5\text{-}HT_{2B}$ serotonin receptors, thereby causing valvular heart disease and severely restricting their use.[9]

I. Predicting drug polypharmacology

Drug polypharmacology has inspired efforts to predict and characterize drug-target associations.[10-15] Several groups have used phenotypic and chemical similarities among molecules to identify those with multiple targets,[16,17] and early drug candidates are screened against molecular target panels.[18] To predict new targets for established drugs, Bork and colleagues looked for side-effects shared between two molecules,[19] while Hopkins and colleagues linked targets by drugs that bind to more than one of them.[20] Indeed, using easily accessible associations, one can map 332 targets by the 290 drugs that bind to at least two of them, resulting in a network with 972 connections (**Figure 1a**). It seemed interesting to calculate a related map that predicts new and unanticipated off-target effects.

a

5-HT1
5-HT2
D1,3,4
α
GABA, BzR
DHFR
TXA
β
M1-5
Na+, K+ channels
5-HT3-7, α, β, D2
Opioid μ, κ, δ
P450
NMDA/AMPA
THR
CA I-V
ER, AR
GR
PR
RAR
EP2
A1-3, P2Y
AT
PDE

Network key

Known association (≤ 1 µM)
Off-target predicted by SEA
Protein target
Drug

b

CA I-V
Na+/K+
β-lactamase, PBP
DNA gyrase
5α
AMPA, KAIN, mGluR
AR, PR
GABA, BzR
HDA
ER
A1-2, P2Y
PDE
TRHR
THR
Folate
RAR
HIV, HSV, HBV
Opioid μ, κ, δ
M1-5, MT
S1P
NET, DAT, 5-HTT, 5-HT
5-HT1
AT1-2, PRP
Ca(2+) channel
PPAR-γ
PGP
α
D1-4
P450
β
α
DNA
FTase
LDH
VDR
CB1-2
EP2

Figure 2.1 Drug-target networks, before and after predicting off-targets

(A) Known drug-target network. Each drug (gold) is linked to its known protein targets (cyan) by a gray edge. Each edge denotes a binding affinity of 1 μM or better for that drug to its target. (B) Predicted drug-target network.

Drugs and proteins are linked as per the known drug-target network, with the addition of red edges representing SEA off-target predictions with E-values ≤ 10^{-10}.

Accordingly, we used a statistics-based chemoinformatics approach to predict new off-targets for 878 purchasable FDA-approved small-molecule drugs and 2,787 pharmaceutical compounds. Unlike bioinformatics methods, which might use the sequence or structural similarity among targets, this Similarity Ensemble Approach (SEA)[21] compares targets by the similarity of the ligands that bind to them, expressed as expectation values, adapting the BLAST algorithms[21-23] (other methods such as naïve Bayesian classifiers[23,24] may also be used, see **Supplementary Table 1**). The approach thus captures ligand-based similarities among what would otherwise be considered disparate proteins. The 3,665 drugs were compared against 65,241 ligands organized into 246 targets drawn from the MDDR database,[25] yielding 901,590 drug-target comparisons.

Most drugs had no significant similarities to most ligand sets. However, 6,928 pairs of drugs and ligand sets were similar, with expectation values (E-values) better than 1×10^{-10}. We analyzed these predictions retrospectively against known associations and prospectively for unreported drug polypharmacology.

II. Retrospective tests of drug-target predictions

We first compared the predicted drug-target associations from the MDDR database against reported associations with affinities better than 1 μM in a second database, the World of Molecular Bioactivity (WOMBAT).[26] For instance, the MDDR annotates Azopt (brinzolamide) only as an "antiglaucoma agent," but WOMBAT reports that it has 3 nM affinity for carbonic

anhydrase II. Correspondingly, when screened internally against all MDDR molecular targets, SEA predicted that this drug is related to "Carbonic anhydrase inhibitors" with an E-value of 8.32×10^{-139}. For 184 of the 746 drugs in WOMBAT, the predicted MDDR target agreed with the annotated WOMBAT target with E-values of 1×10^{-10} or better (**Supplementary Table 2**). These predictions recapitulated 19% of the off-targets for these drugs missing from the MDDR. Another 257 drug-target predictions with E-values $\leq 1 \times 10^{-10}$ were unannotated in either database, and may suggest new polypharmacology.
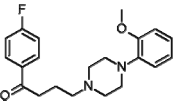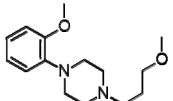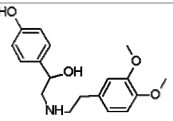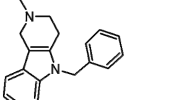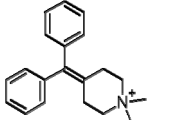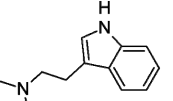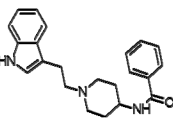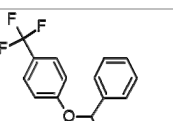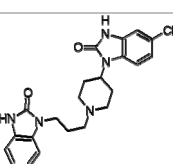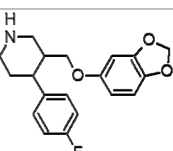
A second retrospective test predicted targets for the 3,665 drugs uncharacterized in either database but known in the literature. Of the 6,928 drug off-targets predicted, we discarded 430 as highly similar by structure to known target ligands, and another 2,666 due to trivial similarities in the reported and predicted activities. This left 3,832 predictions, of which we inspected 184 by literature search and by interrogating other databases. Of these, 42 turned out to be known associations (**Supplementary Table 3**). For instance, when we screened the drug Revanil (lisuride) against the MDDR ligand-target sets, its best E-value was as an α_2 adrenergic antagonist, and when we screened the drug Permax (pergolide) it had an E-value of 8.70×10^{-29} as a 5-HT$_{1D}$ receptor agonist. Consistent with these predictions, Revanil has been reported to bind adrenergic α_2 at 0.055 nM and Permax the 5-HT$_{1D}$ receptor at 13 nM (**Supplementary Table 3**), although neither activity was reported in the MDDR or WOMBAT databases.

III. Prospective tests of new drug-target predictions

For many of these 184 predictions we found no literature precedent. We therefore tested 30 predictions that were experimentally accessible to us. In radioligand competition binding assays, 23 of these (77%) yielded binding affinities (K_i's) less than 15 µM (**Table 1**, **Table 2**, **Supplementary Figure 1**). Fifteen of these 23 were to aminergic GPCRs (**Table 1**) and the

remainder crossed major receptor classification boundaries (**Table 2**). For instance, the α_1 antagonist Doralese was predicted and observed to bind to the dopamine D_4 receptor—both α_1 and D_4 are aminergic GPCRs. Conversely, the HIV-1 reverse transcriptase (enzyme) inhibitor Rescriptor was predicted and observed to bind histamine H_4; this prediction crosses major target boundaries. For several predictions, we tested multiple receptor subtypes because the MDDR left these unspecified; e.g., for a predicted "Adrenergic (α_1) Blocker," we tested the drug at α_{1A}, α_{1B}, and α_{1D} subtypes; we count these as a single target. In total, 14 drugs bound 23 previously unknown targets, with 13 having sub-micromolar and five having sub-100 nM affinities (**Table 1**, **Table 2**). In cases such as Doralese's, the discovered off-target (dopamine D_4) affinity (18 nM) was substantially higher than that for its intended therapeutic target (611 nM for α_{1A} and 226 for α_{1B} adrenergic receptors) (**Figure 2a**).[27]
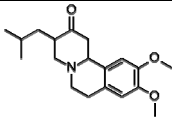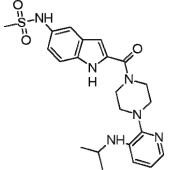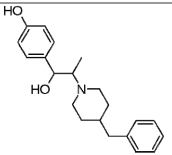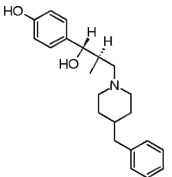
Table 2.1 Prediction and testing of new aminergic GPCR targets for drugs

Drug / MDDR Annotation	E-value	Predicted Target	K_i (nM)	
Sedalande Neuroleptic	8.3×10^{-136}	Adrenergic (α_1) Blocker [†]	α_{1A}	1.2
			α_{1B}	14
			α_{1D}	7
	1.7×10^{-14}	5-HT$_{1D}$ Antagonist		137
Dimetholizine Antihistamine Antihypertensive	1.6×10^{-129}	Adrenergic (α_1) Blocker [†]	α_{1A}	70
			α_{1B}	239
			α_{1D}	171
	2.7×10^{-113}	5-HT$_{1A}$ Antagonist		112
	7.4×10^{-56}	Dopamine (D$_2$) Antagonist		179
Kalgut Cardiotonic	3.1×10^{-79}	Adrenergic (β_3) Agonist		2090
Fabahistin Antihistamine	5.7×10^{-57}	5-HT$_{5A}$ Antagonist		129
Prantal Anticholinergic Antispasmodic	5.5×10^{-32}	δ Agonist		13610
***N,N*-dimethyltryptamine** Serotonergic Hallucinogen	3.1×10^{-21}	5-HT$_{1B}$ Agonist		129
	1.2×10^{-13}	5-HT$_{2A}$ Agonist		127
	1.1×10^{-7}	5-HT$_{5A}$ Antagonist		2135
	5.0×10^{-6}	5-HT$_7$ Modulator		206
Doralese Adrenergic α_1 Blocker Antihypertensive Antimigraine	2.8×10^{-27}	Dopamine (D$_4$) Antagonist		18
Prozac 5-HT Reuptake Inhibitor Antidepressant	3.9×10^{-15}	Adrenergic (β) Blocker [†]	β_1	4385
Motilium Antiemetic Peristaltic Stimulant	4.8×10^{-11}	Adrenergic (α_1) Blocker [†]	α_{1A}	71
			α_{1B}	525
			α_{1D}	705
Paxil 5-HT Reuptake Inhibitor Antidepressive Agent	1.3×10^{-7}	Adrenergic (β) Blocker [†]	β_1	10420

For those targets marked with a dagger (†), the dataset did not specify the receptor subtype, requiring a separate assay for each one. For instance, the MDDR contains an "Adrenergic (α_1) Blocker" target, for which it was necessary to test the α_{1A}, α_{1B}, and α_{1D} subtypes. Note that 5-HT$_{2A}$ is a known target of DMT, but is shown here in gray with its retrospective SEA E-value for comparison purposes.

Table 2.2 Prediction and testing of new cross-boundary targets for drugs

Drug / MDDR Annotation	E-value	Predicted Target	K_i (nM)	
Xenazine VMAT2 (transporter)	1.4×10^{-61}	Adrenergic (α_2) receptor [†] (GPCR)	α_{2A} α_{2C}	959 1318
Rescriptor HIV-1 NNRTI (enzyme)	1.05×10^{-30}	Histamine H$_4$ receptor (GPCR)		5334
Vadilex NMDA Inhibitor (ion channel)	5.14×10^{-13}	μ Opioid receptor (GPCR)		1423
	1.98×10^{-4}	5-HTT; Serotonin transporter (transporter)		77
RO-25-6981 NMDA Inhibitor (ion channel)	1.53×10^{-8}	5-HTT; Serotonin transporter (transporter)		1417
	1.94×10^{-6}	D$_4$ Dopamine receptor (GPCR)		117
	3.61×10^{-6}	NET; Norepinephrine transporter (transporter)		1248
	9.08×10^{-5}	κ Opioid receptor (GPCR)		3128

For the target marked with a dagger (†), the MDDR database did not specify the adrenergic α_2 receptor subtype, requiring a separate assay for each.

How interesting and biologically relevant are these new off-targets? This can be evaluated by the following criteria: when the new targets contribute to the primary activity of the drug, when they may mediate important side effects, or when they are unrelated by sequence,

structure and function to the known, canonical targets. Whereas not all of the newly predicted off-targets fall into these three categories, several fall into each.

IV. Predicted targets as primary mechanism of action

The new targets can improve our understanding of drug action. *N,N*-dimethyltryptamine (DMT) is an endogenous metabolite and a notorious hallucinogen. Recently the molecule was characterized as a σ_1-receptor regulator at micromolar concentrations, an association implicated in its hallucinogenic properties.[28,29] This surprised us because many drugs, including non-hallucinogens, bind promiscuously to the σ_1 receptor with higher affinity than DMT.[30] Also, DMT's hallucinogenic characteristics are consistent with other hallucinogens thought to act through serotonergic receptors, some of which the molecule is known to bind.[31-33] We therefore screened DMT against the 1,133 WOMBAT targets. SEA predicted it to be similar against multiple serotonergic (5-HT) ligand sets, with expectation values ranging from 9.2×10^{-81} to 7.4×10^{-6}. Upon testing in radio-ligand binding assays, we find DMT binds 5-HT$_{1A}$, 5-HT$_{1B}$, 5-HT$_{1D}$, 5-HT$_{2A}$, 5-HT$_{2B}$, 5-HT$_{2C}$, 5-HT$_{5A}$, 5-HT$_6$, and 5-HT$_7$ receptors with affinities from 39 nM to 2.1 µM (**Supplementary Table 4, Supplementary Figure 2**). Of these, three were previously unknown (**Table 1**), and all had affinities for DMT substantially higher than its reported 14.75 µM K_d for σ_1.[28] To further investigate the role of serotonin receptors in DMT-induced hallucination, we turned to a cell-based assay and an animal model that are predictive of hallucinatory actions.[34] Consistent with SEA prediction, we find that DMT not only is a potent partial agonist at 5-HT$_{2A}$ (**Figure 2g**) as has been reported,[31] but also that it induces head twitch response in wild type but not 5-HT$_{2A}$ knockout mice (**Figure 2h**), which is new to this study. The EC$_{50}$ of DMT at 5-HT$_{2A}$ is 100-fold greater than that observed for σ_1.[28] These observations support 5-HT$_{2A}$ as the primary target for DMT's hallucinogenic effects.
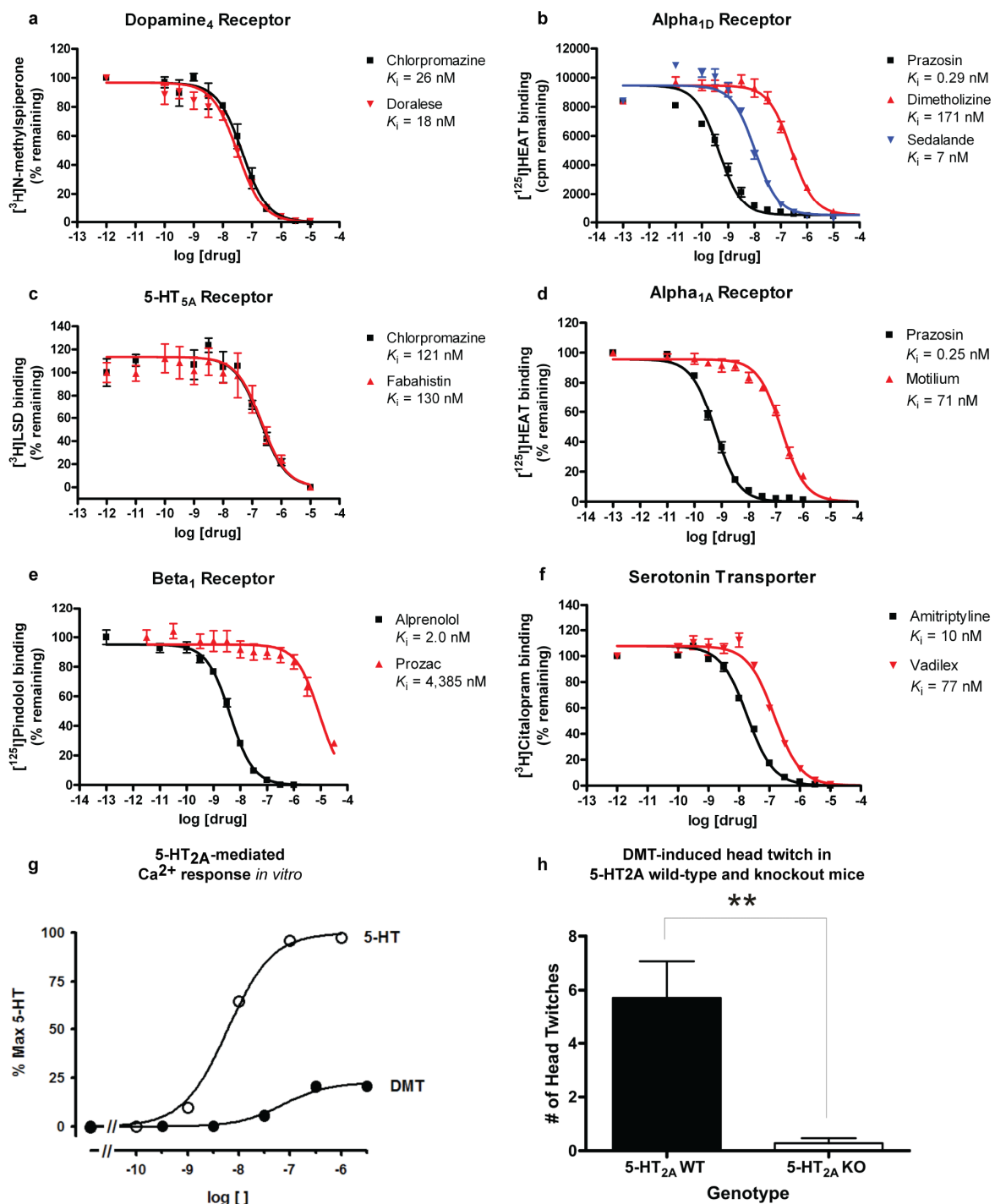
Figure 2.2 Testing new off-target activities

(A-F) Radioligand competition binding assays: (A) Doralese at D_4, (B) Sedalande and Dimetholizine at α_{1D}, (C) Fabahistin at 5-HT$_{5A}$, (D) Motilium at α_{1A}, (E) Prozac at β_1, and (F) Vadilex at the serotonin transporter. (G-H) Investigating 5-HT$_{2A}$ as the target of DMT-

induced hallucination: (G) 5-HT$_{2A}$-mediated Ca^{2+} response was measured after treating HEK 293 cells stably expressing the human 5-HT$_{2A}$ receptor with DMT or 5-HT. DMT's EC$_{50}$ was found to be 118±29 nM (vs. 5-HT's 6.6±0.4 nM baseline, n = 3), with an E$_{max}$ of 23±0.4% (n = 3), confirming that DMT is a potent partial agonist at 5-HT$_{2A}$ receptors. (H) DMT elicited head twitch behavior only in 5-HT$_{2A}$ wild-type mice, confirming that it is a hallucinogenic 5-HT$_{2A}$ agonist. **, p < .01.

Similarly, the new off-targets for Sedalande, a neuroleptic and anxiolytic derived from haloperidol, may illuminate this drug's therapeutic effects. Although subjected to clinical trials with psychiatric patients as far back as the early 1960s,[35] neither its mechanism of action in the central nervous system, nor that of the related Dimetholizine, is well understood. In addition to new activities against α_1 adrenergic receptors (1.2 nM – 239 nM, **Figure 2b**, **Table 1**), Dimetholizine was found to bind the D$_2$ and 5-HT$_{1A}$ receptors and Sedalande to bind the 5-HT$_{1D}$ receptor (**Table 1**, **Supplementary Figure 1**). This likely contributes to the central nervous system activity of both drugs, given the association of the former with anxiety and aggression modulation, and the activity of many antipsychotics against the D$_2$ receptor. We also found analogs of Sedalande that were active against 5-HT$_{1D}$, often at affinities comparable to or greater than those of Sedalande itself (**Supplementary Table 5, Supplementary Figure 3**). This supports the possibility of optimizing these drugs for new indications.

An example of a current drug now being investigated for a new indication is Fabahistin. This drug, used since the 1950s as a symptomatic antihistamine, is now being investigated for Alzheimer's disease. When screened against the 1,133 targets in the WOMBAT database, SEA found an extraordinary similarity to 5-HT$_{5A}$ ligands, with an expectation value of 2.0×10^{-58}. When we measured its binding to the 5-HT$_{5A}$ receptor, Fabahistin had an affinity of 129 nM (**Figure 2c**, **Table 1**). This is an example of a drug whose new, "off-target" affinity is substantially better than that for its canonical H$_1$ receptor target, which is surprisingly low.[36] Its

activity against 5-HT$_{5A}$ and related serotonergic receptors[37] may have implications for

Fabahistin's role as an Alzheimer's disease therapeutic.

V. Off-targets as side-effect mediators

Some of the new off-targets may contribute to a drug's adverse reactions. Motilium is an

antiemetic and dopamine D$_{1/2}$ antagonist that achieves peak plasma concentrations of 2.8 μM[38]

on intravenous administration. This formulation was withdrawn worldwide due to adverse

cardiovascular effects, with the US FDA citing cardiac arrest, arrhythmias, and sudden death.[39]

While Motilium binds the hERG potassium channel with an IC$_{50}$ of 5 μM,[40] the 71 - 705 nM

affinities observed here against α$_{1A}$, α$_{1B}$, and α$_{1D}$ may also contribute to these cardiovascular

effects (**Figure 2d**, **Table 1**, **Supplementary Figure 1**).

Similarly, the micromolar activity against the β-adrenergic receptors of the widely used

selective serotonin reuptake inhibitor (SSRI) antidepressants Prozac and Paxil (**Figure 2e**, **Table

1, Supplementary Figure 1**) may explain several of the adverse effects of these drugs. Abrupt

withdrawal of Paxil raises standing heart rate, a symptom of the SSRI discontinuation

syndrome.[41] This is counterintuitive, as relieving blockade of serotonin reuptake transporters

should reduce the synaptic serotonin available for activation of post-synaptic receptors and such

an effect cannot explain the cardiovascular syndrome.[42] β-blockade by these SSRIs may partially

explain this effect since β-blockers induce a similar rebound tachycardia upon abrupt withdrawal,

due to β receptor up-regulation and sensitization. Despite its higher affinity for β receptors,

Prozac has a longer half-life than Paxil, and its withdrawal does not induce SSRI discontinuation

syndrome. Also, both SSRIs and many β-blockers can induce sexual dysfunction.[43] Since both

serotonergic and adrenergic signaling are involved in sexual response, the finding that both Paxil

and Prozac bind to the β-receptors may explain why they induce greater dysfunction than other SSRIs.

VI. Drug binding across major protein boundaries

Whereas many of the predicted off-targets occur among aminergic GPCRs, a target class for which cross-activity is well-known (see below),[44] four of the drugs bound to targets in entirely different target classes; i.e., those unrelated by sequence or structure (**Table 2**). For instance, the reverse transcriptase (enzyme) inhibitor Rescriptor was predicted and shown to bind to the histamine H_4 receptor, a GPCR. These two targets share no evolutionary history, functional role, or structural similarity whatsoever. As an aside, we note that while Rescriptor's K_i for the H_4 receptor is high at 5.3 μM (**Table 2**, **Supplementary Figure 1**), this is within its steady-state plasma concentration (C_{min} averages 15 μM) and is consistent with the painful rashes associated with Rescriptor use;[45] likewise, H_4 dysregulation has been associated with atopic dermatitis.[46] Similarly, the vesicular monoamine transporter (VMAT) inhibitor[47] Xenazine binds two different GPCRs at sub-micromolar affinity (**Table 2**, **Supplementary Figure 1**). Despite this drug's use over the last 50 years, it has not been reported to bind any GPCR. Finally, the selective ion channel inhibitors Vadilex and RO-25-6981 were predicted and found to bind to GPCRs and to transporters, in addition to their previously known activity against ion channels (**Figure 2f**, **Table 2**, **Supplementary Figure 1**). Whereas these ion channel drugs have known polypharmacology (**Figure 3**), a key point is that the new targets for these four drugs are unrelated to their main therapeutic targets except in the similarity of the ligands that modulate their activities.
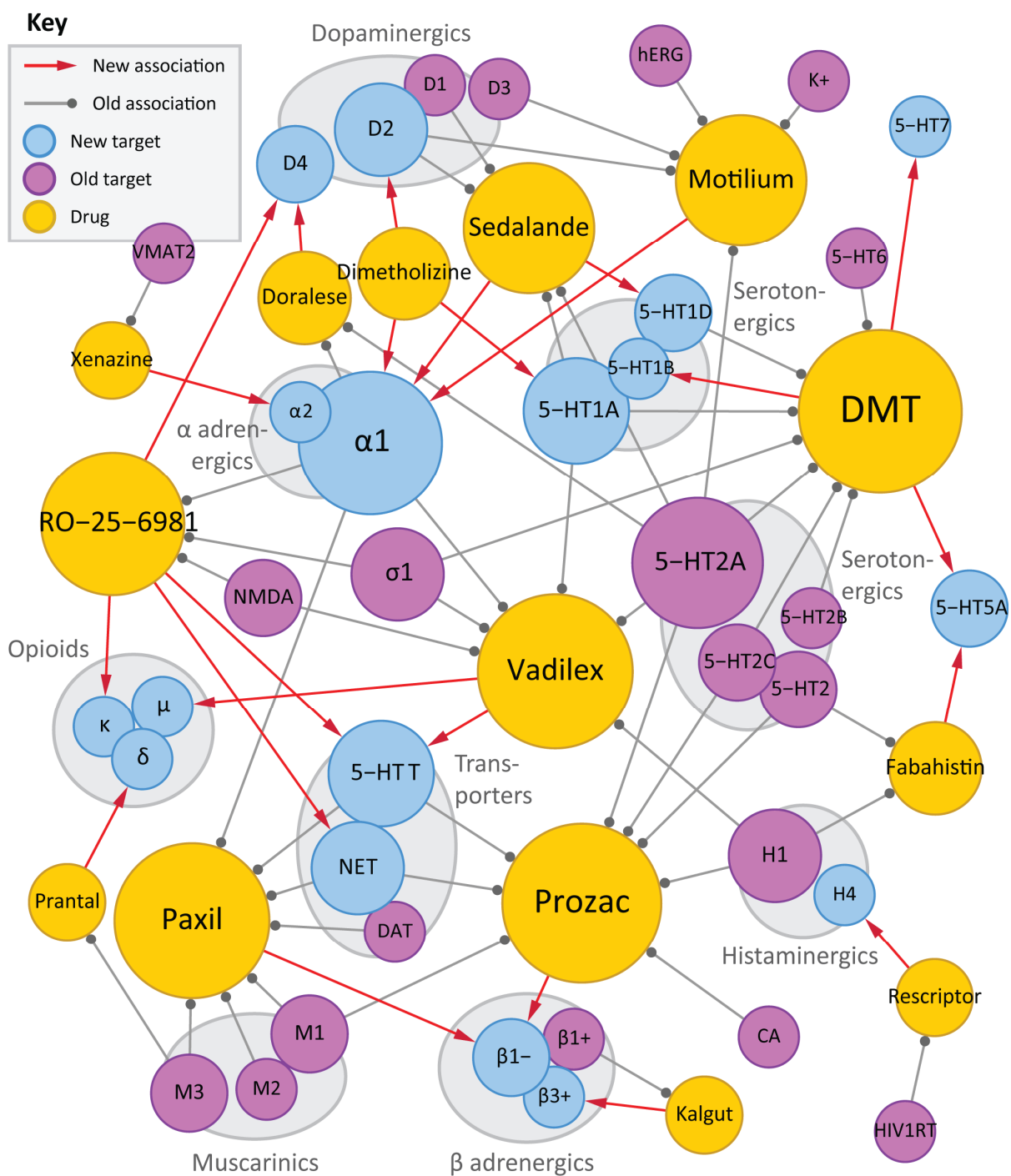
Figure 2.3 Discovered off-targets network

Bipartite network where drugs (gold) are linked by gray edges to their known targets (violet) and by red arrows to their discovered off-targets (cyan). Gray edges denote binding at 1 μM or better, where these affinities are known. Node sizes increase with number of incident edges. Target abbreviations: 5-HTx, serotonin receptor type x; 5-HTT, serotonin transporter; β1+, β1

adrenergic agonist; β1-, β1 adrenergic antagonist; β3+, β3 adrenergic agonist; σ1, σ1-receptor; CA, carbonic anhydrase; DAT, dopamine transporter; HIV1RT, HIV-1 reverse transcriptase; hERG, human Ether-a-go-go Related Gene channel; K+, Potassium channel; NET, norepinephrine transporter; NMDA, N-methyl-D-aspartate receptor; VMAT2, vesicular monoamine transporter 2.

More broadly, the protein target with highest sequence similarity to any of a drug's known targets is rarely predicted by the SEA approach. Rather, the target predicted by ligand similarity is typically well-down in the sequence similarity ranking. Thus for Xenazine, the off-target α_2 adrenergic receptor is 78th most similar to the known target VMAT2 and in fact has no significant similarity at all, with a BLAST E-value of 125 (**Supplementary Table 6**), while for Rescriptor, H_4 is the 167th most similar receptor to HIV RT, and even for Prantal, the aminergic δ-opioid receptor is only 26th most similar to its known muscarinic M_3 target.

VII. Caveats

Certain caveats merit mention. Not all of the new off-targets predicted here would surprise specialists. For instance, Dimetholizine has antihypertensive activity and so its affinity for adrenergic receptors is not wholly unanticipated. Similarly, Kalgut is classified as a "selective β_1 agonist," thought to have little activity on other adrenergic receptors.[48] Whereas the observation that it does bind to the β_3 receptor goes against this classification, structurally this seems easy to credit (**Table 1**, **Supplementary Figure 1**). Indeed, ten of the fourteen drugs reported here are active against aminergic GPCRs (**Figure 3**), and so their cross-activities against other aminergic GPCRs has some precedent.[44] Finally, whereas most of the drugs were active at their predicted off-targets, a third were not; these are examples of the false-positives to which this method is susceptible (**Supplementary Table 7**). Thus, the anxiolytics Valium and Centrax scored well against Cholecystokinin B ligand, the antipsychotic Emilace was predicted to bind 5-HT$_4$, the

anesthetic Duocaine the ϰ-opioid receptor, the antihypertensive Doralese neurokinin receptors, and the narcotic Dromoran and the bradycardic Zatebradine scored well against the D_2 and D_1 receptors. None of these bound their predicted off-targets with affinities better than 10 μM. SEA ignores pharmacophores in its predictions, comparing drugs to ligand sets based on all shared chemical patterns. This is at once a strength, in that it is model-free, and a weakness, in that it may predict activity for drugs that share many features with the ligands of a target, and yet miss a critical chemotype.

VIII. Predicting polypharmacology on a large scale

Notwithstanding these caveats, it is the model-free nature of these predictions that allows a comprehensive exploration of drug-target interactions, most of which remain unexplored. We have focused on a thin slice of pharmacological targets, one dominated by aminergic drugs (**Figure 3**). Stepping back to view the larger space, 364 additional off-targets for 158 drugs are predicted with E-values better than 1×10^{-50}, while 1,853 new off-targets are predicted with E-values better than 1×10^{-10} (**Figure 1b**). This compares to the only 972 off-target activities already annotated in the databases (**Figure 1a**). The Similarity Ensemble Approach and related chemoinformatics methods[16-20] provide tools to explore these associations systematically, both to understand drug effects and explore new opportunities for therapeutic intervention.

2.3 Methods Summary

I. Prediction of off-targets

A collection of 3,665 FDA-approved and investigational drug 2D structures was computationally screened against a panel of over 1,400 protein targets. The drug collection was extracted from the MDL Comprehensive Medicinal Chemistry database. Each target was represented solely by its set of known ligands, which were extracted from three sources of annotated molecules: the MDL Drug Data Report, the World of Molecular Bioactivity (WOMBAT),[26] and the StARlite databases. The structural similarity of each drug to each target's ligand set was quantified as an expectation value (E-value) using the Similarity Ensemble Approach (SEA).[21]

II. Experimental testing

Predicted "off-targets" with strong SEA E-values were evaluated for novelty against orthogonal databases and the literature. Those off-targets without precedent were subjected to radioligand displacement assays using standard techniques[49] at the NIMH Psychoactive Drug Screening Program. The role of 5-HT$_{2A}$ agonism in DMT-induced hallucination was examined in cell-based and in knock-out mouse models.[34] Derivatives of Sedalande were identified in the ZINC[50] database by substructure search, and their affinities for 5-HT$_{1D}$ tested using standard techniques.

III. Drug-target networks and out-group analysis

Comprehensive networks of known drug-target associations (by WOMBAT) and predicted off-targets (by SEA) were constructed. Additionally, SEA off-target predictions were compared to

those derived from Naïve Bayesian classifiers and from PSI-BLAST[21-23] comparisons of a drug's known protein target(s) against the panel of potential protein targets.

2.4 Methods Detail

I. Ligand sets

We extracted ligands from compound databases that annotate molecules by therapeutic or biological category. Multiple ligands in each annotation defined a set of functionally-related molecules. For instance, the 2006.1 MDL Drug Data Report (MDDR) contains 518 compounds annotated as blockers of the α_1 adrenergic receptor, which we grouped together as the "Adrenergic (α_1) Blocker" set.

As reference sources of drug-target ligands, we used three reference databases, as described in **Supplementary Table 8**. The first was a subset of the 2006.1 MDDR, prepared as previously described.[21,23] The second was the 2006.2 World of Molecular Bioactivity (WOMBAT),[26] whose ligands we processed in the same manner as the MDDR. We collapsed targets across species and organized them into inhibitory, activating, and simple binding classes. All compounds with affinity values worse than 1 μM to their targets were removed. This left 1,133 classes built from 191,943 ligands with median and mean of 37 and 169 ligands per target class. The third was StARlite, which we also processed in the same manner as the MDDR. We extracted annotations with the two highest confidence levels (5 and 7), discarded annotations with affinities worse than 1 μM, and organized them into target classes. This yielded 1,158 classes built from 111,329 ligands with median and mean of 43 and 186 ligands per target class.

As a search database of ligand structures for drugs and bioactive molecules, we used the 2004 MDL Comprehensive Medicinal Chemistry database (CMC), which contained 7,517

compounds. All ligands were prepared as above. We then filtered the CMC compounds by vendor availability (as reported in the MDL 2006.3 Available Chemical Directory (ACD), the MDL 2006.1 Screening Compounds Directory, and ZINC[50]), reducing their numbers to 3,665 unique purchasable compounds.

The structures of drugs linking targets in the bipartite drug-target networks (**Figure 1**) were extracted from the 2008 EPA Distributed Structure-Searchable Toxicity (DSSTox) Database at http://www.epa.gov/NCCT/dsstox/, and prepared as above.

II. Ligand activity predictions

We compared each drug individually against each set of reference ligands. All molecules were represented by two topological descriptors: 2048-bit Daylight[51] and 1024-bit folded ECFP_4 fingerprints.[23] We ran the Similarity Ensemble Approach (SEA)[21,23] on each descriptor as a separate screen and chose top-scoring hits (e.g., small E-values) from each screen independently.

As narrated in the main text, our initial SEA screen of 3,665 CMC drugs against 246 MDDR targets yielded 901,590 drug-target comparisons, and this was the screen we subjected to both retrospective literature analysis and prospective empirical testing. During the course of this work, however, we later extended our SEA screen to WOMBAT and StARlite databases, comprising some 4,152,445 and 4,244,070 drug-target comparisons, respectively. We have not as yet made an effort to mine these expanded SEA screens for retrospective literature validation, and instead conducted prospective empirical testing. **Supplementary Table 8** delineates the particular screen (i.e., database) from which each prediction in **Table 1** and **Table 2** is derived.

For comparison of SEA "off-target" predictions against those of naïve Bayesian classifiers (**Supplementary Table 1**), we implemented our own Laplacian-corrected naïve Bayesian classifier with Avidon weighting, as previously described.[23]

III. Drug-target and target-target networks

The drug-target networks in **Figure 1** are bipartite: Along any given path, the nodes alternate between drug targets and drugs. All targets are from WOMBAT, and all drugs from the EPA DSSTox collection. Red edges denote SEA predictions between drug and target nodes, with E-value $\leq 10^{-10}$. Predictions already reported to have affinity ≤ 1 μM in WOMBAT comprise gray edges. All network figures were generated in Cytoscape 2.6.1.[52]

The discovered off-targets network (**Figure 3**) is a bipartite graph linking drugs from **Table 1** and **Table 2** with their targets. Gray edges link drugs to their known targets, and were built by manual literature and database search.

IV. WOMBAT out-group analysis

We mapped 204 MDDR activity classes to WOMBAT targets in two phases. In the first, we mapped 87 MDDR activity classes using EC numbers from the Schuffenhauer ontology[25] to those present in WOMBAT. We then mapped a further 118 GPCR, ligand-gated ion channel, and nuclear hormone receptor MDDR activity classes by supervised sub-phrase matching (**Supplementary Table 9**, **Supplementary Table 10**). While this mapping is not guaranteed to be exhaustive, it is correct to the best of our knowledge.

We extracted all compounds from WOMBAT where the "drug" field was set to "1" (746 unique drugs). We then ran these 746 drugs blindly against the mapped MDDR classes, and discarded all trivial hits (e.g., those where the ligand is already known by the database to be a member of the predicted set). Of those that remained, we then asked how many of the "new" predictions (e.g., not so annotated in the MDDR) were in fact substantiated by existing WOMBAT annotations at affinities ≤ 1 μM.

V. Sequence similarity comparison

For each drug in **Figure 3**, we associated each of its known and newly-discovered targets with human protein sequences in FASTA format from http://www.uniprot.org. We ran these sequences via PSI-BLAST (BLAST version 2.2.14)[21-23] with default parameters against a database built from a subset of the targets in the MDDR, as previously described.[21] For each novel SEA off-target prediction, we reported the best direct PSI-BLAST match (along with its E-value and ranking) from any of that drug's previously known targets, with the predicted off-target (**Supplementary Table 6**). Our goal was to address the question, "Were we to start with the best choice from among a drug's known protein targets, how likely would we be to recapitulate, solely by sequence similarity, each SEA 'off-target' prediction?"

VI. Experimental testing

Radioligand binding and functional assays were performed as previously described.[49,53] Detailed experimental protocols are available on the NIMH PDSP website at http://pdsp.med.unc.edu/UNC-CH%20Protocol%20Book.pdf.

VII. Mice

All experiments were approved by the Institutional Animal Care and Use Committee at the University of North Carolina, Chapel Hill. Mice were housed under standard conditions – 12 hour light/dark cycle and food and water ad libitum.

VIII. Head Twitch

Littermate pairs of 5-HT$_{2A}$ wild type and knockout mice were pretreated for two hours with 75 mg/kg pargyline, i.p., prepared in sterile saline (.9% NaCl) (P8013, Sigma-Aldrich, St. Louis, MO). Mice were then injected with sterile saline or 1.0 mg/kg DMT, i.p., prepared in sterile

saline and moved to a new cage. Head twitch behavior, which consists of a rapid, rotational flick

of the head about the axis of the neck, was counted over 15 minutes. We have determined that

trained observers count the same number of head twitches whether blinded or unblinded to

genotype (data not shown). We confirmed that this was the case with three littermate pairs, and

the rest of the studies were performed by one unblinded observer.[34]

2.5 Acknowledgements

I. Author contributions

B.K.S., J.J.I, and M.J.K. developed the ideas for SEA. M.J.K. wrote the SEA algorithms,

undertook the calculations, and identified the off-targets reported here, typically vetted with J.J.I.

and B.K.S., unless otherwise noted below. M.J.K. wrote the Naïve Bayesian classifier algorithms

with assistance from J.H. With assistance from B.K.S. and J.J.I., C.L. identified off-targets for

Fabahistin, K.L.H.T. identified off-targets for Prozac and Paxil, and D.D.E. identified the off-

target for Rescriptor. V.S. and B.L.R. designed empirical tests of the predictions, analyzed and

interpreted data, and performed experiments. T.B.T., R.W., R.C.M., A.A., N.H.J., and M.B.K. performed empirical testing of the predictions. S.J.H. and R.A.G. generated materials for the experiments. M.J.K. and B.K.S. wrote the manuscript with contributions and review from B.L.R. and V.S. All authors discussed the results and commented on the manuscript.

II. Author information

The authors M.J.K., J.J.I., and B.K.S. declare competing financial interests. Correspondence and requests for materials should be addressed to B.L.R. (bryan_roth@med.unc.edu) or B.K.S. (shoichet@cgl.ucsf.edu).

2.6 References

1   Ehrlich, P. The Theory and Practice of Chemotherapy. *Folia Serologica* **7**, 697-714 (1911).

2   Peterson, R. T. Chemical biology and the limits of reductionism. *Nat Chem Biol* **4**, 635-638 (2008).

3   Nobeli, I., Favia, A. D. & Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat Biotechnol* **27**, 157-167 (2009).

4   Marona-Lewicka, D. & Nichols, D. E. Further evidence that the delayed temporal dopaminergic effects of LSD are mediated by a mechanism different than the first temporal phase of action. *Pharmacol Biochem Behav* **87**, 453-461 (2007).

5   Marona-Lewicka, D. & Nichols, D. E. WAY 100635 produces discriminative stimulus effects in rats mediated by dopamine D(4) receptor activation. *Behav Pharmacol* **20**, 114-118 (2009).

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

7   Rix, U. *et al.* Chemical proteomic profiles of the BCR-ABL inhibitors imatinib, nilotinib, and dasatinib reveal novel kinase and nonkinase targets. *Blood* **110**, 4055-4063 (2007).

8   Hopkins, A. L. Network pharmacology. *Nat Biotechnol* **25**, 1110-1111 (2007).

9   Roth, B. L. Drugs and valvular heart disease. *The New England journal of medicine* **356**, 6-9 (2007).

10 Bajorath, J. Computational analysis of ligand relationships within target families. *Curr Opin Chem Biol* **12**, 352-358 (2008).

11 Oprea, T. I., Tropsha, A., Faulon, J. L. & Rintoul, M. D. Systems chemical biology. *Nat Chem Biol* **3**, 447-450 (2007).

12 Newman, D. J. Natural products as leads to potential drugs: an old process or the new hope for drug discovery? *J Med Chem* **51**, 2589-2599 (2008).

13 Siegel, M. G. & Vieth, M. Drugs in other drugs: a new look at drugs as fragments. *Drug Discov Today* **12**, 71-79 (2007).

14 Miller, J. R. *et al.* A class of selective antibacterials derived from a protein kinase inhibitor pharmacophore. *Proc Natl Acad Sci U S A* **106**, 1737-1742 (2009).

15 Walsh, C. T. & Fischbach, M. A. Repurposing libraries of eukaryotic protein kinase inhibitors for antibiotic discovery. *Proc Natl Acad Sci U S A* **106**, 1689-1690 (2009).

16 Young, D. W. *et al.* Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat Chem Biol* **4**, 59-68 (2008).

17 Wagner, B. K. *et al.* Large-scale chemical dissection of mitochondrial function. *Nat Biotechnol* **26**, 343-351 (2008).

18 Krejsa, C. M. *et al.* Predicting ADME properties and side effects: the BioPrint approach. *Curr Opin Drug Discov Devel* **6**, 470-480 (2003).

19 Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263-266 (2008).

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

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

22 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

23 Hert, J., Keiser, M. J., Irwin, J. J., Oprea, T. I. & Shoichet, B. K. Quantifying the relationships among drug classes. *J Chem Inf Model* **48**, 755-765 (2008).

24 Nigsch, F., Bender, A., Jenkins, J. L. & Mitchell, J. B. Ligand-target prediction using Winnow and naive Bayesian algorithms and the implications of overall performance statistics. *J Chem Inf Model* **48**, 2313-2325 (2008).

25 Schuffenhauer, A. *et al.* An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J Chem Inf Comput Sci* **42**, 947-955 (2002).

26 Oprea, T. I. *Chemoinformatics in drug discovery*. (Wiley-VCH, 2005).

27 Lomasney, J. W. *et al.* Molecular cloning and expression of the cDNA for the alpha 1A-adrenergic receptor. The gene for which is located on human chromosome 5. *J Biol Chem* **266**, 6365-6369 (1991).

28 Fontanilla, D. *et al.* The hallucinogen N,N-dimethyltryptamine (DMT) is an endogenous sigma-1 receptor regulator. *Science* **323**, 934-937 (2009).

29 Su, T. P., Hayashi, T. & Vaupel, D. B. When the endogenous hallucinogenic trace amine N,N-dimethyltryptamine meets the sigma-1 receptor. *Sci Signal* **2**, pe12 (2009).

30 Roth, B. L., Kroeze, W. K., Patel, S. & Lopez, E. The Multiplicity of Serotonin Receptors: Uselessly diverse molecules or an embarrasment of riches? *The Neuroscientist* **6**, 252-262 (2000).

31 Smith, R. L., Canton, H., Barrett, R. J. & Sanders-Bush, E. Agonist properties of N,N-dimethyltryptamine at serotonin 5-HT2A and 5-HT2C receptors. *Pharmacol Biochem Behav* **61**, 323-330 (1998).

32 Kohen, R. *et al.* Cloning, characterization, and chromosomal localization of a human 5-HT6 serotonin receptor. *J Neurochem* **66**, 47-56 (1996).

33 Pierce, P. A. & Peroutka, S. J. Hallucinogenic drug interactions with neurotransmitter receptor binding sites in human cortex. *Psychopharmacology (Berl)* **97**, 118-122 (1989).

34 Abbas, A. I. *et al.* PSD-95 is essential for hallucinogen and atypical antipsychotic drug actions at serotonin receptors. *J Neurosci* **29**, 7124-7136 (2009).

35 Kurland, A. A., Mc, C. K. & Michaux, W. W. Clinical trial of haloanisone (R-2028) with hospitalized psychiatric patients. *J New Drugs* **2**, 352-360 (1962).

36 Gankina, E. M. *et al.* [Effect of some antihistamine preparations on binding of3H-mepyramine and3H-cimetidine to histamine receptors in rat brain]. *Khimiko-farmatsevticheskii Zhurnal* **26**, 9-11 (1992).

37 Gankina, E. M. *et al.* [The effect of antihistaminic preparations on the binding of labelled mepyramine, ketanserin and quinuclidinyl benzilate in the rat brain]. *Eksp Klin Farmakol* **56**, 22-24 (1993).

38 Heykants, J. *et al.* On the pharmacokinetics of domperidone in animals and man. IV. The pharmacokinetics of intravenous domperidone and its bioavailability in man following intramuscular, oral and rectal administration. *Eur J Drug Metab Pharmacokinet* **6**, 61-70 (1981).

39 FDA Warns Against Women Using Unapproved Drug, Domperidone, to Increase Milk Production. *US Food and Drug Administration Talk Paper* (2004).

40 Stork, D. *et al.* State dependent dissociation of HERG channel inhibitors. *Br J Pharmacol* **151**, 1368-1376 (2007).

41 Michelson, D. *et al.* Interruption of selective serotonin reuptake inhibitor treatment. Double-blind, placebo-controlled trial. *Br J Psychiatry* **176**, 363-368 (2000).

42 Berger, M., Gray, J. A. & Roth, B. L. The Extended Pharmacology of Serotonin. *Annual Reviews in Medicine* **60**, 355-366 (2009).

43 Waldinger, M. D., Hengeveld, M. W., Zwinderman, A. H. & Olivier, B. Effect of SSRI antidepressants on ejaculation: a double-blind, randomized, placebo-controlled study with fluoxetine, fluvoxamine, paroxetine, and sertraline. *J Clin Psychopharmacol* **18**, 274-281 (1998).

44 Peters, J. U., Schnider, P., Mattei, P. & Kansy, M. Pharmacological promiscuity: dependence on compound properties and target specificity in a set of recent Roche compounds. *ChemMedChem* **4**, 680-686 (2009).

45 Scott, L. J. & Perry, C. M. Delavirdine: a review of its use in HIV infection. *Drugs* **60**, 1411-1444 (2000).

46 Dijkstra, D. *et al.* Human inflammatory dendritic epidermal cells express a functional histamine H4 receptor. *J Invest Dermatol* **128**, 1696-1703 (2008).

47 Mehvar, R., Jamali, F., Watson, M. W. & Skelton, D. Pharmacokinetics of tetrabenazine and its major metabolite in man and rat. Bioavailability and dose dependency studies. *Drug Metab Dispos* **15**, 250-255 (1987).

48 Inamasu, M., Totsuka, T., Ikeo, T., Nagao, T. & Takeyama, S. Beta 1-adrenergic selectivity of the new cardiotonic agent denopamine in its stimulating effects on adenylate cyclase. *Biochem Pharmacol* **36**, 1947-1954 (1987).

49 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 (2008).

50 Irwin, J. J. & Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**, 177-182 (2005).

51 James, C., Weininger, D. & Delany, J. *Daylight theory manual.* (Daylight Chemical Information Systems Inc, 1992-2005).

52 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).

53 Roth, B. L. *et al.* Salvinorin A: a potent naturally occurring nonnitrogenous kappa opioid selective agonist. *Proc Natl Acad Sci U S A* **99**, 11934-11939 (2002).

Gloss to Chapter 3

One characteristic of small molecules that may not have been emphasized in prior chapters is precisely that they are not all drugs. The debate over the hallucinogen *N,N*-dimethyltryptamine's endogenous role in the human body, touched on in Chapter 2, is a case in which we indirectly address this realm. But what of venturing there directly? Endogenous signaling molecules course through our cells and swarm the extra-cellular matrix; cycles of energy exchange, amino acid biosynthesis, and protein salvage are driven by the turnover of small molecules down branching metabolic pathways. Instead of grouping drugs and drug-like ligands by their trans-membrane receptors or channels, what if we were to instead group small-molecule substrates, cofactors, and products by the reactions in which they participated? Whereas this would narrow our scope to core metabolism, recent genomics efforts have provided both curated and putative metabolic reaction and pathway molecule data for hundreds of organisms.

In Chapter 3, we mapped drug space to metabolic space using the Similarity Ensemble Approach. We primarily investigated within the interface between drug-like and metabolic molecules, from which several conclusions emerged. For the first, we found that we could identify meaningful similarity patterns between sets comprised of drug-like ligands and those of metabolites. While a necessary basis for this work, this was not a foregone conclusion—especially as many metabolic reactions operate on molecules of typically smaller size than drugs. Secondly, we could identify reactions and pathway regions within an organism wherein the metabolites were most similar to the ligands of known drug targets (drug "effect space"), such as histidine biosynthesis in methicillin-resistant *Staphylococcus aureus*. Against pathogenic species,

these regions may be useful starting places for drug discovery efforts based on current drug chemotypes, but where they occur in human metabolism, these regions may suggest toxicity arising from off-target action at these essential enzymes. Chapter 3 thus concludes with consideration of "differential" effect spaces, which may help identify metabolic targets in pathogens that are both absent in humans and amenable to existing drugs.

The remainder of this introduction provides a brief overview of our early observations of chemical similarity patterns within metabolic space.

I. Compact metabolic space

The most striking feature of the first SEA metabolic networks was the presence in each of a few massive hubs, formed by single reactions that demonstrated extraordinarily high similarity to a whole host of others (**Figure 3.i.b**). These "hub" reactions were not particularly central to metabolic pathways, but rather consisted primarily of molecules common to many metabolic reactions, typically cofactors. These so-called "common carriers," such as ATP and GTP, while indeed present in many reactions and absolutely essential to their function, proved uninformative for differentiating *among* reactions. Just as BLAST filters out the "noise" of statistically overrepresented coil-coil regions during protein sequence alignment, we found that it was common practice to filter out common carriers in metabolism; we likewise follow this convention in Chapter 3.

a Comparing drug vs. metabolic spaces

b *E. coli* metabolic space

1 ADENYL-KIN-RXN:
Consists of ATP,
ADP, & AMP

c *E. coli* metabolic space
sans common carriers

d *S. cerevisiae* metabolic space
sans common carriers

1 Transketolase
2 Asparagine synthetase
(glutamine-hydrolysing)
3 Acetolate synthase

Figure 3.i Preliminary metabolic similarity analyses

(a) Comparative plot of uncorrected raw SEA scores from databases of drug-like ligands (MDDR and ZINC leadlike subset)[1] compared to those of metabolites (EcoCyc, YeastCyc, and KEGG).[2, 3] The metabolites, even when corrected for common carriers, are 1-2 orders of magnitude more similar to each other on average than are the drug-like ligands. Additionally, as is qualitatively apparent from this plot, each individual metabolite collection's raw scores demonstrate smaller standard deviation than those from drug-like collections. (b) Initial SEA metabolic network, wherein each node represents a reaction as comprised of its substrates, cofactors, and products; each edge is a significant SEA E-value present between reactions. These networks are Floyd-Warshall[4, 5] graphs, which may be thought of as the consensus network built from successive overlays of all the shortest paths from any given

reaction to another across the strongly-connected network built from the full matrix of SEA similarities among all reactions. This network was built from molecule data extracted from EcoCyc. (c-d) SEA metabolic networks built from reaction sets with the common carriers removed. While still more internally similar than drug networks (as demonstrated in panel (a)), these metabolic networks lack the massive hubs of (b). It is of note, however, that where hubs are present, they appear to be present across species; in this case, *E. coli* (from EcoCyc) and *S. cerevisiae* (from YeastCyc).

Even after the removal of common carriers (**Figure 3.i.c-d**), we found that metabolic space remained both "smaller" and "denser" than drug space. This is qualitatively apparent from **Figure 3.i.a**, wherein the inter-molecule similarity scores within metabolic collections (each represented by its own color) are both higher (therefore at smaller distances from each other) and narrower (therefore denser). **Table 3.i** quantifies these differences, showing that drugs yield SEA chemical backgrounds with higher mean similarity among ligands and lower standard deviation exponents than metabolism does—again suggesting that metabolites sample less space, but do so more densely. This may make intuitive sense, as the cost of sampling new chemotypes is presumably greater for an organism, which must provide and subsequently maintain the enzymatic machinery to do so, than it is for the medicinal chemist, whose choices are driven by other concerns.

Table 3.i Internal chemical similarity patterns of metabolic vs. drug collections

Type	Collection	μ	σ	σ′
Drug	MDDR	9.60×10^{-4}	0.014	0.61
	ZINC	7.60×10^{-4}	0.012	0.58
Metabolic (sans common carriers)	KEGG	0.049	0.14	0.67
	EcoCyc (-)	0.064	0.16	0.67
	YeastCyc (-)	0.074	0.19	0.68
Metabolic (with common carriers)	YeastCyc	0.14	0.34	0.69
	EcoCyc	0.14	0.35	0.69

The μ, σ, and σ' columns are with respect to the random chemical backgrounds that SEA automatically calculates for each molecule collection. For each background, SEA fits an equation of the form "$y = \mu \times x + b$" to determine the mean chemical similarity expected at any particular value x, where x corresponds to the total number of random molecules being compared, e.g., the x-axis of **Figure 3.i.a**. Similarly, SEA fits the standard deviation of the random chemical similarity expected at x to an equation of the form "$y = \sigma \times x^{\sigma'} + b$" (note that the exponent of x is no longer 1, but rather σ'). For a more detailed treatment of SEA random background calculations, see Appendix B.

II. References

1. Irwin, J.J. & Shoichet, B.K. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**, 177-182 (2005).

2. Caspi, R. et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **36**, D623-631 (2008).

3. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

4. Floyd, R.W. Algorithm 97: Shortest Path. *Communications of the ACM* **5**, 345 (1962).

5. Warshall, S. A theorem on Boolean matrices. *Journal of the ACM* **9**, 11-12 (1962).

Chapter 3:

A mapping of drug space from the viewpoint of small molecule metabolism

James Corey Adams[1†], Michael J. Keiser[2†], Li Basuino[3], Henry F. Chambers[3], Deok-Sun Lee[4,5,6], Olaf G. Wiest[7], Patricia C. Babbitt[8*]

1 Graduate Program in Pharmaceutical Sciences and Pharmacogenomics, University of California, San Francisco, CA 94158-2550

2 Graduate Program in Bioinformatics, University of California, San Francisco, CA 94158-2517

3 San Francisco General Hospital, University of California San Francisco, San Francisco CA 94110

4 Center for Complex Network Research and Departments of Physics, Biology, and Computer Science, Northeastern University, Boston, MA 02115

5 Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115

6 Department of Natural Medical Sciences, Inha University, Incheon 402-751, Korea

7 Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, 46556

8 Departments of Bioengineering and Therapeutic Sciences, Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California, San Francisco, CA 94158-2330

† Co-first authors

* Corresponding author

3.1 Abstract

Small molecule drugs target many core metabolic enzymes in humans and pathogens, often mimicking endogenous ligands. The effects may be therapeutic or toxic, but are frequently unexpected. A large-scale mapping of the intersection between drugs and metabolism is needed to better guide drug discovery. To map the intersection between drugs and metabolism, we have grouped drugs and metabolites by their associated targets and enzymes using ligand-based set signatures created to quantify their degree of similarity in chemical space. The results reveal the chemical space that has been explored for metabolic targets, where successful drugs have been found, and what novel territory remains. To aid other researchers in their drug discovery efforts, we have created an online resource of interactive maps linking drugs to metabolism. These maps predict the "effect space" comprising likely target enzymes for each of the 246 MDDR drug classes in humans. The online resource also provides species-specific interactive drug-metabolism maps for each of the 385 model organisms and pathogens in the BioCyc database collection. Chemical similarity links between drugs and metabolites predict potential toxicity, suggest routes of metabolism, and reveal drug polypharmacology. The metabolic maps enable interactive navigation of the vast biological data on potential metabolic drug targets and the drug chemistry currently available to prosecute those targets. Thus, this work provides a large-scale approach to ligand-based prediction of drug action in small molecule metabolism.

3.2 Author Summary

All humans, plants, and animals use enzymes to metabolize food for energy, build and maintain the body, and get rid of toxins. Drugs used to clear infections or cure cancer often target enzymes in bacteria or cancer cells, but the drugs can interfere with the proper function of

human enzymes as well. Recent studies have mapped drugs to enzymes and many other targets in humans and other organisms, but have not focused on metabolism. In this study, we present a new method to predict what enzymes drugs might affect based on the chemical similarity between classes of drugs and the natural chemicals used by enzymes. We have applied the method to 246 known drug classes and a collection of 385 organisms (including 65 National Institutes of Health Priority Pathogens) to create maps of potential drug action in metabolism. We also show how the predicted connections can be used to find new ways to kill pathogens and to avoid unintentionally interfering with human enzymes.

3.3 Introduction

Drug developers have long mined small molecule metabolism for new drug targets and chemical strategies for inhibition. The approach leverages the "chemical similarity principle"[1] which states that similar molecules likely have similar properties. Applied to small molecule metabolism, this principle has motivated the search for enzyme inhibitors chemically similar to their endogenous substrates. The approach has yielded many successes, including antimetabolites such as the folate derivatives used in cancer therapy and the nucleoside analog pro-drugs used for antiviral therapy. However, drug discovery efforts also frequently falter due to unacceptable metabolic side-effect profiles or incomplete genomic information for poorly characterized pathogens.[2-4]

With the recent availability of large datasets of drugs and drug-like molecules, computational profiling of small molecules has been performed to create global maps of pharmacological activity. This in turn provides a larger context for evaluation of metabolic targets. For example, Paolini et al.[5] identified 727 human drug targets associated with ligands exhibiting potency at concentrations below 10 μM, thereby creating a polypharmacology interaction network organized by the similarity between ligand binding profiles. Keiser et al.[6]

organized known drug targets into biologically sensible clusters based solely upon the bond topology of 65,000 biologically active ligands. The results revealed new and unexpected pharmacological relationships, three of which involved GPCRs and their predicted ligands that were subsequently confirmed *in vitro*. Cleves et al.[7] also rationalized several known drug side effects and drug-drug interactions based upon three-dimensional modeling of 979 approved drugs. However, despite the clear rationale and past successes in applying ligand-based approaches to drug discovery, global mapping between drugs and small molecule metabolism, the goal of this study, has been hindered by both methodological challenges and incomplete genomic information. The relatively recent availability of metabolomes for numerous organisms allows a fresh look on a large scale.[8-13]

In this work, we link the chemistry of drugs to the chemistry of small molecule metabolites to investigate the intersection between small molecule metabolism and drugs. The Similarity Ensemble Approach (SEA)[6] was used to link metabolic reactions and drug classes by their chemical similarity, measured by comparing bond topology patterns between sets of molecules. Two types of molecule sets are used in this work. The first comprises drug-like molecules known to act at a specific protein target, and the second comprises the known substrates and products of an enzymatic reaction. While this approach is complementary to target and disease focused methods,[5, 14-23] neither protein structure nor sequence information is used in the comparisons. Thus, these links provide an orthogonal view of metabolism based only upon the chemical similarity between existing drug classes and endogenous metabolites.

To provide the results in the context of metabolism, drug "effect-space" maps were also created. For each of the 246 drug classes investigated in this work, effect-space maps enable visualization of the chemical similarities between drugs and metabolites painted onto human metabolic pathways, allowing a unique assessment of potential drug action in humans. In

addition, to aid target discovery in pathogens, 385 species-specific effect-space maps were created to show the predicted effect-space of currently marketed drugs, painted onto metabolic pathways representing target reactions in model organisms and pathogens. Examples of these maps are provided below and their applications in predicting drug action, toxicity, and routes of metabolism are discussed. To enable facile exploration of the drug-metabolite links established by this analysis, interactive versions of both sets of maps are available at http://sea.docking.org/metabolism.

Finally, using methicillin-resistant *Staphylococcus aureus* (MRSA), a major pathogen causing both hospital- and community-acquired infections that is resistant to at least one of the antibiotics most commonly used for treatment[24-28] as an example, we show by retrospective analysis the use of species-specific maps for discovery and evaluation of drug targets. This also illustrates how additional types of biological information can be incorporated to enhance the value of these analyses.

3.4 Results

I. Drug-metabolite links reproduce known drug-target interactions

To evaluate the chemical similarity between drug classes and metabolic reactions, links between sets of metabolic ligands and sets of drugs were generated according to SEA (**Figure 1**).[6] The similarity metric consists of a descriptor, represented by standard two-dimensional topological fingerprints, and a similarity criterion, the Tanimoto coefficient (Tc). Expectation (E) values were calculated for each set pair by comparing the raw scores to a background distribution generated using sets of randomly selected molecules (see **Methods** for further details). To represent metabolic ligand sets, the MetaCyc database, which includes enzymes from more than

900 different organisms catalyzing over 6,000 reactions, was used.[12] The substrates and products of each enzymatic reaction were combined to form a reaction set, each of which was required to contain at least two unique compounds (**Datasets S1** and **S2**). Ubiquitous molecules called common carriers, which frequently play critical roles in reaction chemistry but do not distinguish the function of a specific enzyme, were removed, leaving a total of 5,056 reactions involving 4,998 unique compounds. To represent drugs, a subset of 246 targets of the MDL Drug Data Report (MDDR) collection, which annotates ligands according to the targets they modulate, was used.[29] These sets contain 65,241 unique ligands with a median and mean of 124 and 289 ligands per target, respectively. Overall, 246 drug versus 5,056 reaction set comparisons involving 1.39×10^9 pairwise comparisons were made.
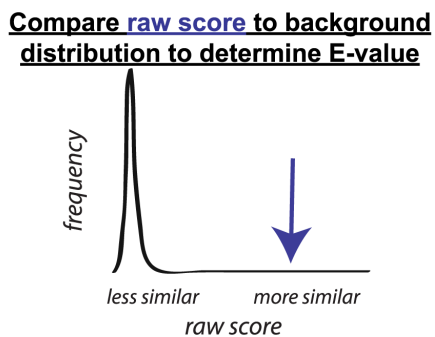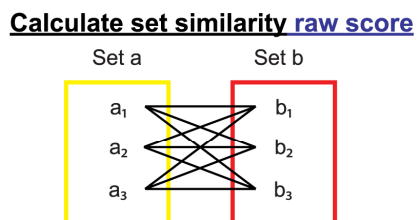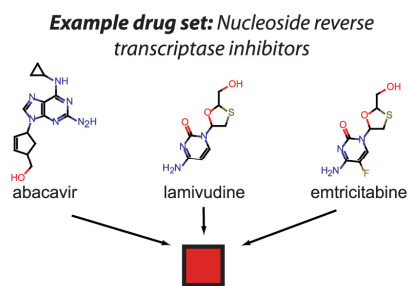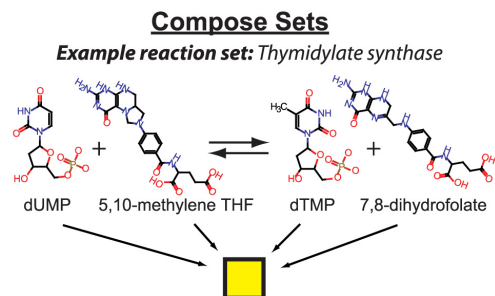
Figure 3.1 Similarity Ensemble Approach (SEA)

SEA compares groups of ligands based upon bond topology. Example ligand sets include the thymidylate synthase reaction set, composed of the reaction substrates and products, and the nucleotide reverse transcriptase inhibitor (NRTI) drug set, which includes known

inhibitors of the nucleoside reverse transcriptase enzyme. Fingerprints representing the bond topology of each molecule are generated. Raw scores between sets are calculated based upon Tanimoto coefficients between fingerprints for all molecule pairs. Finally, the raw scores are compared to a background distribution to determine the expectation value (E) representing the chemical similarity between sets. See **Methods** for further details.

Although drugs and metabolites typically differ in their physiochemical properties, significant and specific similarity links nonetheless emerged. Using SEA at an expectation value cutoff of $E = 1.0 \times 10^{-10}$, a previously reported cutoff for significance,[6] 54% (132 of 246) of drug sets link to an average of 43.7 (median = 10) or 0.9% of metabolic reactions. None of the remaining 46% (114 of 246) of drug sets link to any metabolic reaction sets. For instance, while the α-glucosidase drug set links to the α-glucosidase reaction ($E = 1.00 \times 10^{-51}$), the thrombin inhibitor drug set does not link significantly with any metabolic reaction. The thrombin inhibitor drug set targets the serine protease thrombin, which does not participate in small molecule metabolism, but rather plays a role in the coagulation signaling cascade. Likewise, 40% (2,044 of 5,056) of metabolic reactions hit an average of 2.8 (median = 2) or 1.1% of drug sets at expectation value $E = 1.0 \times 10^{-10}$ or better. For instance, the metabolite set for retinal dehydrogenase reaction set links, as expected, to the retinoid drugs at $E = 3.05 \times E^{-98}$, but the valine decarboxylase reaction, which is not an MDDR drug target, does not link significantly to any drug sets. These strikingly similar results suggest both broad coverage (54% of drug sets and 40% of metabolite sets) and specificity (single sets link to just 0.9% of metabolite sets and 1.1% of drug sets, respectively). For full results, see **Dataset S3**.

To determine the utility of the method for recovery of known drug-target interactions, it was hypothesized that chemical similarity between MetaCyc reaction sets and corresponding MDDR drug sets could specifically recover the known drug-target interactions. The 246 MDDR drug set targets include 62 enzymes that could be mapped to MetaCyc via the Enzyme

Commission (EC) number[30] describing the overall reaction catalyzed.[31] The results show that all

62 reaction sets for these targets link to at least one MDDR drug set. The majority of best hits

(42 out of 62) were found at expectation values of $E = 1.0 \times 10^{-10}$ or better (**Table 1**). At

expectation values better than $E = 1.0 \times 10^{-25}$, 61% (19 of 31) of best hits recover either the

specific known target or another enzyme in the same pathway. Examples of specific compounds

linked by this analysis are given in **Figure 2** for a selected group of these best-scoring hits.

Table 3.1 Metabolic enzyme targets and their best links to MDDR

Enzyme Target[a]	EC#	Best Hit MDDR Drug Set	Best Hit E-value
Adenosine kinase	*2.7.1.20*	*S-Adenosyl-L-Homocysteine Hydrolase Inhibitor*	*4.38E-288*
Adenosylmethionine decarboxylase	4.1.1.50	S-Adenosyl-L-Homocysteine Hydrolase Inhibitor	2.71E-216
Thromboxane-A synthase	*5.3.99.5*	*Prostaglandin*	*1.66E-204*
Adenosylhomocysteinase	**3.3.1.1**	**S-Adenosyl-L-Homocysteine Hydrolase Inhibitor**	**4.73E-203**
Adenosine deaminase	3.5.4.4	Adenosine (A1) Agonist	7.69E-159
Thymidine kinase	**2.7.1.21**	**Thymidine Kinase Inhibitor**	**3.19E-151**
Dihydrofolate reductase	1.5.1.3	Glycinamide Ribonucleotide Formyltransferase Inhibitor	1.02E-134
Catechol O-methyltransferase	2.1.1.6	S-Adenosyl-L-Homocysteine Hydrolase Inhibitor	4.67E-127
Prostaglandin-endoperoxide synthase	**1.14.99.1**	**Prostaglandin**	**8.57E-110**
Purine-nucleoside phosphorylase	2.4.2.1	Adenosine (A1) Agonist	8.35E-105
Ribose-phosphate pyrophosphokinase	2.7.6.1	S-Adenosyl-L-Homocysteine Hydrolase Inhibitor	4.33E-91
Phosphoribosylglycinamide formyltransferase	**2.1.2.2**	**Glycinamide Ribonucleotide Formyltransferase Inhibitor**	**1.55E-82**
Phosphoribosylaminoimidazolecar boxamide formyltransferase	*2.1.2.3*	*Glycinamide Ribonucleotide Formyltransferase Inhibitor*	*9.12E-80*
3',5'-cyclic-nucleotide phosphodiesterase	3.1.4.17	S-Adenosyl-L-Homocysteine Hydrolase Inhibitor	1.23E-77
Thymidylate synthase	**2.1.1.45**	**Thymidylate Synthetase Inhibitor**	**2.54E-75**
Steryl-sulfatase	*3.1.6.2*	*Aromatase Inhibitor*	*4.90E-62*

Enzyme Target[a]	EC#	Best Hit MDDR Drug Set	Best Hit E-value
Guanylate cyclase	4.6.1.2	Purine Nucleoside Phosphorylase Inhibitor	2.68E-60
Cholestenone 5-alpha-reductase	**1.3.1.22**	**Steroid (5alpha) Reductase Inhibitor**	**3.63E-60**
Steroid 17-alpha-monooxygenase	*1.14.99.9*	*Steroid (5alpha) Reductase Inhibitor*	*1.37E-58*
RNA-directed DNA polymerase	2.7.7.49	S-Adenosyl-L-Homocysteine Hydrolase Inhibitor	1.06E-52
Alpha-glucosidase	**3.2.1.20**	**Glucosidase (alpha) Inhibitor**	**1.00E-51**
Farnesyl-diphosphate farnesyltransferase	**2.5.1.21**	**Squalene Synthase Inhibitor**	**2.12E-46**
Beta-galactosidase	*3.2.1.23*	*Glucosidase (alpha) Inhibitor*	*4.04E-46*
Sterol esterase	3.1.1.13	Phospholipase A2 Inhibitor	3.18E-44
Leukotriene-A4 hydrolase	*3.3.2.6*	*Prostaglandin*	*5.16E-40*
Squalene monooxygenase	*1.14.99.7*	*Squalene Synthase Inhibitor*	*7.59E-40*
Ribonucleoside-diphosphate reductase	1.17.4.1	S-Adenosyl-L-Homocysteine Hydrolase Inhibitor	2.47E-38
3-hydroxyanthranilate 3,4-dioxygenase	**1.13.11.6**	**3-Hydroxyanthranilate Oxygenase Inhibitor**	**1.14E-33**
Dihydroorotase	**3.5.2.3**	**Dihydroorotase Inhibitor**	**2.25E-32**
Nitric-oxide synthase	**1.14.13.39**	**Nitric Oxide Synthase Inhibitor**	**8.86E-28**
Phospholipase A2	**3.1.1.4**	**Phospholipase A2 Inhibitor**	**9.82E-26**
Diaminopimelate epimerase	5.1.1.7	Nitric Oxide Synthase Inhibitor	2.43E-24
Membrane dipeptidase	3.4.13.19	Nitric Oxide Synthase Inhibitor	2.81E-23
3-alpha(or 20-beta)-hydroxysteroid dehydrogenase	*1.1.1.53*	*Aromatase Inhibitor*	*1.51E-22*
Sterol O-acyltransferase	2.3.1.26	Adenosine (A2) Agonist	4.95E-22
Hydroxymethylglutaryl-CoA reductase (NADPH)	1.1.1.34	Adenosine (A2) Agonist	4.95E-22
IMP dehydrogenase	1.1.1.205	Adenosine (A1) Agonist	8.98E-17
ATP-citrate (pro-S-)-lyase	4.1.3.8	Adenosine (A2) Agonist	1.83E-15
Glutamate--cysteine ligase	6.3.2.2	Nitric Oxide Synthase Inhibitor	2.71E-11
Dopamine-beta-monooxygenase	1.14.17.1	Adrenergic (beta1) Agonist	3.81E-11
Lanosterol synthase	*5.4.99.7*	*Squalene Synthase Inhibitor*	*1.38E-10*
Nucleoside-diphosphate kinase	2.7.4.6	P2T Purinoreceptor Antagonist	2.76E-10

[a]Exact matches (the enzyme is the canonical target of the best MDDR hit) are shown in **bold** type, pathway matches (the enzyme shares the same pathway as the canonical target of the best MDDR hit) are shown in *italic* type, and enzymes not in the same pathway as the canonical target are shown in regular type.
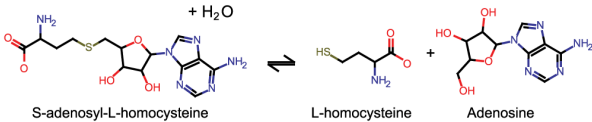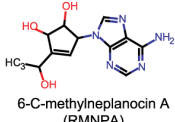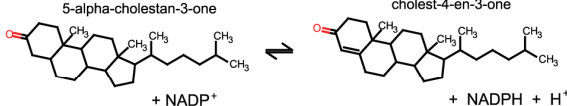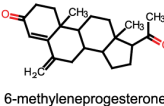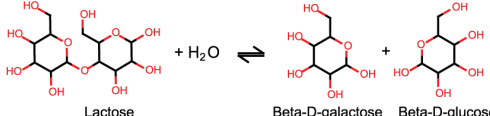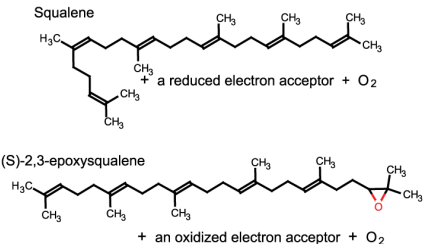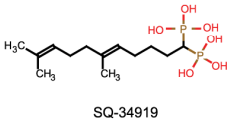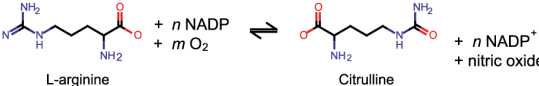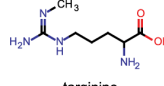
Enzyme Target	Substrates and products	E-value of best hit	Representative Inhibitor
S-Adenosyl-L-Homocysteine Hydrolase	S-adenosyl-L-homocysteine + H_2O ⇌ L-homocysteine + Adenosine	4.73E-203 S-Adenosyl-L-Homocysteine Hydrolase Inhibitor	6-C-methylneplanocin A (RMNPA)
Cholestenone-5-alpha-reductase	5-alpha-cholestan-3-one + NADP$^+$ ⇌ cholest-4-en-3-one + NADPH + H$^+$	3.63E-60 Steroid-5-alpha reductase inhibitors	6-methyleneprogesterone
Lactase	Lactose + H_2O ⇌ Beta-D-galactose + Beta-D-glucose	4.04E-46 Glucosidase-alpha inhibitor	Camiglibose
Squalene monooxygenase	Squalene + a reduced electron acceptor + O_2 ⇌ (S)-2,3-epoxysqualene + an oxidized electron acceptor + O_2	7.59E-40 Squalene synthase inhibitor	SQ-34919
Nitric oxide synthase (NOS)	L-arginine + n NADP + m O_2 ⇌ Citrulline + n NADP$^+$ + nitric oxide	8.86E-28 Nitric oxide synthase inhibitors	targinine

Figure 3.2 Selected best hits between MetaCyc reaction sets and MDDR drug sets

Other links recovered off-pathway hits, which often reflect known polypharmacology that is well-documented. For example, the glycinamide ribonucleotide formyltransferase (GART) inhibitor drug set hits both the GART reaction set ($E = 1.55 \times 10^{-82}$) and the off-pathway but pharmacologically related antifolate target dihydrofolate reductase (DHFR) ($E = 1.02 \times 10^{-134}$). Other off-pathway hits reflect biological connections, or physical connections, between targets. For example, the adenosine deaminase reaction set links to the A_1 adenosine receptor agonist drug set ($E = 7.69 \times 10^{-159}$) (**Table 1**) capturing the known interaction between A_1 adenosine receptors and adenosine deaminase on the cell surface of smooth muscle cells.[32] Considering only the stringent case of exact matches based on EC numbers, a Mann-Whitney rank-sum test

(also referred to as the U-test) shows that the expectation values for links between reaction sets and drug sets of known drug target enzymes were significantly better than the expectation values for links to reaction sets of non-target enzymes, i.e., 62 known enzyme targets were recovered in a background of 4,920 non-target "other" enzymes at a statistical significance of $P = 2.01 \times 10^{-6}$.

In addition to recapitulating many known drug-target interactions, the links identified by these comparisons also suggest new hypotheses about drug-target interactions. One such new prediction involves the phospholipase A2 (PLA2) inhibitor drug class. The substrates and products of PLA2 recapitulate its known link to the PLA2 inhibitor drug set ($E = 9.82 \times 10^{-26}$), however, the sterol esterase reaction returns an even better score against the PLA2 inhibitor set ($E = 3.18 \times 10^{-44}$) (**Table 1**). Although this predicted pharmacological relationship has, to our knowledge, not been previously documented, the result is consistent with the known biological relationship between PLA2 and sterol esterase. Both enzymes are secreted by the pancreas and require phosphatidylcholine hydrolysis to facilitate intestinal cholesterol uptake.[33] Thus, this link suggests that therapeutic agents directed against PLA2 may also inhibit sterol esterase, perhaps even more strongly than their intended target.

II. Human drug "effect-space" maps detail interactions between drug classes and enzyme targets

To present links between small molecule metabolites and drugs in the context of their known (and potential) metabolic targets, metabolic "effect-space" maps for currently marketed drugs were generated for each of the 246 drug classes investigated in this work. These maps enable visualization of the chemical similarities between drugs and metabolites painted onto human metabolic pathways, illustrating potential interactions between an individual drug class and specific metabolic enzymes in humans. Examples include the nucleoside reverse transcriptase,

dihydrofolate reductase, and thymidylate synthase inhibitors which target pyrimidine nucleotide

metabolism and biosynthesis of the essential coenzyme folate (**Figure 3 and Table 2**). Using

the canonical human metabolic pathways from HumanCyc,[34] a subset of the BioCyc[12] database

collection, reactions in each metabolic network have been colored according to their similarity to

known drug classes (**Figure 3**). While **Table 1** presents only the top link for each of 62 enzyme

targets in MetaCyc against the 246 MDDR drug classes, the networks in **Figure 3** detail all

significant hits for selected drug classes against the pyrimidine and folate pathways. Interactive

versions of these maps, one for each of the 246 drug classes included in our analysis, are

available online (see below).

Table 3.2 Links between selected drug classes and top ranked metabolic reactions

Rank	Thymidylate Synthetase (TS) Inhibitor	E-value
1	Dihydrofolate reductase (DHFR)	1.96E-123
2	Methyltetrahydrofolate-corrinoid-iron-sulfur protein methyltransferase	3.58E-102
3	Methionyl-tRNA formyltransferase	1.97E-99
4	Methylenetetrahydrofolate reductase	2.67E-86
5	Thymidylate synthase (TS)	2.54E-75
6	Formate-tetrahydrofolate ligase	1.44E-74
7	Dihydrofolate synthetase	1.35E-70
8	Aminomethyltransferase	7.13E-63
9	5-methyltetrahydrofolate-homocysteine S-methyltransferase	2.80E-62
10	Phosphoribosylaminoimidazolecarboxamide (AICAR) formyltransferase	1.50E-60
11	Phosphoribosylglycinamide formyltransferase (GART)	1.50E-60
Rank	Dihydrofolate Reductase (DHFR) Inhibitor	E-value
1	Dihydrofolate reductase (DHFR)	1.46E-82
2	Methyltetrahydrofolate-corrinoid-iron-sulfur protein methyltransferase	2.84E-75
3	Methylenetetrahydrofolate reductase	6.01E-73
4	Methionyl-tRNA formyltransferase	7.00E-66
5	Aminomethyltransferase	6.90E-55
6	Formate-tetrahydrofolate ligase	6.15E-49
7	Thymidylate synthase (TS)	1.91E-48
8	5-methyltetrahydrofolate-homocysteine S-methyltransferase	2.60E-45

Rank		E-value
9	3-methyl-2-oxobutanoate hydroxymethyltransferase	2.68E-44
10	Glycine decarboxylase	2.68E-44
11	Glycine hydroxymethyltransferase (SHMT)	2.68E-44
12	Dihydrofolate synthetase	9.65E-42
13	Phosphoribosylaminoimidazolecarboxamide (AICAR) formyltransferase	2.21E-39
14	Phosphoribosylglycinamide formyltransferase (GART)	2.21E-39
Rank	**Nucleoside Reverse Transcriptase Inhibitor (NRTI)**	**E-value**
1	Thymidylate kinase	7.48E-28
2	Thymidine kinase	3.48E-26
3	Deoxythymidine diphosphate kinase	1.54E-24
4	Ribonucleoside-triphosphate reductase	2.88E-14
5	Deoxyuridine triphosphate pyrophosphatase	5.60E-12
6	Deoxyuridine kinase	1.14E-11
7	Deoxyuridine diphosphate kinase	1.45E-11
8	Thymidylate synthase (TS)	5.68E-11
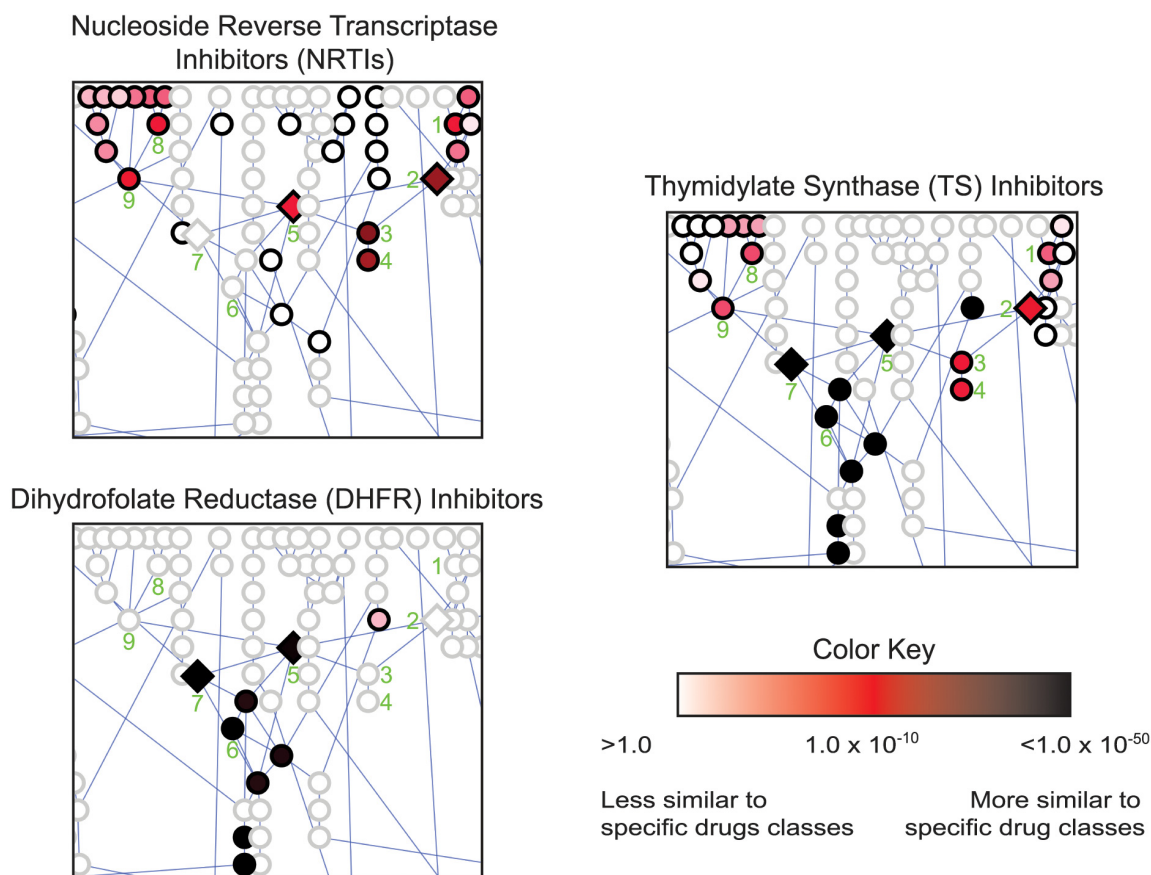
Figure 3.3 Effect-space map showing chemical similarity between specific drug classes and metabolites in human folate and pyrimidine biosynthesis

Each node represents one reaction set – the substrates and products of a single human metabolic reaction. Edges connect the reactions in the canonical pathway as annotated in HumanCyc[34]. As given in the color key, each reaction is colored according to the expectation value indicating the strength of similarity between that target reaction set and the respective MDDR drug set. Diamond shaped nodes indicate reactions catalyzed by enzymes annotated as known drug targets in the MDDR; circles indicate reactions catalyzed by enzymes not annotated as targets. Reaction key: 1. Deoxyuridine kinase 2. Thymidine kinase 3. Thymidylate kinase 4. Deoxythymidine diphosphate kinase 5. Thymidylate synthase (TS) 6. Methylene tetrahydrofolate reductase 7. Dihydrofolate reductase (DHFR) 8. Deoxyuridine diphosphate kinase 9. Deoxyuridine triphosphate diphosphatase.

It has previously been shown that chemical similarity between known drugs often suggests novel drug-target interactions.[5-7, 14] Consistent with these observations, effect-space maps such as those shown in **Figure 3** can also be used to exploit chemical similarities between drugs and metabolites to indicate potential routes of drug metabolism and toxicity.[3, 11, 35, 36] For example, the nucleotide reverse transcriptase inhibitors (NRTIs) used in HIV therapy are administered as pro-drugs. The effect-space map reflects this route of NRTI metabolism leading to viral inhibition. The top three hits yielded by the NRTI drug set queried against human metabolism – thymidine kinase ($E = 3.48 \times 10^{-26}$), thymidylate kinase ($E = 7.48 \times 10^{-28}$), and deoxythymidine diphosphate kinase ($E = 1.54 \times 10^{-24}$) (**Figure 3** reaction numbers 2, 3, and 4; additional results in **Table 2**) – successively phosphorylate the NRTI pro-drugs into the pharmacologically active NRTI triphosphates.[37, 38] The viral reverse transcriptase enzyme then incorporates the fully phosphorylated NRTIs into the growing DNA strand, thereby terminating transcription of the viral DNA. In this example, considerable toxicity mitigates the therapeutic value of inhibiting viral DNA transcription since the phosphorylated NRTIs directly inhibit human nucleotide kinases and mitochondrial DNA pol-γ. They also may be incorporated by pol-γ into the growing human mitochondrial DNA strand, and once incorporated are inefficiently excised by DNA pol-γ exonuclease.[39] Thus, the effect-space map illustrates both the route of metabolism and a mechanism of toxicity for NRTIs in humans.

Drug effect-space maps also offer a broad glimpse of potential human metabolic interactions predicting new "polypharmacology". From the ligand perspective, "drug polypharmacology" refers to a single drug or drug class that hits multiple targets. For example, dihydrofolate reductase (DHFR, reaction number 7 in **Figure 3**) uses NADPH to reduce 7,8-dihydrofolate to tetrahydrofolate. Antifolate drugs inhibit DHFR, and, as expected, the DHFR drug set recovers the DHFR reaction substrates and products as the top similarity hit in human

metabolism ($E = 1.46 \times 10^{-82}$) (**Figure 3, Table 2, Figure 4**). However, at least 20 other

reactions also use folate coenzymes in human metabolism.[40-42] Accordingly, SEA finds additional

links between the DHFR drug set and established antifolate targets outside the pyrimidine and

folate biosynthesis pathways such as serine hydroxymethyltransferase (SHMT, $E = 2.68 \times 10^{-44}$),

phosphoribosyl-aminoimidazole-carboxamide formyltransferase (AICAR transformylase, $E = 2.21 \times 10^{-39}$), and phosphoribosyl-glycinamide formyltransferase (GART, $E = 2.21 \times 10^{-39}$)

(**Table 2**). The effect-space maps in **Figure 3** illustrate the results from **Table 2** and **Figure 4** in

a single view, illustrating drug polypharmacology with respect to critical metabolic pathways.
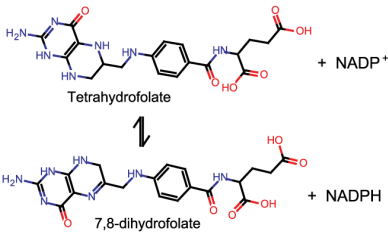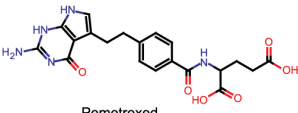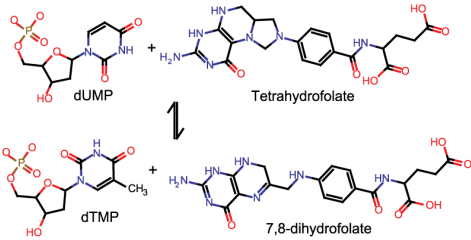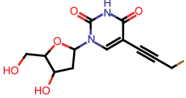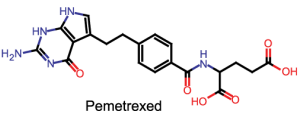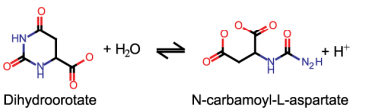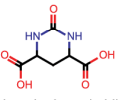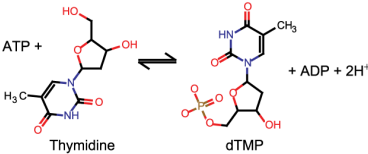
Figure 3.4 Selected links between MDDR drug classes and human folate and pyrimidine metabolism

Alternatively, from the target perspective, "target polypharmacology" may refer to a single target being modulated by multiple classes of drugs. For instance, thymidylate synthase (TS) is another classic antifolate target that uses a folate coenzyme to methylate deoxyuridine phosphate, generating deoxythymidine phosphate.[43-46] As expected, the TS reaction links to known antifolate drug classes such as GART inhibitors (E = 4.76 x 10^-73) and DHFR inhibitors (E = 1.91 x 10^-48) (**Table 3 and Figure 4**). However, TS is also effectively inhibited by uracil analogs such as fluoropropynyl deoxyuridine, which is not a folate, but rather a pyrimidine analog. Accordingly, the TS reaction also links to reverse transcriptase inhibitors, which include

fluoropropynyl deoxyuridine and additional pyrimidine analogs such as azidothymidine (AZT) ($E = 5.68 \times 10^{-11}$) (**Figure 4**). The target polypharmacology of the thymidylate synthase enzyme is mirrored by the drug polypharmacology of the thymidylate synthase inhibitors. The TS inhibitors link not only to the reactions of deoxyribonucleotide biosynthesis including thymidylate synthase ($E = 2.54 \times 10^{-75}$), but also the GART ($E = 1.50 \times 10^{-60}$) and DHFR ($E = 1.96 \times 10^{-123}$) reactions (**Figure 3 and Table 2**). Thus, SEA recapitulates the known polypharmacology of TS. Effect-space maps illustrate and clarify these pharmacological relationships.

Table 3.3 Links between selected metabolic reactions and top ranked drug classes

Rank	Thymidylate Synthetase (TS) Reaction	E-value
1	Thymidylate synthase inhibitor (TS)	2.54E-75
2	Glycinamide ribonucleotide formyltransferase inhibitor (GART)	4.76E-73
3	Thymidine kinase inhibitor (TK)	1.18E-62
4	Dihydrofolate reductase inhibitor (DHFR)	1.91E-48
5	Folylpolyglutamate synthetase inhibitor	2.27E-31
6	Nucleoside reverse transcriptase inhibitor (NRTI)	5.68E-11
Rank	**Dihydrofolate Reductase (DHFR) Reaction**	**E-value**
1	Glycinamide Ribonucleotide Formyltransferase Inhibitor	1.02E-134
2	Thymidylate Synthetase Inhibitor	1.96E-123
3	Dihydrofolate Reductase Inhibitor	1.46E-82
4	Folylpolyglutamate Synthetase Inhibitor	3.15E-62

III. Species-specific effect-space maps for pathogens and model organisms

The great diversity of metabolic strategies, pathways, and enzymes present in humans, model organisms, and pathogenic species presents both opportunities and significant barriers to drug discovery. To address these issues, species-specific effect-space maps were created for each of 385 organisms from the BioCyc Database Collection. Target reactions existing in common and

differentially between each of these species and humans are shown in these metabolic maps. As with the human effect-space maps, this set of maps is available in interactive form online. To show how these maps may be used to provide a context for drug discovery, MRSA is used as an example (**Figure 5**). The global view of drugs and metabolism provided by this species-specific map illustrates some of the daunting challenges to the selection of tractable metabolic drug targets in this organism.
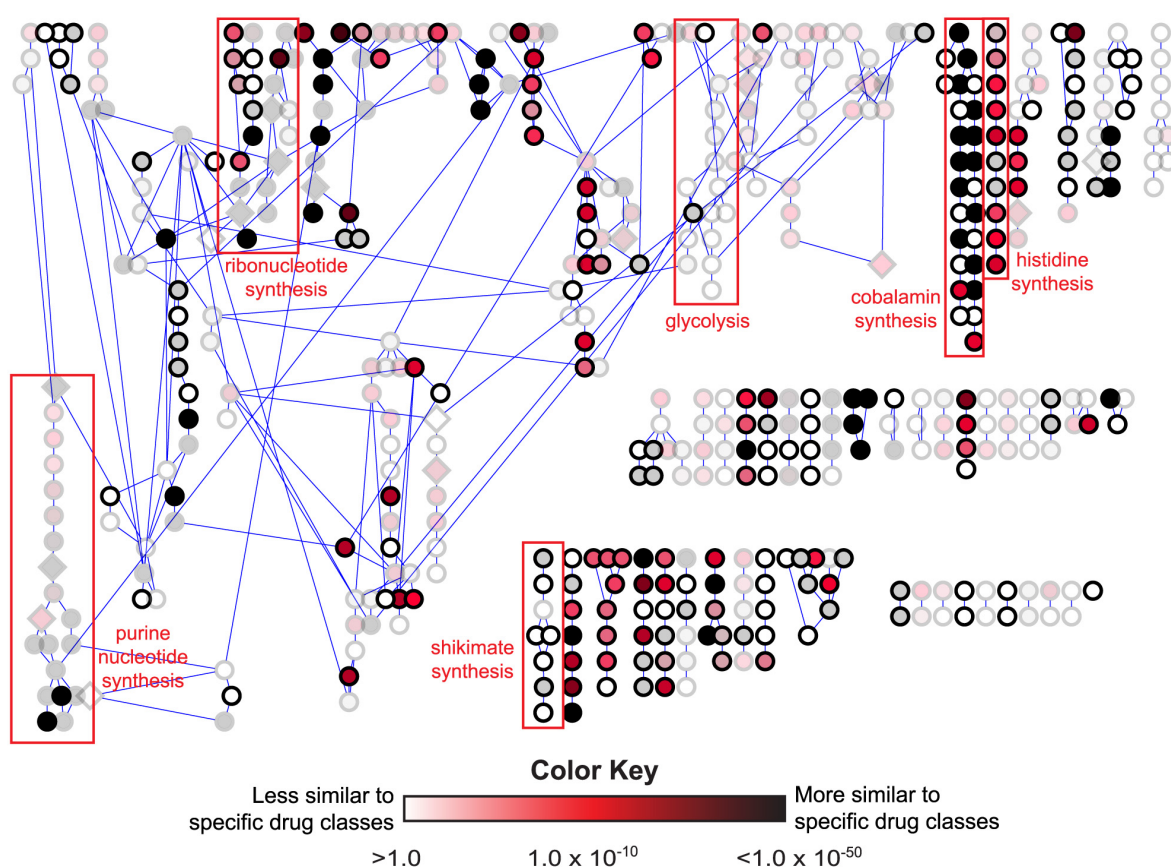
Figure 3.5 Effect-space map showing chemical similarity between drugs and metabolites in MRSA

Canonical pathway representation of methicillin-resistant *Staphylococcus aureus* (MRSA) [12] small molecule metabolism colored by expectation value of the best hit against MDDR. Reactions that are also present in humans have been faded. Layout based upon the Cytoscape 2.5 y-

files hierarchical layout. Edge lengths are not significant. For ease of viewing, reactions are not labeled but can be identified in the interactive versions of the maps available at the online resource.

As described for **Figure 3**, each node in the MRSA network in **Figure 5** represents one reaction set, the substrates and products of a single metabolic reaction. Edges connect the reactions according to canonical BioCyc MRSA pathways. Each reaction in the network has been colored according the expectation value of the best link between the reaction set and any of the 246 MDDR drug sets. Lighter colored nodes have higher expectation values indicating less drug-like reaction sets, while darker colored nodes indicate more drug-like reaction sets. To provide therapeutic context, reactions that are also present in human metabolism have been faded, indicating that drug sets targeting these enzymes in MRSA may have the undesirable potential to inhibit the human enzymes as well. As with the other organisms represented in our online maps, most reactions in the MRSA subset have little chemical similarity to any MDDR drug set. Although 74% of the 469 MRSA metabolic reactions have measurable similarity to at least one MDDR drug set, only 36% of these links had expectation values of $E = 1.0 \times 10^{-10}$ or better. Several complete pathways of diverse chemical classes, including shikimic acid, phospholipid, peptidoglycan, teichoic acid, and molybdenum cofactor biosynthesis, lack links to any drug set at all. Only 18 of the 469 MRSA metabolic reactions are already known to be drug targets in MDDR. Fourteen of these are represented in **Figure 5** (as diamonds), but all 18 of these also appear in humans. Enzymes that catalyze these reactions in humans would likely be vulnerable to inhibitors developed against these MRSA targets, putting those drugs at risk for toxicity.

Figure 6 illustrates how additional information can be used to further filter potential metabolic targets by painting additional biological or genomic data onto a species-specific map.

Since successful modulation of a target may not alone be sufficient to kill a pathogen due to the presence of redundant pathways for the formation of critical metabolites, integration of such additional information into a metabolic map may provide added value in addressing the multi-dimensional challenges of drug discovery. Flux balance analysis of metabolic networks was used by several of the authors of this work to identify essential enzymes and metabolites required for the formation of all necessary biomass components for 13 strains of *Staphylococcus aureus*, including the methicillin-resistant N315 strain (MRSA).[47] Using these results, 39 essential reactions and 19 synthetic lethal reaction pairs could be mapped to our dataset (**Figure 6**), highlighting those reactions for which inhibition is most likely to result in the death of the organism. Several of these reactions have been successfully targeted by currently marketed drugs, such as the previously discussed antifolate targets DHFR (E = 1.02×10^{-134}), thymidylate synthase (E = 2.54×10^{-75}), and dihydrofolate synthase (E = 1.35×10^{-70}). This retrospective result illustrates the potential of such additional information in enriching for targets and drug chemistry that have been proven accessible. Other targets and pathways have not yet yielded successful drugs but are under investigation in MRSA or other pathogens, such as the shikimate pathway[48] in aromatic amino acid biosynthesis and the histidine biosynthesis pathway.[49]
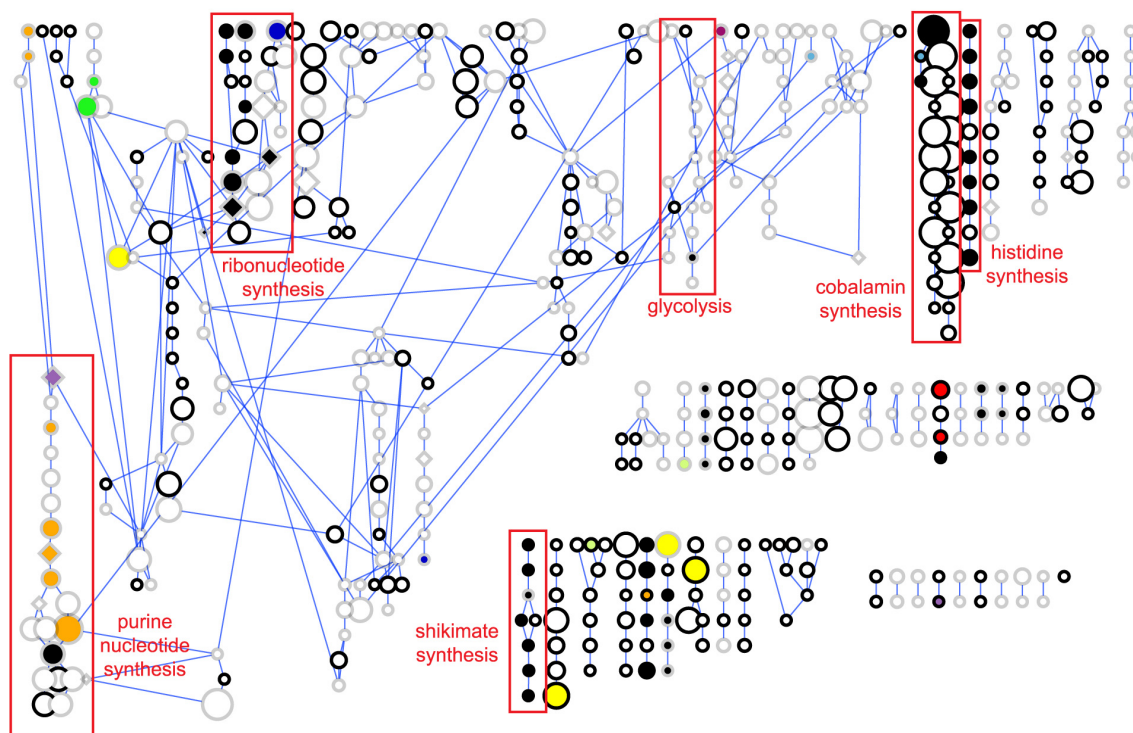
Figure 3.6 Essential and synthetic lethal map of MRSA metabolism

Canonical pathway representation of methicillin-resistant *Staphylococcus aureus* (MRSA) small molecule metabolism colored by essentiality and synthetic lethality of reactions. Key: black = essential reaction; other colors = synthetic lethal reaction pairs; node size = similarity to top MDDR hit (bigger is more drug-like); diamond shape = MDDR drug target; faded border = human reaction.

The combination of the essentiality data with the drug space mapping emphasizes the challenges to drug discovery against MRSA. Thus, while species-specific antifolates do exist, many antifolates such as methotrexate used in cancer therapy cause severe toxicity.[42] To avoid such toxicity, 14 of the 39 essential MRSA reactions that are also present in humans can be excluded from further consideration as drug targets in MRSA.

A compilation of all of the metabolic network maps generated in this study is available at http://sea.docking.org/metabolism. These include interactive versions of the human effect-

space maps shown in **Figure 3**, one for each of the 246 MDDR drug classes analyzed in this work, and 385 species-specific maps such as that shown in **Figure 5**. The species-specific maps were generated from the BioCyc database public collection, a compendium of 385 model organisms and pathogens whose genomes have been sequenced and their metabolomes deciphered. Of these, 65 have been designated as Priority Pathogens by the National Institute of Allergy and Infectious Diseases (NIAID) and include *Bacillus anthracis*, *Brucella melitensis*, *Cryptosporidium parvum*, *Salmonella*, SARS, *Toxoplasma gondii*, *Vibrio cholerae*, and *Yersinia pestis*.[50] Browse and similarity search tools are also provided, allowing exploration of the metabolic reaction sets and current drug classes used in this work, as well as comparison to user-defined custom ligand sets. These interactive tools enable facile exploration between the vast biological data on potential metabolic drug targets in these organisms and the drug chemistry currently available to prosecute those targets.

3.5 Discussion

A key product of this study is the construction of drug-metabolite correspondence maps that provide both a global view and a more contextual picture of predicted drug action in human metabolism than has been previously available. Several aspects of these maps deserve particular emphasis. First, despite the differences in physiochemical properties of most drugs and small molecule metabolites, numerous links arise between drugs and metabolism. Viewed in the context of metabolic networks, the pharmacological relationships predicted by these links can be readily interpreted in a way that is biologically sensible. Moreover, as shown by both the drug effect space maps and species-specific maps, our retrospective analyses confirm that biologically and pharmacologically significant connections can be recovered, capturing known polypharmacology and revealing the relevant chemotypes previously explored in drug

development. The metabolome-wide exploratory tools provided with these map sets also enable a new way to interrogate the links between drugs and metabolism that will likely be useful for prediction of new targets and to indicate routes of drug metabolism and toxicity. Further, by integrating biological information such as essentiality and synthetic lethal analyses with the metabolic context, our approach allows users to focus evaluation of potential targets around specific types of data simply by painting the results on to metabolic maps.

With respect to the coverage of drug links across small molecule metabolism that this study provides, we note that the SEA method relies solely upon the chemical similarity of ligands to establish links between drug sets and reaction sets. Based on these links, and the biologically sensible connections shown in the results, we infer that a particular drug class may act on a certain target. However, drugs may also act against an enzyme active site without resembling the endogenous substrate, or by allosteric regulation at an entirely different site. The SEA method, as applied here to the substrates and products of metabolic reactions, does not capture these additional drug-target links. Other viable strategies are available for targeting metabolic enzyme active sites that use principles unrelated to the ligand-drug similarities that are the focus of our approach.[51-54] For instance, Tondi et al. designed novel inhibitors of thymidylate synthase that complemented the three dimensional structure of the active site. Five high-scoring compounds selected for testing were dissimilar to the substrate but bound competitively with it.[54] While many classical kinase inhibitors interact directly with the ATP binding site, imatinib (tradename Gleevec) represents a new generation of allosteric protein kinase inhibitors that alter the kinase conformation to prevent ATP binding. Other allosteric kinase inhibitors prevent the protein substrate from loading.[51]

While a quantitative determination of the proportion of drug-target links that cannot be accessed by our approach is beyond the scope of this study, we can provide a rough estimate for

the frequency of such cases based on the results reported in **Table 1**. Of the 62 known enzyme targets in MetaCyc, 42 (68%) the substrate/product metabolite sets show significant chemical similarity to at least one MDDR drug set, establishing a reasonable first pass estimate for the percentage of current enzyme targets accessible to this approach. Furthermore, 40% (2,044 of 5,056) of all MetaCyc reaction sets linked at $E = 1.0 \times 10^{-10}$ or better to MDDR, with each reaction linking to an average of just 2.8 MDDR drug sets. These results indicate broad and specific coverage of metabolism, and suggest that numerous additional enzyme targets may be accessible by the method presented here.

3.6 Conclusion

Using the SEA method, we have shown that comparison between ligand sets representing MDDR drug classes and ligand sets representing the substrates and products of metabolic reactions yields statistically significant links between known drugs and enzyme targets. Because the method is based on chemical similarity and requires only information from these molecule sets rather than the sequence, structure or physiochemistry of the targets, this ligand-based approach is independent from, and complementary to, protein structure and sequence based methods. Our results also suggest the potential of this method for predicting previously unknown interactions between drug classes and metabolic targets, recovering routes of metabolism and toxicity in humans, and identifying potential drug targets (as well as challenges for target discovery) in emerging pathogens. Thus, by mapping the chemical diversity of drugs to small molecule metabolism using ligand topology, this work establishes a computational framework for ligand-based prediction of drug class action, metabolism, and toxicity.

3.7 Methods

I. Compound sets

All compounds, both drugs and metabolites, are represented using Daylight SMILES strings.[55] Sets comprised of isomers with unique compound names were retained, even though stereochemistry was later removed as part of the molecule fingerprinting process.

II. Ligand sets

Reaction sets were extracted from the 8.15.2007 release of MetaCyc based upon the substrates and products annotated to each reaction. Two filters were applied. First, the ten most common metabolites based on the number of occurrences in the MetaCyc metabolic network were removed: water, ATP, ADP, NAD, pyrophosphate, NADH, carbon dioxide, AMP, glutamate, and pyruvate. Second, each reaction set was required to include at least two unique compounds, as indicated by a MetaCyc or a MDDR unique compound id.

III. Drug sets

Drug sets were extracted from the MDDR, a compilation of about 169,000 drug-like ligands in 688 activity classes, each targeting a specific enzyme (designated by the Enzyme Commission (E.C.) number). The subset of this database for which mappings between enzymes and the MDDR drug classes were available was used. These mappings were based on a previous study that maps E.C. numbers, GPCRs, ion channels and nuclear receptors to MDDR activity classes.[31] Only sets containing five or more ligands were used. Salts and fragments were removed, ligand protonation was normalized and duplicate molecules were removed. Of the 688 targets in the MDDR, 97 were excluded as having too few ligands (<5), and another 345 targets

were excluded because their definitions did not describe a molecular target, e.g., drugs associated only with an annotation such as "Anticancer" were not used. The remaining 246 enzyme targets were together associated with a total of 65,241 unique ligands, with a median and mean of 124 and 289 drug ligands per target. For further details, see Keiser et al.[6]

IV. Set comparisons

All pairs of ligands between any two sets were compared using a pair-wise similarity metric, which consists of a descriptor and a similarity criterion. For the similarity descriptor, standard two-dimensional topological fingerprints were computed using the Scitegic ECFP4 fingerprint.[56] The similarity criterion was the widely used Tanimoto coefficient (Tc).[57] For set comparisons, all pair-wise Tcs between elements across sets were calculated, and those scoring above a threshold were summed, giving a raw score relating the two sets. The Tanimoto coefficient threshold of 0.32 was determined according to a previously published method based upon fit to an extreme value distribution.[6] A model for random similarity similar to that used by BLAST[58] was used to generate expectation values (E) which are used to describe the strengths of relationships discovered using this protocol.[6] All scores reported here are based upon the background distribution and cutoff scores generated using the drug sets extracted from the MDDR collection. For further details, see Keiser et al.[6] Network visualization was performed in Cytoscape 2.6.2[59] using the γ-files hierarchical layout algorithm.

V. MRSA essentiality and synthetic lethal analysis

Essentiality and synthetic lethal data generated as described earlier.[47] Briefly, the metabolic network was reconstructed from the genome to include all reactions that have an active flux The essentiality of a given enzyme was then assessed by the effect of the removal of that enzyme on

biomass production. Similarly, synthetic lethal pairs can be identified by systematic pairwise deletion of enzymes and recalculation of biomass production in an ideally rich medium.

3.8 Acknowledgments

3.9 References

1. Johnson, M., Lajiness, M. & Maggiora, G. Molecular similarity: a basis for designing drug screening programs. *Prog Clin Biol Res* **291**, 167-171 (1989).

2. Payne, D.J., Gwynn, M.N., Holmes, D.J. & Pompliano, D.L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov* **6**, 29-40 (2007).

3. Kramer, J.A., Sagartz, J.E. & Morris, D.L. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nat Rev Drug Discov* **6**, 636-649 (2007).

4. Drews, J. Case histories, magic bullets and the state of drug discovery. *Nat Rev Drug Discov* **5**, 635-640 (2006).

5. 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).

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

7. Cleves, A.E. & Jain, A.N. Robust ligand-based modeling of the biological targets of known drugs. *J Med Chem* **49**, 2921-2938 (2006).

8. Watkins, S.M. & German, J.B. Metabolomics and biochemical profiling in drug discovery and development. *Curr Opin Mol Ther* **4**, 224-228 (2002).

9. Shyur, L.F. & Yang, N.S. Metabolomics for phytomedicine research and drug development. *Curr Opin Chem Biol* **12**, 66-71 (2008).

10. Rochfort, S. Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. *J Nat Prod* **68**, 1813-1820 (2005).

11. Kell, D.B. Systems biology, metabolic modelling and metabolomics in drug

discovery and development. *Drug Discov Today* **11**, 1085-1092 (2006).

12. Caspi, R. et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **36**, D623-631 (2008).

13. Dobson, C.M. Chemical space and biology. *Nature* **432**, 824-828 (2004).

14. Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L. & Vidal, M. Drug-target network. *Nat Biotechnol* **25**, 1119-1126 (2007).

15. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. & Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **24**, i232-240 (2008).

16. Cheng, A.C. et al. Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* **25**, 71-75 (2007).

17. Goh, K.I. et al. The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-8690 (2007).

18. Hajduk, P.J., Huth, J.R. & Tse, C. Predicting protein druggability. *Drug Discov Today* **10**, 1675-1682 (2005).

19. Hopkins, A.L. & Groom, C.R. The druggable genome. *Nat Rev Drug Discov* **1**, 727-730 (2002).

20. Imming, P., Sinning, C. & Meyer, A. Drugs, their targets and the nature and number of

drug targets. *Nat Rev Drug Discov* **5**, 821-834 (2006).

21. Meisner, N.C. et al. The chemical hunt for the identification of drugable targets. *Curr Opin Chem Biol* **8**, 424-431 (2004).

22. Russ, A.P. & Lampel, S. The druggable genome: an update. *Drug Discov Today* **10**, 1607-1610 (2005).

23. Lee, D.S. et al. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci U S A* **105**, 9880-9885 (2008).

24. Navarro, M.B., Huttner, B. & Harbarth, S. Methicillin-resistant Staphylococcus aureus control in the 21st century: beyond the acute care hospital. *Curr Opin Infect Dis* **21**, 372-379 (2008).

25. Powell, J.P. & Wenzel, R.P. Antibiotic options for treating community-acquired MRSA. *Expert Rev Anti Infect Ther* **6**, 299-307 (2008).

26. Clements, A. et al. Overcrowding and understaffing in modern health-care systems: key determinants in meticillin-resistant Staphylococcus aureus transmission. *Lancet Infect Dis* **8**, 427-434 (2008).

27. Avdic, E. & Cosgrove, S.E. Management and control strategies for community-associated methicillin-resistant Staphylococcus aureus. *Expert Opin Pharmacother* **9**, 1463-1479 (2008).

28. Nicasio, A.M., Kuti, J.L. & Nicolau, D.P. The current state of multidrug-resistant gram-negative bacilli in North America. *Pharmacotherapy* **28**, 235-249 (2008).

29. MDL Information Systems, I. MDL Drug Data Report. (MDL Information Systems, Inc., San Leandro, CA; 2006).

30. Tipton, K.F. (NC-IUBMB, New York; 1992).

31. Schuffenhauer, A. et al. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J Chem Inf Comput Sci* **42**, 947-955 (2002).

32. Ciruela, F. et al. Adenosine deaminase affects ligand-induced signalling by interacting with cell surface adenosine receptors. *FEBS Lett* **380**, 219-223 (1996).

33. Mackay, K., Starr, J.R., Lawn, R.M. & Ellsworth, J.L. Phosphatidylcholine hydrolysis is required for pancreatic cholesterol esterase- and phospholipase A2-facilitated cholesterol uptake into intestinal Caco-2 cells. *J Biol Chem* **272**, 13380-13389 (1997).

34. Romero, P. et al. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6**, R2 (2005).

35. Martin, R., Rose, D., Yu, K. & Barros, S. Toxicogenomics strategies for predicting drug toxicity. *Pharmacogenomics* **7**, 1003-1016 (2006).

36. Ekins, S. et al. Computational prediction of human drug metabolism. *Expert Opin Drug Metab Toxicol* **1**, 303-324 (2005).

37. Lewis, W. Cardiomyopathy, nucleoside reverse transcriptase inhibitors and mitochondria are linked through AIDS and its therapy. *Mitochondrion* **4**, 141-152 (2004).

38. Petit, F., Fromenty, B., Owen, A. & Estaquier, J. Mitochondria are sensors for HIV drugs. *Trends Pharmacol Sci* **26**, 258-264 (2005).

39. Lewis, W. et al. Antiretroviral nucleosides, deoxynucleotide carrier and mitochondrial DNA: evidence supporting the DNA pol gamma hypothesis. *Aids* **20**, 675-684 (2006).

40. Kisliuk, R.L. Synergistic interactions among antifolates. *Pharmacol Ther* **85**, 183-190 (2000).

41. Faessel, H.M., Slocum, H.K., Rustum, Y.M. & Greco, W.R. Folic acid-enhanced synergy for the combination of trimetrexate plus the glycinamide ribonucleotide formyltransferase inhibitor 4-[2-(2-amino-4-oxo-4,6,7,8-tetrahydro-3H-pyrimidino[5,4,6][1,4]thiazin -6-yl)-(S)-ethyl]-2,5-thienoylamino-L-glutamic acid (AG2034): comparison across sensitive and resistant human tumor cell lines. *Biochem Pharmacol* **57**, 567-577 (1999).

42. Chan, D.C. & Anderson, A.C. Towards species-specific antifolates. *Curr Med Chem* **13**, 377-398 (2006).

43. Costi, M.P. et al. Thymidylate synthase structure, function and implication in drug discovery. *Curr Med Chem* **12**, 2241-2258 (2005).

44. Gmeiner, W.H. Novel chemical strategies for thymidylate synthase inhibition. *Curr Med Chem* **12**, 191-202 (2005).

45. McGuire, J.J. Anticancer antifolates: current status and future directions. *Curr Pharm Des* **9**, 2593-2613 (2003).

46. Chu, E., Callender, M.A., Farrell, M.P. & Schmitz, J.C. Thymidylate synthase inhibitors as anticancer agents: from bench to bedside. *Cancer Chemother Pharmacol* **52 Suppl 1**, S80-89 (2003).

47. Lee, D.S. et al. Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. *J Bacteriol* **191**, 4015-4024 (2009).

48. Dias, M.V. et al. Chorismate synthase: an attractive target for drug development against orphan diseases. *Curr Drug Targets* **8**, 437-444 (2007).

49. Cho, Y., Ioerger, T.R. & Sacchettini, J.C. Discovery of novel nitrobenzothiazole inhibitors for Mycobacterium tuberculosis ATP phosphoribosyl transferase (HisG) through virtual screening. *J Med Chem* **51**, 5984-5992 (2008).

50. Zhang, C. et al. An emerging cyberinfrastructure for biodefense pathogen and pathogen-host data. *Nucleic Acids Res* **36**, D884-891 (2008).

51. Bogoyevitch, M.A. & Fairlie, D.P. A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding. *Drug Discov Today* **12**, 622-633 (2007).

52. Ciulli, A. & Abell, C. Fragment-based approaches to enzyme inhibition. *Curr Opin Biotechnol* **18**, 489-496 (2007).

53. Moore, E.C., Hurlbert, R.B., Boss, G.R. & Massia, S.P. Inhibition of two enzymes in de novo purine nucleotide synthesis by triciribine phosphate (TCN-P). *Biochem Pharmacol* **38**, 4045-4051 (1989).

54. Tondi, D. et al. Structure-based discovery and in-parallel optimization of novel competitive inhibitors of thymidylate synthase. *Chem Biol* **6**, 319-331 (1999).

55. James, C., Weininger, D. & Delaney, J. (Daylight Chemical Information Systems Inc., Mission Viejo, CA; 1992-2005).

56. Hert, J. et al. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org Biomol Chem* **2**, 3256-3266 (2004).

57. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* **11**, 1046-1053 (2006).

58. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment

search tool. *Jour. Mol. Biol.* **215**, 403-410 (1990).

59. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).

Chapter 4:

Future Directions

Whither SEA? More broadly, to what uses will we put statistical chemical similarity as both processing power and bioactivity data become ever more available? I touch here on three directions for current and future exploration. Lacking claim to Delphic foresight, I have chosen these by scientific criteria of impact, feasibility, and the extent to which the prospect of their continued existence as "directions" instead of "Chapters" exasperates me.

In the first direction (**4.1**), I revisit the protein function identification engine alluded to in the Introduction. In the second (**4.2**-**4.3**), I argue that similarity need not exclude novelty and illustrate this by example. In the last (**4.4**-**4.5**), I highlight several baffling patterns of chemical similarity that, like mice in the night, have left little more than their paw-print tracks of implication across the pages of my lab books. I suspect that, cornered, some such mice may roar.

4.1 From DOCK hits to protein function

"Form finds function" – so went the tagline on the cover of *Nature* in August 2007 when Johannes Hermann, in collaboration with Dr. Raushel's group, demonstrated the use of a DOCK[1] hit list[§] to deduce a protein's function.[2] What if we could find patterns in these lists automatically, and annotate structural genomics proteins on a broad scale? This "future" direction was in fact my primary motivation for developing SEA, and Brian Shoichet sketched out its general form when we discussed my rotation project five years ago. If the Similarity Ensemble Approach is now the core of the "protein identification engine" described in the Introduction, then DOCK was to be its fuel. After all, why apply SEA to the high challenge of protein function, when doing so presupposes that many ligands already be known for each "uncharacterized" protein? Wouldn't we already have a good idea of what that protein does by the time we had so many ligands for it, thereby defeating our own purpose? But on the other hand—do we actually need to the true ligands, or would putative ones be enough? Docking[1, 3] could certainly provide these.

I hypothesize that *sets of putative ligands will inform on their targets' pharmacology, even when any given ligand prediction itself is unreliable.* For putative ligands we turn to virtual screening, in which we dock libraries of candidate small molecules against target protein structures.[1, 3] Whereas the "hit lists" that result are prone to error in their individual predictions, the molecules that rank highly are typically sensible overall. Indeed the hit list as a whole may be more reliable than its individual components. This in itself would be a novel use of virtual screening and would provide valuable starting points for structure-based protein annotation.

[§] A target's "hit list" is the list of molecules predicted by a virtual screening program such as DOCK to bind to that protein target. It consists only of putative ligands, many of which will not actually bind.

I envision two main ways to exploit virtual screening against proteins of uncharacterized pharmacology. Both ways, however, share the following common framework:

1. The input is the docking hit list for the uncharacterized protein.

2. The output is that hit list's similarity to ligand sets for characterized proteins.

I. Approach 1: Hit lists vs. hit lists

Hypothesis: *The docking hit lists for two pharmacologically similar proteins are similar by SEA.*

The glucocorticoid and progesterone nuclear receptors both bind hormones, and indeed progesterone acts an antagonist at the glucocorticoid receptor. It may be no surprise, then, that the docking hit lists for each would have some overlap in their predicted ligands, and that the one list would, on the whole, resemble the other. Initial results suggest that the docking hit lists for these two receptors have strong similarity to each other by SEA (**Figure 4.1**). This is consistent with the hypothesis that DOCK hit lists may serve in lieu of known-ligand sets as a means to represent a protein target—and that the similarity among such targets may likewise be amenable to quantification by SEA.

Target key

ACE	angiotensin-converting enzyme	HIVPR	HIV protease
AChE	acetylcholinesterase	HIVRT	HIV reverse transcriptase
ADA	adenosine deaminase	HMGA	hydroxymethylglutaryl-CoA reductase
ALR2	aldose reductase	HSP90	human heat shock protein 90
AmpC	AmpC beta-lactamase	InhA	enoyl ACP reductase
AR	androgen receptor	MR	mineralocorticoid receptor
CDK2	cyclin-dependent kinase 2	NEUA	neuraminidase
COMT	catechol O-methyltransferase	P38	P38 mitogen activated protein
COX1	cyclooxygenase-1	PARP	poly(ADP-ribose) polymerase
COX2	cyclooxygenase-2	PDE5	phosphodiesterase 5
DHFR	dihydrofolate reductase	PDGFrb	platelet derived growth factor receptor kinase
EGFr	epidermal growth factor receptor	PNP	purine nucleoside phosphorylase
ER_agonist	estrogen receptor (agonist conformation)	PPAR	peroxisome proliferator activated receptor gamma
ER_antagonist	estrogen receptor	PR	progesterone receptor
FGFr1	fibroblast growth factor receptor kinase	RXR	retinoic X receptor alpha
FXa	factor Xa	SAHH	S-adenosyl-homocysteine hydrolase
GART	glycinamide ribonucleotide transformylase	SRC	tyrosine kinase SRC
GPB	glycogen phosphorylase beta	TK	thymidine kinase
GR	glucocorticoid receptor	VEGFr2	vascular endothelial growth factor receptor

Figure 4.1 SEA similarity matrix for forty docking hit lists

Initial foray into using SEA to quantify the similarities among docking hit lists, circa May 2008. Each target was represented by its top 500 hits from DOCK, as run previously by Niu Huang.[4] In the matrix (which is symmetric about the diagonal), darker blue cells indicate stronger SEA E-values between docking hit lists, and white indicates no appreciable similarity. *Caveat lector*—while these data were convenient for initial proof of concept testing, they are likely not the most appropriate collection of docking hit lists to use, as they are from an early version of the Directory of Useful Decoys (DUD, http://dud.docking.org). DUD was designed as a DOCK benchmarking tool, and the consequences for its use in 2D similarity searches with SEA are unclear.

Nevertheless, this work is only beginning and significant questions remain. In **Figure 4.1**'s matrix, a large number of targets show at least weak similarity to each other; is this informative or just noise? What database of DOCK molecules would be optimal, for use on a truly broad scale irrespective of protein class? Fragments may be appropriate (and would be quick to dock), but focused libraries may give better discrimination, where functional starting points could be inferred from the protein's sequence or fold. In terms of focused libraries, Johannes Hermann docked high-energy intermediates (HEI) rather than ground states, but we cannot expect to exhaustively compute such intermediates across a fragment library in any general purpose manner.

Perhaps a broader question is whether this approach should work at all, when we expect the majority of putative ligands selected by docking to fail the ultimate test of binding. I would argue, however, that whether or not DOCK is actually *correct* in its prediction of ligand binding is irrelevant. As a component of the "protein function identification engine" proposed here, DOCK need only be *consistent* in its predictions. For this use, docking hit lists are merely computational "signatures" for entire protein active sites—ones that are motivated by binding. In such signatures, even pathological errors are statistical signal, as long as they are deterministic.

II. Approach 2: Hit lists vs. known ligands

Hypothesis: *For two pharmacologically similar proteins, the first protein's docking hit list is similar by SEA to the second protein's set of known binders.*

This second approach to protein function prediction differs from the first in only one respect: Rather than comparing docking hit lists entirely among themselves, we would compare each to sets of "ideal" data; namely, to sets of well-defined, validated, known ligands for each reference protein, such as were used in earlier Chapters. Those concerned with the fidelity of virtual screening's binding predictions may find reassurance in this refuge to a "gold standard" dataset. Those considering the proposal in terms of information signal may take pause. While we would expect this approach to increase the signal-to-noise ratio in an absolute sense, i.e., with respect to each target's *true* ligands, it can no longer benefit from the presence of consistent error in DOCK—and we know that such error exists.[**]

Preliminary results comparing docking hit lists to MDDR activity classes reveal strong SEA similarities in only a limited number of cases (**Figure 4.2**). These successes may be where DOCK runs achieved greater enrichment for that target's true ligands. Alternatively, the ligand type characteristic of these targets may simply have contained more structural signal—such as the presence of an adenosine scaffold, which is a strong topology signal by SEA. This remains an open question.

[**] Although the reader may note that we do not know that such error is in fact consistent, and this I concede.

Figure 4.2 SEA similarities between docking hit lists and matching MDDR ligand sets

First attempts at using SEA to quantify the similarity between a docking hit list and its corresponding MDDR target(s). Docking hit lists are arranged horizontally and MDDR activity classes vertically. The matrix is clustered by SEA similarity, where darker red cells denote stronger SEA E-values. Targets with strong patterns in their ligands, such as those containing adenosine substructures or folates are predicted well by this approach. Many others are not.

4.2 Should similarity negate novelty?

Chemical similarity as a technique seems shackled to the past. It is powerful because it largely circumvents the need for a full understanding of the physical principles of macromolecular binding—in fact, one could imagine successful chemical similarity campaigns prosecuted in the complete absence of any information about the actual protein target. It is weak for the same reason: It does not encode a full understanding. Indeed, how could a similarity method infer a truly novel chemical scaffold for a target or disease class?

A challenge for similarity methods has long been the choice of the "similar-enough" cutoff[5]—set it too low and predictions are too noisy, too high and they are trivial. This puts pragmatism and novelty at odds. SEA may offer two advances here; in the first, the similarity cutoffs that it uses are more inclusive than would be practical for direct ligand-ligand similarity. For Daylight and Tanimoto coefficients, we find 0.5-0.6 Tc to be optimal, compared to the 0.85-0.9 Tc industry standard.[5] The second is more abstract. As a strong SEA E-value denotes only the presence of set-wise similarity that would be unlikely by random chance alone, it does not require that any particular ligand pair be very similar to another. A weak-but-prevalent similarity between sets is sufficient for a strong E-value. This may allow for semi-random recombination of recurrent ligand topology patterns, and make the method more robust to minor perturbations. This gives us a chance at novelty after all.

One way to leverage this property systematically would be to focus on strong SEA predictions where the very best ligand-ligand match between sets is poor; these cases necessarily demonstrate weak yet broad-based similarities. We have not comprehensively attempted this. However, the following case study demonstrates how we have used SEA to find a novel ligand scaffold. Along with Amanda DeGraw in Dr. Mark Distefano's group, we are preparing this

study for submission. I provide here a narration of the computational side of the story, along with some preliminary results from Amanda to illustrate its biological motivations.

4.3 Case study: Prediction and evaluation of protein farnesyltransferase inhibition by commercial drugs

Amanda J. DeGraw[1], Michael J. Keiser[2], Brian K. Shoichet[2], Mark D. Distefano[1]

(1. Department of Chemistry, University of Minnesota, Minneapolis, MN 55455; 2. Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143-2550.)

I. Abstract

One of the major obstacles to overcome in pharmaceutical development is predicting compound activity *in vivo*. Often drugs have unexpected effects due to their undesirable interaction with other receptors or biochemical pathways. These unexpected activities can be harmful, leading to toxicity, or beneficial, suggesting new therapeutic indications. The Similarity Ensemble Approach (SEA) is a program that relates proteins based on the set-wise chemical similarity among their ligands.[6] It can be used to rapidly search large compound databases and to build cross-target similarity maps. The emerging maps relate targets in ways that reveal relationships one might not recognize based on sequence or structural similarities alone. SEA was used to look for potential off-target activity of commercially available drugs. Two families of compounds emerged as potential inhibitors of protein farnesyltransferase (PFTase): Desloratadine-based H_1 receptor antagonists and azole antifungals. Here we present the evaluation of two common drugs, Loratadine and Miconazole, and their structural analogues for off-target PFTase inhibition activity.

II. Preliminary results

One startling result of this work is that even commercially marketed drugs may have unexpected but biologically relevant activity at human enzymes—and that these "off-target" activities can be predicted computationally. Protein farnesyltransferase (PFTase) catalyzes protein prenylation, and has emerged as a cancer target due to the high prevalence of mutated Ras oncogenes in human tumors. The Similarity Ensemble Approach (SEA) has previously uncovered novel off-target drug activity among drugs that target aminergic G-protein coupled receptors (GPCRs),[6] and we ask here how prevalent such off-target activity may be among drugs that target enzymes. In this work, we have focused on PFTase, using SEA to compare 746 commercial drugs against ligand sets built from the 1,640 known non-peptide PFTase ligands reported in the literature.

But to do so, we must first ask what affinity constitutes relevant off-target activity at human PFTase. Many known inhibitors have 10-20 µM affinity for this enzyme. To be comprehensive, we considered three thresholds of PFTase inhibitor (FTI) affinity, each at increasingly greater stringency. In the first instance, we considered all those 1,692 FTIs known to have 100 µM or better affinity, reasoning that this would allow for the greatest breadth of predictions. We then narrowed our focus to include only those 1,423 FTIs with 10 µM or better affinity, and finally excluded all but the 1,188 FTIs with affinity $\leq$ 1 µM. We considered each threshold independently, and later extracted each commercial drug's best SEA match with the set of known PFTase ligands at any of the three thresholds. For example, Loratadine matched most strongly against the FTIs known to have $\leq$ 10 µM for their target, with a SEA expectation value (E-value) of 1.07×10^{-81} (**Table 4.1**). On the other hand, Ubenimex was most similar to the $\leq$ 100 µM FTIs, with an E-value of 1.53×10^{-16} for them (**Table 4.1**), compared to weaker E-values of 4.97×10^{-13} and 7.23×10^{-9} for its similarity against the $\leq$ 10 µM and $\leq$ 1 µM FTIs, respectively (**Table 4.2**). An E-value, much like a p-value, denotes the likelihood that a particular

130

event—in this case the degree of chemical similarity for a commercial drug against the set of ligands for protein farnesyltransferase—would have been found that strong or better by random chance alone. When applied across all 746 commercial drugs, this analysis found 13 of them (comprising 1.9% of the total drugs screened) to have measurable similarity to at least one of the three FTI sets (**Table 4.1**).

Table 4.1 Top SEA predictions of off-target PFTase binding for commercial drugs
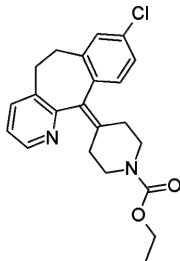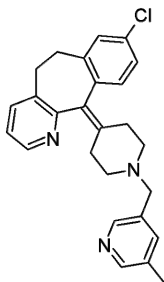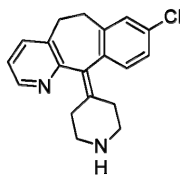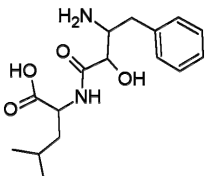
	Drug	Best SEA E-value	Best FTI Match
1	Loratadine	1.07×10^{-81}	10 µM
2	Rupatadine	1.10×10^{-49}	10 µM
3	Desloratadine	1.22×10^{-30}	10 µM
4	Ubenimex	1.53×10^{-16}	100 µM
5	Azatadine	2.68×10^{-11}	100 µM
6	Phenylalanine S	1.70×10^{-4}	100 µM
7	Miconazole	2.00×10^{-4}	100 µM
8	Diazepam	5.52×10^{-4}	1 µM
9	Temazepam	1.21×10^{-3}	1 µM
10	Thymopentin	2.10×10^{-3}	100 µM
11	Cortisone acetate	6.57×10^{-3}	100 µM
12	Prednisone	3.81×10^{-2}	100 µM
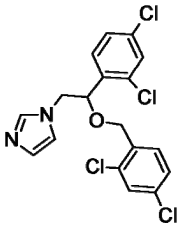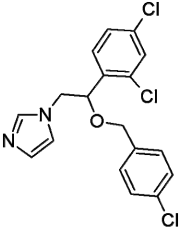
Drugs in blue have PFTase activity already reported in the literature. Drugs in gray were either peptides or unavailable for purchase.

Of the thirteen commercial drugs predicted to have off-target PFTase binding by SEA (**Table 4.1**), eight already had literature precedent or could not be tested (see **Methods**). Three of the five remaining predicted FTIs had low-to-mid micromolar affinity for PFTase in our binding assays (**Table 4.2**). The five commercial drugs comprised three histamine H_1 antagonists (Loratadine, Desloratadine, and Rupatadine), an antineoplastic (Ubenimex), and an azole antifungal (Miconazole). Of these, a subset of the antihistamines bound PFTase, as did the

antifungal (**Table 4.2**). For all but Ubenimex, the 100 µM and 10 µM FTI sets yielded the

strongest SEA predictions, with little difference in prediction strength between the two affinity

classes (**Table 4.2**). This was consistent with their PFTase IC_{50} values, which we found to be

between 20-80 µM—with the exception of Desloratadine and Ubenimex, which did not bind

PFTase up to 100 µM and 200 µM, respectively.

Table 4.2 Predicting and testing PFTase binding against known FTIs

Drug	FTIs by affinity threshold			IC_{50} (µM)
	1 µM	10 µM	100 µM	
Loratadine **1**	7.87×10^{-53}	$\mathbf{1.07 \times 10^{-81}}$	$\mathbf{1.53 \times 10^{-81}}$	13.3 ± 1.8
Rupatadine **2**	5.90×10^{-41}	$\mathbf{1.10 \times 10^{-49}}$	$\mathbf{8.15 \times 10^{-49}}$	76 ± 18
Desloratadine **3**	3.45×10^{-27}	$\mathbf{1.22 \times 10^{-30}}$	$\mathbf{1.83 \times 10^{-30}}$	> 100
Ubenimex **4**	7.23×10^{-9}	4.97×10^{-13}	$\mathbf{1.53 \times 10^{-16}}$	> 200

	Miconazole				
5		$8.26 \times 10^{+2}$	**2.86×10^{-3}**	**2.00×10^{-4}**	18.9 ± 3.6
	Econazole				
6		N/A	N/A	N/A	23.3 ± 2.0

Listing of SEA E-values denoting the statistical significance of each drug's similarity to known protein farnesyltransferase inhibitors (FTIs). The closer the E-value approaches to zero, the more significant the similarity; the strongest prediction for each drug is bold. Each drug was compared against three different sets of known FTIs. For instance, for the "1 µM FTIs" set, we only considered those FTIs known to have 1 µM affinity or greater for PFTase. For the "10 µM FTIs" set, we considered all FTIs known to have 10 µM or greater affinity for FT, etc. Where a drug's PFTase predictions by SEA are very close (within approx a single order of magnitude), both E-values are bolded. Note also that fluconazole, ketoconazole, clotrimazole, & TIPT did not inhibit PFTase up to 200 µM, and none of these were predicted to do so by SEA.

III. Discussion

Whereas much of the previous drug cross-talk predicted by SEA focuses on drugs that target aminergic GPCRs,[6] Loratadine and Miconazole were predicted to bind PFTase in defiance of these traditional target-class boundaries. Loratadine represents one of the first uses of this approach to demonstrate that a commercial drug thought to bind only a GPCR also binds an enzyme, and Miconazole represents the approach's first enzyme-enzyme cross-talk prediction in commercial drugs. Both of these drugs not only inhibit PFTase *in vitro* (**Table 4.2**, assay data not

shown), but also disrupt localization of H-Ras to the cell membrane (**Figure 4.3**), consistent with inhibition of PFTase *in vivo*.

a DMSO Control
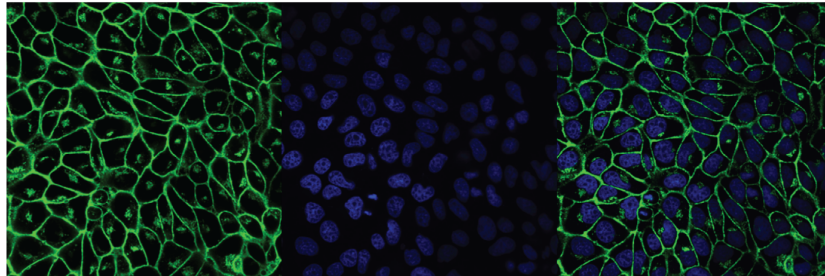
b Claritin 24 hours, 25 uM
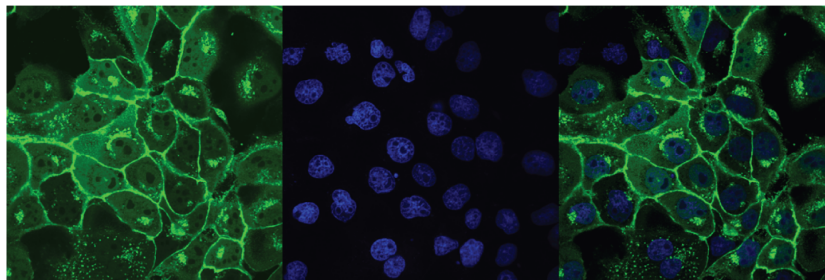
c Miconazole 24 hours, 10 uM
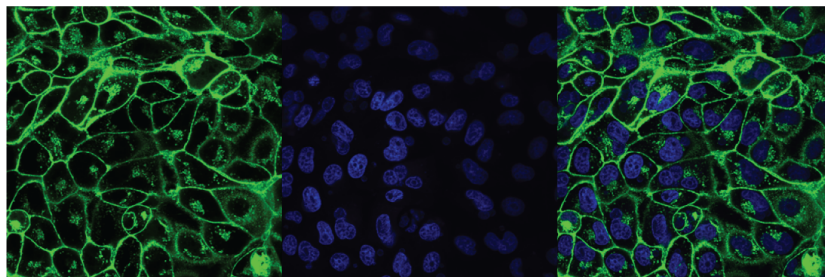
d Miconazole 48 hours, 10 uM
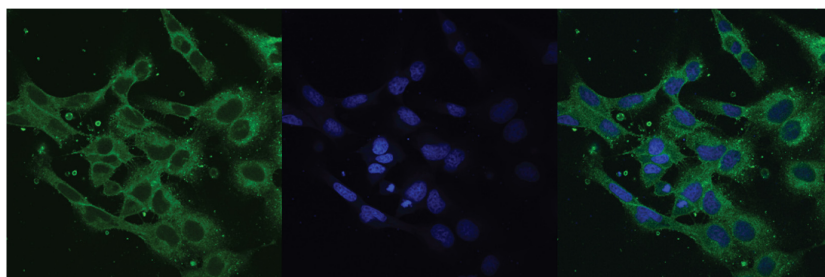
Figure 4.3 Claritin and Miconazole as inhibitors of PFTase *in vivo*

Mark Distefano's group has a MDCK cell line that expresses a GFP-H-Ras chimera, which is used here to visualize H-Ras processing. H-Ras's normal localization to the cell membrane is dependent on its prenylation by PFTase. These images are preliminary results from Amanda DeGraw, demonstrating that (b) Loratadine (Claritin) and (c-d) Miconazole markedly disrupt normal localization (a) of the GFP-H-Ras chimera to the cellular membrane—consistent with their inhibition of PFTase observed *in vitro* by Amanda by fluorescent and HPLC assay (not shown). Note that Miconazole only successfully does so after 48 hours, and this is accompanied by some cell death.

Miconazole (Monistat) and Econazole (Spectazole) are topical imidazole antifungals that increase cell membrane permeability in fungi, resulting in leakage of cellular contents. Both of these anti-fungals interact with 14-α demethylase, a cytochrome P-450 enzyme necessary to convert lanosterol to ergosterol, which is an essential component of cell membranes.[7] By SEA, we found that Miconazole had weak yet significant similarity to the set of 100 μM PFTase inhibitors, with an E-value of 2.00×10^{-4} (**Table 4.1**). Upon validating Miconazole's 19 μM IC_{50} for PFTase, we also tested Econazole, which has a highly similar chemical structure, and found it to have 23 μM IC_{50} for this target (**Table 4.2**). It is intriguing to consider that these interactions with protein farnesyltransferase may complement the drugs' anti-fungal activity, as farnesyl is necessary to secure Ras to cell membranes (**Figure 4.3.d**). This cross-talk may suggest new directions for anti-fungal development.

Loratadine (Claritin) is a second-generation anti-histamine that binds the H_1 receptor—a cross-membrane target that shares no evolutionary history, functional role, or structural similarity with the enzyme PFTase. We found Loratadine to have an exceptionally strong E-value of 1.07×10^{-81} for PFTase ligands in the 10 μM – 100 μM range (**Table 4.2**), and SEA also predicted strong scores for two of its analogs, Rupatadine and Desloratadine (1.10×10^{-49} and 1.22×10^{-30}, respectively, **Table 4.2**). These off-target predictions were confirmed for Loratadine and Rupatadine, at 13 μM and 76 μM affinities, whereas Desloratadine showed no binding up to

100 μM. We were subsequently able to further confirm Loratadine's role in FTI development.[8] Rupatadine (Rupafin), however, is known to be active only at the histamine H_1 and platelet-activating factor (PAF) receptors, and has shown good safety profiles in prolonged treatment periods.[9, 10] The presence of such off-target drug activity across GPCR and enzyme class boundaries even among well-studied commercial drugs suggests new opportunities for both drug development and side-effect management.

IV. Methods

A. Sources of known protein farnesyltransferase ligands

We used several subsets of the known protein farnesyltransferase (PFTase) inhibitors as our reference sets. To do so, we built the set of known ligands corresponding to each major drug target in the literature extracted from the World of Molecular BioAcTivity (WOMBAT) 2006.2 database, as in previous work.[11] After removal of duplicates, molecules that we could not process, and ligands with affinities worse than 100 μM for their protein targets, this database comprised 169,046 molecules annotated into 1,469 target-function sets (e.g., the PFTase *inhibitors* and the PFTase *binders* of unknown function comprised two distinct sets).

We then extracted the 1,723 molecules from this collection that were annotated as PFTase inhibitors (1,648 molecules) or PFTase binders (75 molecules), and filtered out all molecules containing two sequential peptide bonds along a standard peptide backbone, using a SMARTS filter in Scitegic PipelinePilot. We further subdivided these PFTase binders (collectively termed "FTIs" or "FTI sets" in main text) by their affinities for PFTase, into 100 μM, 10 μM, and 1 μM FTI sets, containing 1,692 molecules, 1,423 molecules, and 1,188 molecules, respectively. The remaining 1,467 target-function sets from WOMBAT were not considered in this analysis.

B. Collection of commercial drugs

We extracted all molecules annotated as marketed drugs in the WOMBAT 2006.2 database, and processed them as above (excepting peptide and 100 μM affinity filtering). This yielded 746 commercial drugs, each of which we screened individually against the FTI sets using SEA.

C. Choice of drugs for testing

We excluded Diazepam and Temazepam because Diazepam's weak PFTase activity is already reported,[12] and the steroids because Cortisone's and Prednisone's PFTase activities are likewise known.[13] We could not obtain Azatadine, and did not test Thymopentin or Phenylalanine-S because they are peptides. We excluded both the known peptide FTIs and the predicted peptide drugs because SEA's statistical models were built using small-molecule drug chemical similarity descriptors. As peptides contain oft-repeated chemical patterns in their backbones—and thus strong opportunities for *uninformative* similarity—they may have the potential to skew small-molecule similarity models if included. We nonetheless tested one peptide prediction, Ubenimex, because it appeared highly similar to a known FTI.

D. Similarity measures

We used 1024-bit folded Scitegic ECFP_4 topological fingerprints as previously described.[11] Although we later tested 2048-bit Daylight[6, 11] fingerprints, these resulted in a narrower and weaker subset of the PFTase predictions found via ECFP_4, and are not reported here. As before, we used Tanimoto coefficients to calculate pair-wise similarity between fingerprints.[6, 11]

E. Predictions of PFTase binding using the Similarity Ensemble Approach (SEA)

We ran SEA as previously described.[6] The query collection consisted of the 746 commercial drugs, each drug as its own "set" of one molecule. The reference collection comprised the three overlapping sets of PFTase ligands ("FTIs"), binned into each set at (a) 1 µM, (b) 10 µM, or (c) 100 µM affinity thresholds. A FTI set with a weaker affinity threshold, such as 100 µM, comprised an all-inclusive superset of those sets at stronger affinity threshold (both the 1 µM and 10 µM sets, in this example). After each drug was compared independently against each of

the three FTI sets by SEA, their E-values were compared (e.g., see **Table 4.2**), and all

commercial drugs with measurable best-match E-values across sets were reported (**Table 4.1**).

4.4 The chemical SEA analytical

Substantial opportunity yet exists for improvement of SEA's theoretical underpinnings. For instance, its current statistical model is entirely empirical. To define the random background of chemical similarity for any given molecule collection, SEA samples the distribution explicitly. While this is eminently tractable, thanks to excellent bitwise fingerprint-comparison libraries adapted years ago by Michael Mysinger, it also begs the question, what are the analytical principles underlying these distributions? Indeed, certain patterns *always* occur in these SEA models, regardless of choice of molecule collection, molecular representation, or comparison coefficient. I endeavor here to highlight some of these patterns and my thoughts as to their implications. I do not yet know why they occur—but perhaps you, gentle reader, may be more fortunate.

I. Background distribution shapes and a Tc_{50}

SEA corrects for the "uninformative" random chemical similarity in any target-target or drug-target chemical similarity comparison. It does so with respect to two parameters. For the first, recall that we calculate similarity with respect to a particular molecule collection such as the MDDR or WOMBAT (see Chapter 2 for descriptions of these databases). This collection defines the "chemical space" of our comparison, and is roughly analogous to the "non redundant" sequence collection in BLAST. The second parameter is our choice of Tanimoto coefficient threshold above which ligand-ligand scores are included in the "raw score" (see Appendix B for details).

One way to choose a threshold is to determine how well the distribution of z-scores that results from it resembles an Extreme Value distribution (desired), as compared to how well it

resembles a Normal distribution (undesired). For instance, see **Figure A.1.5** in Appendix A. Consider the goodness-of-fit chi-square values for the EVD in particular (red); it worsens at a threshold choice of 0.2 Tc, then undergoes slow improvement until it bottoms out near 0.55 Tc. We consequently chose a 0.56 Tc cutoff, based solely on the argument that the best fit to an EVD was ideal.[††] Oddly, this general pattern for an EVD—initially poor fit, bump around 0.2 Tc, improvement thereafter followed by slow worsening—is present in *all* such SEA goodness-of-fit plots I have ever made. I do not think it is merely a property of the molecular fingerprint or comparison coefficient; I have encountered it across many fingerprints, and also with the Dice coefficient instead of Tanimoto (data not shown).

What if we leave aside the distribution shapes, and instead consider only the simple fits to means and standard deviations of random chemical similarity (**Figure 4.4.a**)? Intriguingly, these fit parameters follow familiar patterns, and this is sensible because these fits underlie the z-score distributions. Of particular note is a SEA threshold choice of 0.2 Tc, where both mean and standard deviation coefficients undergo a substantial shift (**Figure 4.4.a**). This correlates with the earlier observation that most random ligand-ligand pairs score in the 0.2 Tc region (as mentioned in the Introduction). Compare also the shape of either standard deviation coefficient curve (cyan or blue line) to the goodness-of-fit EVD distribution described above (**Figure A.1.5**, red line). These curves all experience substantial shifts in the 0.2 Tc region.

[††] See following subsection (**4.4.II**) for subsequent empirical results consistent with this hypothesis.

a MDDR2006.1 and WOMBAT2006.2 Fits

Coefficients

Exponents

W mu coeff
M mu coeff
W sig coeff
M sig coeff
W sig exp
M sig exp

Tc Threshold

b MDDR2006.1 Fits Over Random Data

Avg Raw Score at 1

Avg Raw Score at 100

avg at 1
avg at 100

Tc Threshold
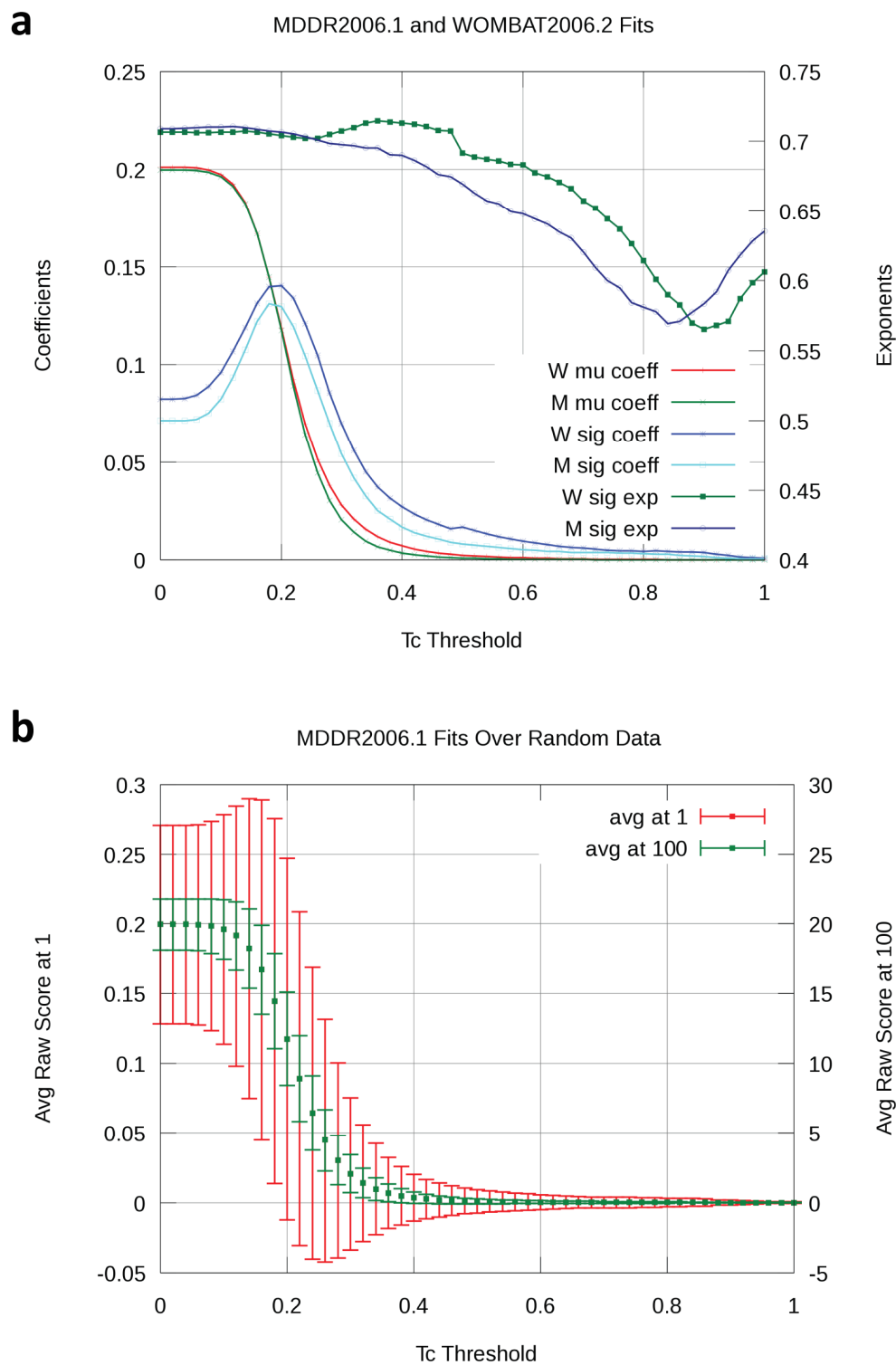
Figure 4.4 Patterns in random background fits

(a) Summary plot of optimal coefficients and exponents for random background model fits, using Daylight fingerprints and Tanimoto coefficients. If the mean ("mu," left y-axis) random chemical similarity is fit to the simple formula $y = \mathbf{m} \times x + b$, and the standard deviation ("sig," right y-axis) to $y = \mathbf{m} \times x^{\mathbf{e}} + b$, then $\mathbf{m}$ is the coefficient ("coeff") and $\mathbf{e}$ is the exponent ("exp"). See Appendix B for a description of background model calculation. The fits are shown for two different molecule collections, the MDDR (M) and WOMBAT (W). See Chapter 2 for descriptions of these collections. (b) Excerpt view of average random background raw scores for MDDR only, at product-set-size = 1 (red, left y-axis) and product-set-size = 100 (green, right y-axis). As the product of set sizes grows, the absolute value of the random expected chemical similarity grows too, while the proportionate standard deviations shrink.

Likewise, the raw scores over which the background fits are defined also demonstrate the most interesting behavior at 0.2 Tc. In a striking shift visually analogous to an IC_{50} curve (**Figure 4.4.b**), average random chemical similarity scores transition from a stable baseline when Tc thresholds are below 0.2 Tc to a new one that approaches zero asymptotically. Again, 0.2 Tc is the inflection point of this shift; we could consider it to define a "Tc_{50}" for this combination of molecule collection, representation, and comparison coefficient.[‡] What would such a Tc_{50} tell us? Is it merely representative of the most over-represented scores for random molecule comparisons? But then why is it so consistent ($Tc_{50} \approx 0.2$ Tc)?

II. How well does background model theory correlate with empirical success?

As alluded to in the previous subsection, SEA's background model depends essentially on only two parameters: (1) the choice of molecular database to represent "chemical space," and (2) the choice of Tanimoto coefficient (Tc) threshold, above which pair-wise scores contribute to the uncorrected "raw score." In practice, we choose our Tc threshold to achieve the best fit of the

---

[‡] Although the x-axis scale is linear, not logarithmic.

random chemical background to an EVD (Chapter 1, Appendix B). While conceptually satisfying, however, does this approach actually work in practice?

The answer, in my experience, is yes. In extensive retrospective leave-one-out and *k*-fold cross-fold validation experiments across large datasets such as MDDR or WOMBAT, we find that the Tc threshold chosen *solely by SEA's ability to correctly assign ligands to their annotated target sets* converges on the same threshold as that chosen by the background distribution's best fit to an Extreme Value distribution (**Figure 4.5**). This seems a resounding validation of the random background model theory—the threshold derived from theory alone independently converges with that derived from empirical performance tests.

MDDR2006.1 Schuf Single-Molecule 3-fold AUCs
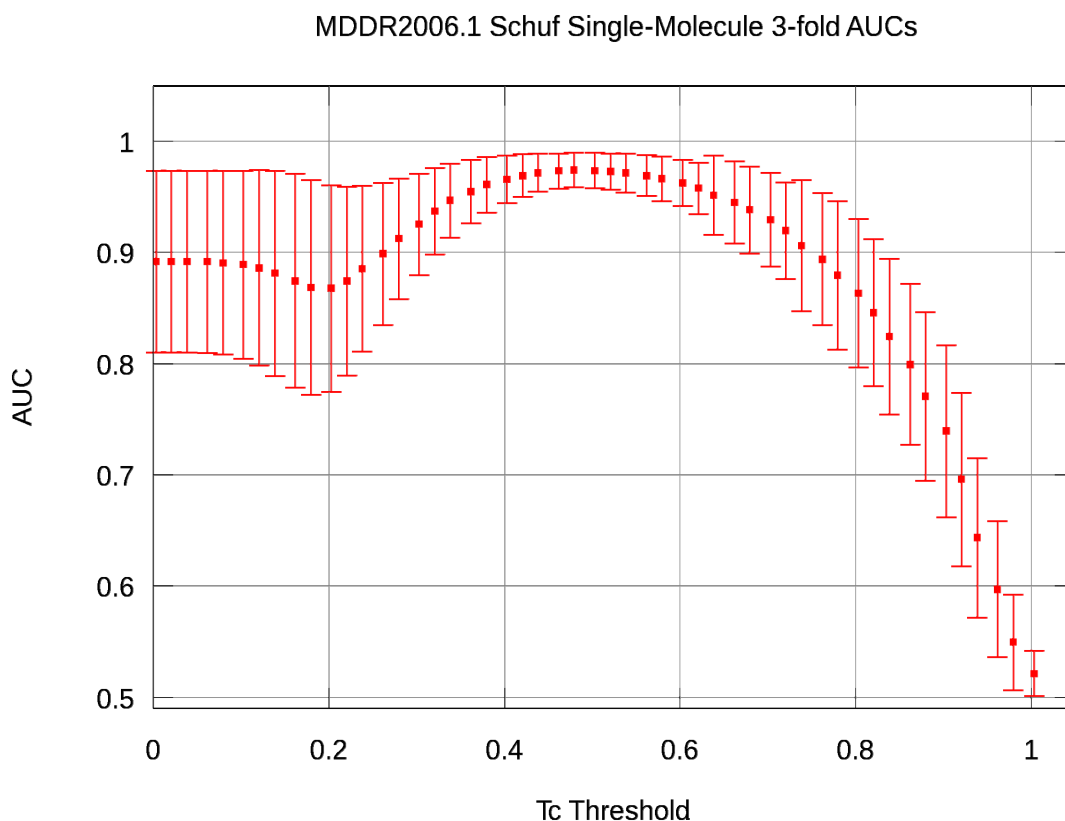
Figure 4.5 Dependence of SEA's ability to recapitulate known ligand annotations on the choice of raw score Tc threshold

Summary of experiment wherein a full random background model is built for SEA at each of the Tc thresholds shown, and SEA is subjected to rigorous k-fold validation testing on the MDDR using that background mode. In the k-fold validation process, each MDDR activity class is randomly broken up into k slices, the first of which are fused into a single test set and the remainder of which become the training set. Each molecule in the test set is compared against the training set and also against all other 244 MDDR "decoy" sets. The sensitivity and specificity rates for these SEA predictions (e.g., how often did it assign the molecule back to its correct activity class) are condensed into a single receiver-operating characteristic (ROC) area-under-the-curve score (AUC). For this AUC, a score of 1.0 means all predictions were perfect, and 0.0 means none we perfect; a score of 0.5

AUC is equivalent to a disgruntled monkey randomly choosing targets with its eyes closed. Each such experiment is then repeated in full for each of the $k - 1$ other slices for that annotation, and each AUC stored. This is then repeated over all 245 activity classes in the MDDR, for a total of $245 \times k$ AUCs calculated *per Tc threshold on the x-axis*. In this plot, $k = 10$. Each data point thus represents the average and standard deviation of approximately 2,500 retrospective AUCs, comprised of $> 65,000$ SEA predictions, or more than 3 million predictions in total.

Note that this experiment was run using ECFP4 instead of Daylight fingerprints, and thus the optimal threshold choice (higher AUC) is closer to 0.44 Tc than the 0.56 Tc optimal threshold of **Figure A.1.5**, which used Daylight.

One future direction arising from this may be to determine and prove how optimization for EVD background distributions results in this better discrimination ability. Secondly, as we have two objective means of determining whether we are setting our parameters appropriately—namely, the theoretical goodness-of-fits and the quantifiable retrospective testing—could we benefit from introducing and evaluating additional parameters into the SEA background models? After all, it is not strictly true to say SEA's background depends only on the two parameters already mentioned; rather, certain "parameters" are set implicitly by design decisions we have made in SEA's development. The description and early exploration of one such reconsidered design decision follows below.

III. How should we weight Tanimoto coefficients?

In all prior uses of SEA described in this thesis or published elsewhere, we have made the implicit decision to weight each Tanimoto coefficient linearly. In each raw score, then, a ligand-ligand score of 0.25 Tc is worth half of one that is worth 0.5 Tc. But is this sensible? Or should we give higher scores extra emphasis? One way to do this would be to consider raising the Tanimoto coefficients to some power, such as squaring them, so that higher scores would count for more. In fact, if we could find the right formula to do this, perhaps we could dispense with step-function Tc thresholds entirely; after all, why should a Daylight ligand-ligand pair-wise score of 0.56 Tc count towards the raw score in full, when a 0.54 Tc counts not at all? Of course, it is not clear what this formula, or weighting function, *should* be—and this remains a challenge.

One simple approach to test this hypothesis is to test it with several simple functions that make some intuitive sense. To this end, we tested a power-based Tanimoto weighting function; i.e., "weighted Tc" = Tc^x, where $2 \leq x \leq 6$. Results for $4 \leq x \leq 6$ background model fits (**Figure 4.6.a**) and their retrospective cross-fold testing (**Figure 4.6.b**) are shown here. Again, theoretical and empirical results reliably converge; both find $x = 4.5$ to be the most optimal choice of power-weighting (compared to the implicit default of $x = 1$, when Tc weighting is linear). This may imply that the choice of Tanimoto weighting merits further investigation.[§§]

[§§] *Caveat lector*—any attempt to add new parameters to an existing model must ultimately take the data's existing degrees of freedom into account, but I do not fully address this here.

a

MDDR2006.1 Distribution Fits by Tc Powers

b

MDDR2006.1 Schuf 3-fold tests: Choice of Tc exponent
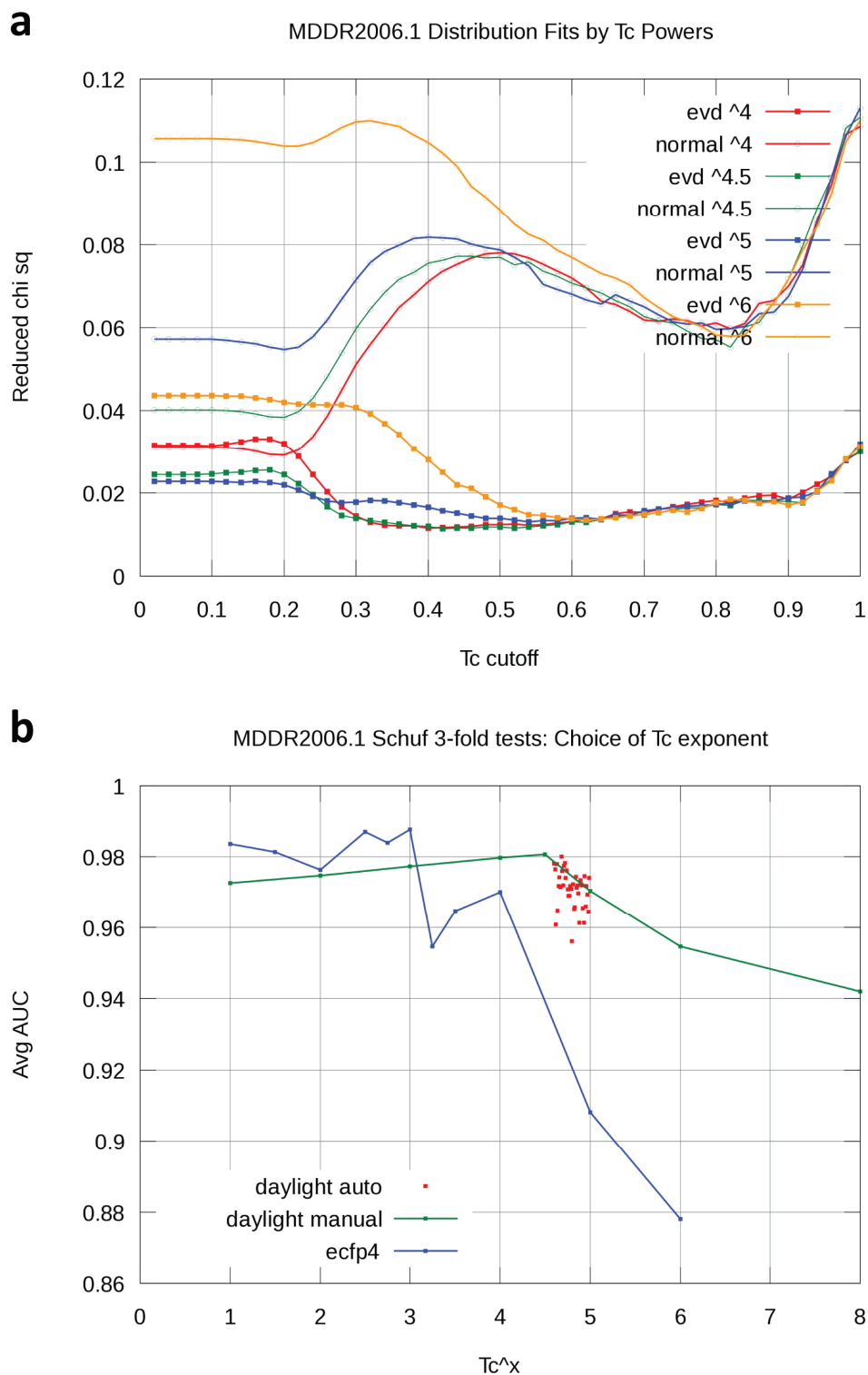
Figure 4.6 Evaluation of power-weighting in SEA raw scores

(a) SEA random chemical background distributions are calculated against the MDDR, using Daylight fingerprints, at four different choices of power weighting. E.g., "evd ^ 4" is the goodness-of-fit for an extreme value distribution, where each Tc score is first raised to the 4th power before being summed into the raw score. Note that power = 4.5 (i.e., $Tc^{4.5}$) yields the best (lowest chi-square) EVD fit. (b) Empirical k-fold testing ($k = 3$) of the same Tc-power choices, again confirming that the $Tc^{4.5}$ weighting is optimal for Daylight, as measured by its high average AUC. The red dots show several individual datapoints for AUCs automatically calculated to sample this 4.5-5.0 power region in greater detail.

4.5 Prediction is very hard, especially about the future

With a nod to Yogi Berra, this closing section comprises a brief overview of several immediate

challenges and opportunities that remain, to my mind, incompletely addressed. It is likely that

SEA's most intriguing implications and future developments will be precisely those that I do not

expect, and hence have found no place in these pages—but I present the following directions

nonetheless.

I. Weighted set membership and K_i

In a SEA set, all ligands are given equal importance. But clearly this is not the case; perhaps we

could weight ligand-ligand scores by the logs of their K_i's. But then what is the appropriate

random background?

II. Molecule representations

SEA typically uses standard fingerprints (Daylight or ECFP4) and similarity coefficients

(Tanimoto, and a few tests of Dice coefficients). Jérôme Hert has compared the performance of

several such fingerprints for SEA,[11] but many more molecular representations remain untested.

For instance, what of a fingerprint that encodes a DOCK pose?

III. Toxicity, transport, and metabolism

We have used SEA to uncover potential side effect mechanisms, where these are mediated by

particular protein targets, in Chapters 1 and 2. But can we use it to identify chemical topology

patterns for toxicities? Or would these be better addressed by explicit pharmacophores? What of

a ligand's susceptibility to active transport, when this may be subject to general physical

properties? Additionally, could we use SEA to determine a drug's susceptibility to enzymatic metabolism?

IV. A foolish consistency

Not all ligand sets have the same self-similarity by SEA; some have greater internal "consistency." Could we deconvolute mere historical bias, such as that from medicinal chemistry optimization, from fundamental conclusions about a protein target's binding patterns? Would greater consistency also reflect greater success in ligand prediction, in a docking hit list for instance? Previous efforts in this direction have failed; perhaps indeed "A foolish consistency is the hobgoblin of little minds" (Ralph Waldo Emerson).

V. Targets of phenotypic screens

Can we use SEA to identify the targets of promising molecules identified in phenotypic screens? Early results suggest so. How applicable are these successes, however, to the broad scale, and to what extent could we automate the process?

VI. Sequence and structure and SEA

We often highlight the difference between SEA's ability to link protein targets by a *pharmacological* logic compared to the evolutionary logic that motivates protein sequence comparisons. Yet clearly these forms of orthogonal information could be combined to mutual benefit. Where the approaches disagree, is it because one method achieves an informative discrimination or rather because it fails to make a true association that the other finds? How could we combine them?

4.6 References

1. Lorber, D.M. & Shoichet, B.K. Flexible ligand docking using conformational ensembles. *Protein Sci* **7**, 938-950 (1998).

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

3. Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. & Perry, K.M. Structure-based discovery of inhibitors of thymidylate synthase. *Science* **259**, 1445-1450 (1993).

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

5. Brown, R.D. & Martin, Y.C. Use of structure activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comp Sci* **36**, 572-584 (1996).

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

7. DRUGDEX (Thomson Reuters, Greenwood Village, Colorado; 2008).

8. Njoroge, F.G. et al. Discovery of novel nonpeptide tricyclic inhibitors of Ras farnesyl protein transferase. *Bioorg Med Chem* **5**, 101-113 (1997).

9. Picado, C. Rupatadine: pharmacological profile and its use in the treatment of allergic disorders. *Expert Opin Pharmacother* **7**, 1989-2001 (2006).

10. Mullol, J. et al. Rupatadine in allergic rhinitis and chronic urticaria. *Allergy* **63 Suppl 87**, 5-28 (2008).

11. Hert, J., Keiser, M.J., Irwin, J.J., Oprea, T.I. & Shoichet, B.K. Quantifying the relationships among drug classes. *J Chem Inf Model* **48**, 755-765 (2008).

12. Roskoski, R., Jr. & Ritchie, P.A. Time-dependent inhibition of protein farnesyltransferase by a benzodiazepine peptide mimetic. *Biochemistry* **40**, 9329-9335 (2001).

13. Lingham, R.B. et al. Clavaric acid and steroidal analogues as Ras- and FPP-directed inhibitors of human farnesyl-protein transferase. *J Med Chem* **41**, 4492-4501 (1998).

Appendix A:

Supplementary figures and tables

A.1 Supplementary material for Chapter 1

I. Supplementary methods

I.A Transformation of Raw scores to Z-scores

A raw score was transformed into a z-score by taking its difference to the raw score expected at random for that combination of set sizes and dividing the result by the random standard deviation:

Given,

$$rs(S_1,S_2) = \text{raw score of set } S_1 \text{ vs. set } S_2$$

$$n(S_1,S_2) = \text{size}(S_1) \times \text{size}(S_2)$$

$$\mu(x) \approx (4.24 \times 10^{-4})\, x \qquad \text{[Expected raw score mean, \textbf{Supp. Table 5}]}$$

$$\sigma(x) \approx (4.49 \times 10^{-3})\, x^{\,0.665} \qquad \text{[Expected raw score std. dev., \textbf{Supp. Table 5}]}$$

Then,

$$z = (rs(S_1,S_2) - \mu(n(S_1,S_2))) \,/\, \sigma(n(S_1,S_2))$$

For the comparison of the sets Dihydrofolate reductase inhibitor (DHFR) vs. thymidylate synthase inhibitor (TS):

$$rs(S_{\text{DHFR}}, S_{\text{TS}}) = 772.25$$

$$n(S_{\text{DHFR}}, S_{\text{TS}}) = 216 \times 253 = 54{,}648$$

$$z = (772.3\text{-}12.241)/6.349 \approx 119$$

I.B Transformation of Z-scores to E-values

The background distribution of random-set raw scores of MDDR compounds conformed to an extreme value distribution (**Supplementary Figure 1**, see also **Methods**). Given this observation, the probability of obtaining the same or better raw score by random chance alone can be calculated by:[9]

$$P(Z > z) = 1 - \exp(\text{-}e^{\text{-}z\pi/\text{sqrt}(6)-\Gamma'(1)})$$

Where $\Gamma'(1)$ is the Euler-Mascheroni constant (≈ 0.577215665).

Then it follows that:

$$E(z) = P(z)N_{db}$$

Where N_{db} is the number of set comparisons made in the database search.

Note, however, that for Z-scores > 28, $P(Z > z)$ exceeds the numerical precision of most computing languages (e.g., Python), and the following numerical approximation following a Taylor expansion may be used instead:[10]

$$x = \text{-}\exp(\text{-}z\pi \,/\, \text{sqrt}(6) - \Gamma'(1))$$

---

[9] Pearson, W.R. Empirical statistical estimates for sequence similarity searches. *J Mol Biol* **276**, 71-84 (1998).

[10] Valiant, P. *Personal communication*. 2004.

$$P(Z > z) = -(x + (x^2)/2 + (x^3)/6)$$

For the comparison of the sets Dihydrofolate reductase inhibitor (DHFR) vs. thymidylate synthase inhibitor (TS), $z \approx 119$, so:

$$P(Z > 119) \approx 2.92 \times 10^{-67}$$

$$E(119) = (2.92 \times 10^{-67})(246^2) \approx 1.77 \times 10^{-62}$$

The astute reader will note that this value does not perfectly match that reported in the **Results** for this comparison (1.11×10^{-61}) as the calculations in this example use fewer significant digits and the approximation error has accumulated.

II. Supplementary figures
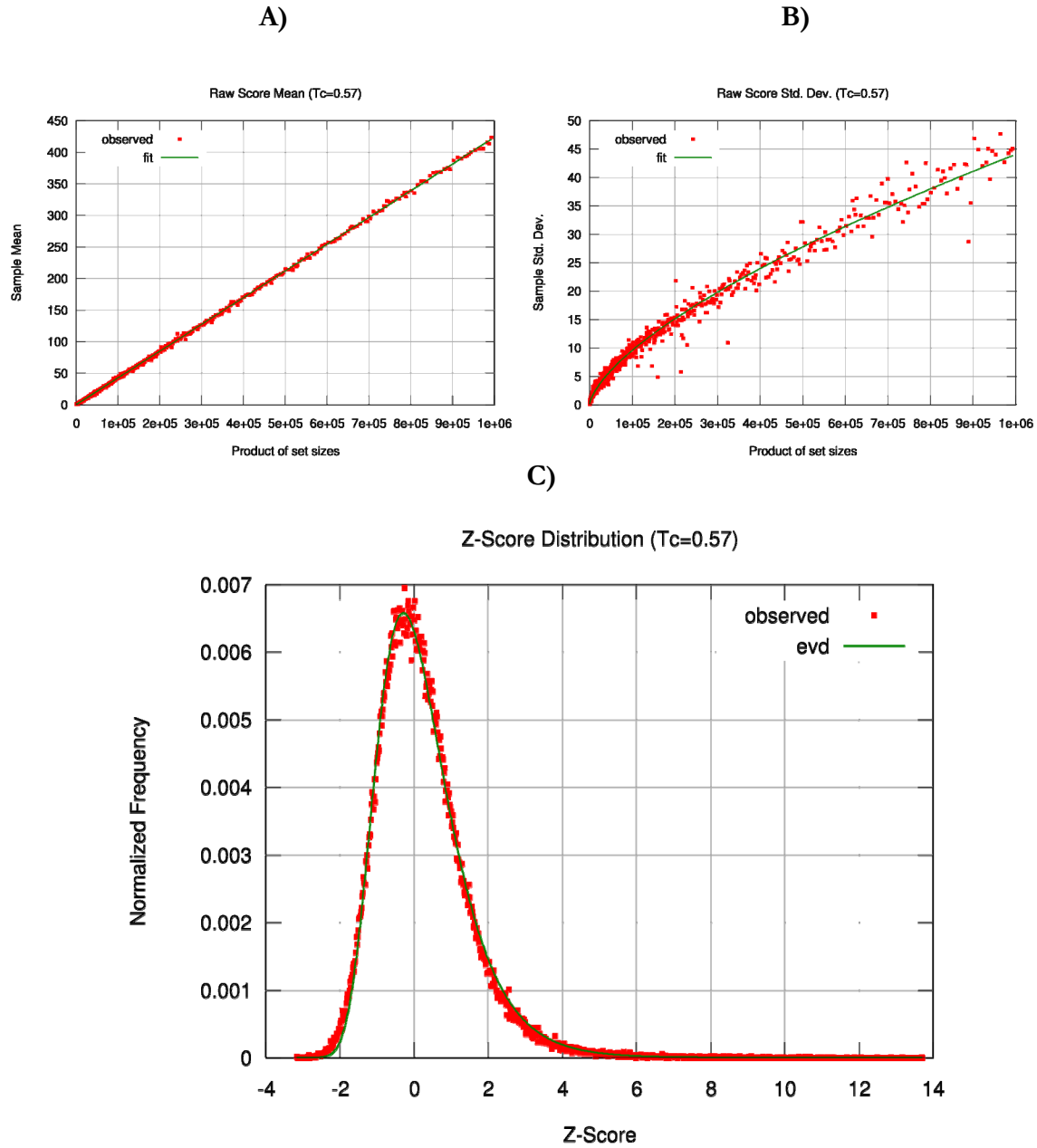
Figure A.1.1 Statistical model fits for MDDR

Plots and fits for (a) mean, (b) standard deviation, and (c) z-score distribution of the random background statistical model calculated from the filtered MDDR database. $N = 1,421$ for all plots.
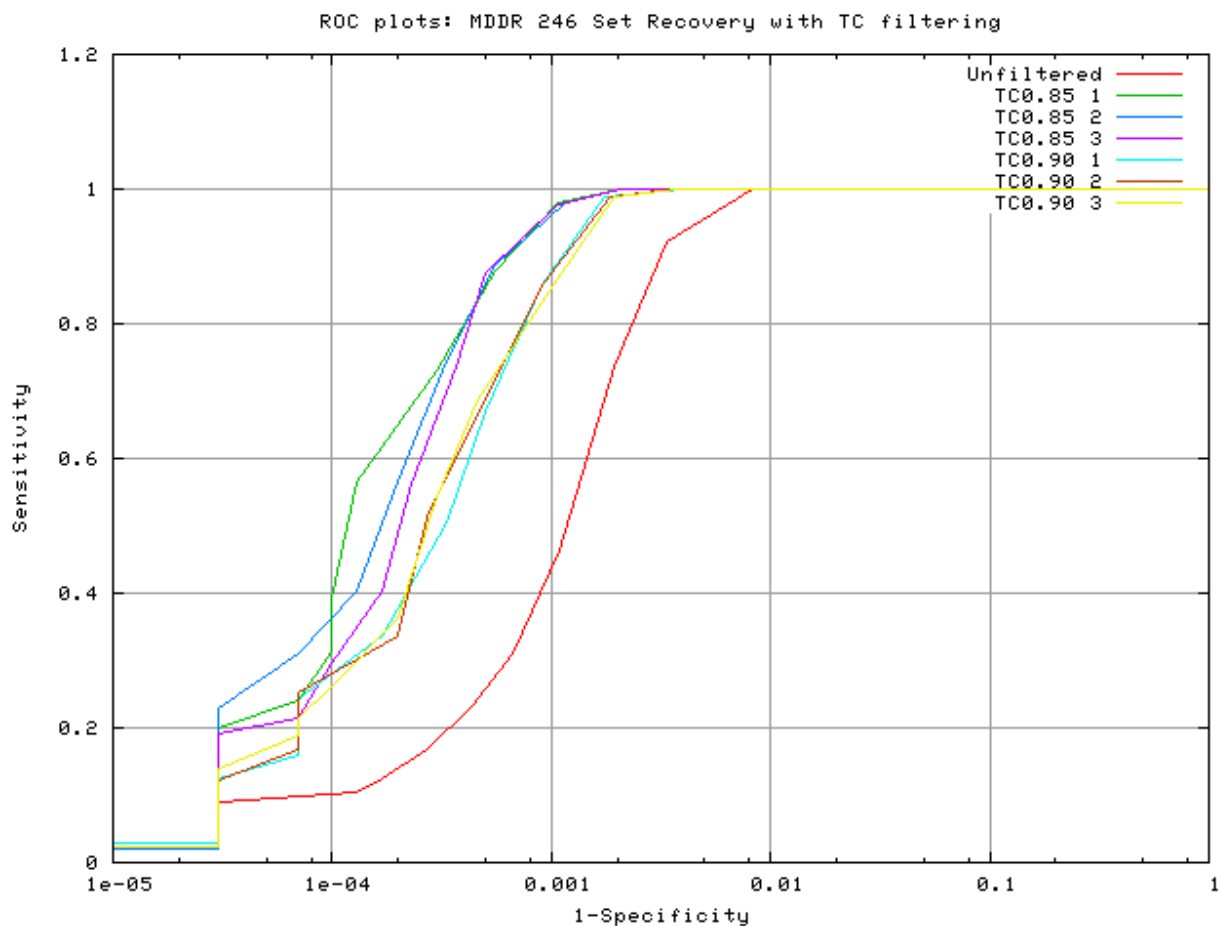
Figure A.1.2 Set recovery in database search after TC-chemotype filtering

We were interested to determine the effect of redundancy in medicinal chemistry and drug databases on SEA. Compounds from the 246 MDDR activity classes were placed in a randomly-ordered list. Starting with the first compound, any other within 0.90 Tc was removed. This procedure was also repeated at a more stringent 0.85 Tc radius. These procedures were repeated three times at each radius, and the resulting collections of filtered sets were compared and plotted. The ROC plot measures the ability of the scoring technique to recover the correct match of the query activity class to itself in the reference collection, over a sliding threshold. Note that the x-axis is displayed on the log scale, as no visual differentiation of the curves was otherwise possible. The comparison of the filtered collection versus the unfiltered MDDR also yields a slight performance increase (data not shown). We conclude that Tc-based chemotype filtering slightly improves MDDR activity class recall and thus SEA does not depend on redundancy, as one would expect decreased performance with the removal of redundant compounds, were this the case.
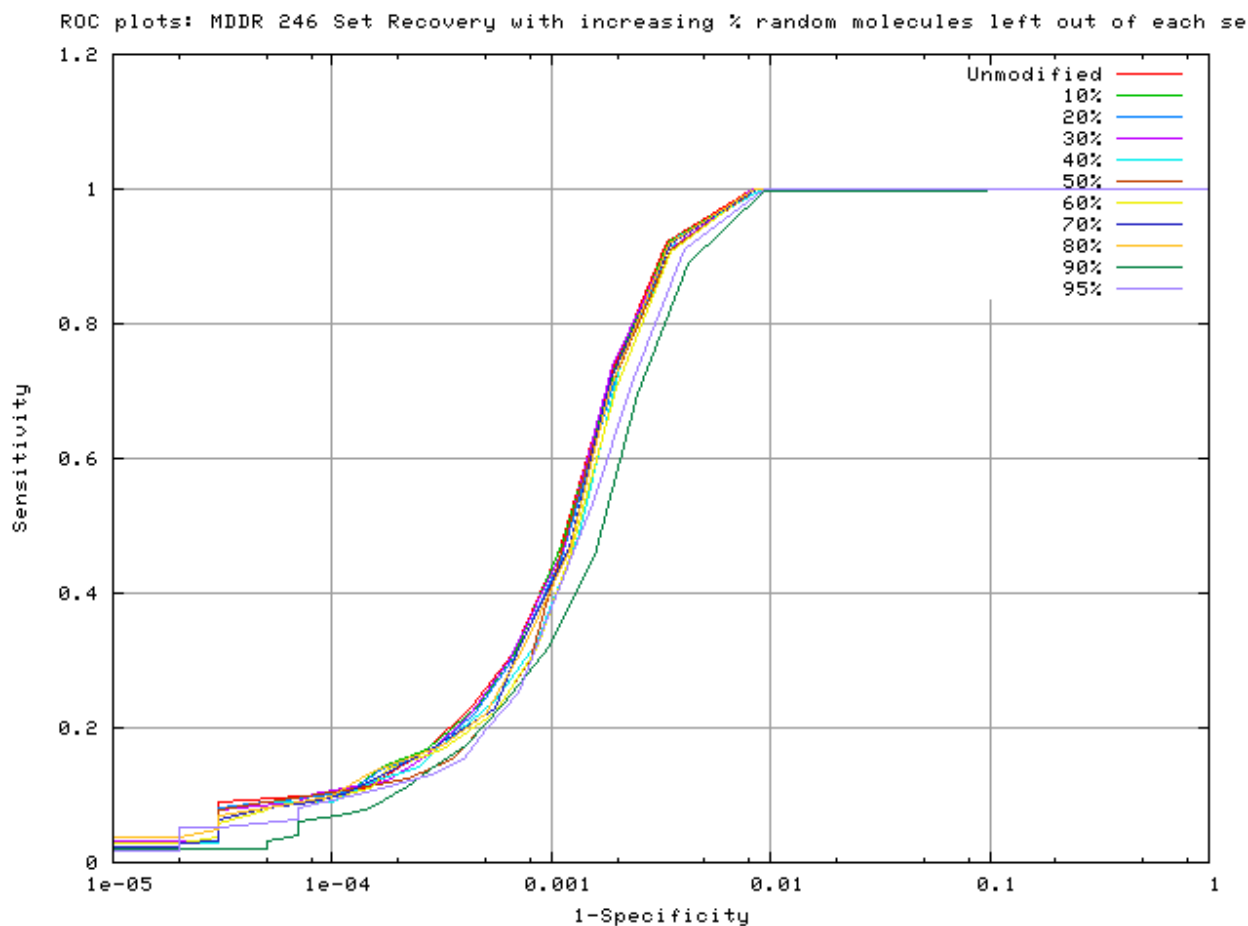
Figure A.1.3 Set recovery in database search with progressive random removal of compounds from query set

As an additional check on the dependence of SEA on redundancy or particular component compounds, a second experiment was performed, in which increasingly large percentages of ligand sets were randomly removed. In this procedure, each set had a random 10% of its ligands for the first run, a different random 20% for the next, and so on. These depleted set collections were compared against the unmodified MDDR set collection (as sets with 90% of compounds removed, for

instance, became prohibitively small to compare against each other statistically). The data for one such representative run, from 10% removed to 90% removed is plotted as a ROC curve with a log-scale x-axis, showing no appreciable performance decrease, and this was the case with all runs computed (data not shown). We conclude that SEA signal is robust to random and extensive removal of individual compounds from the sets.
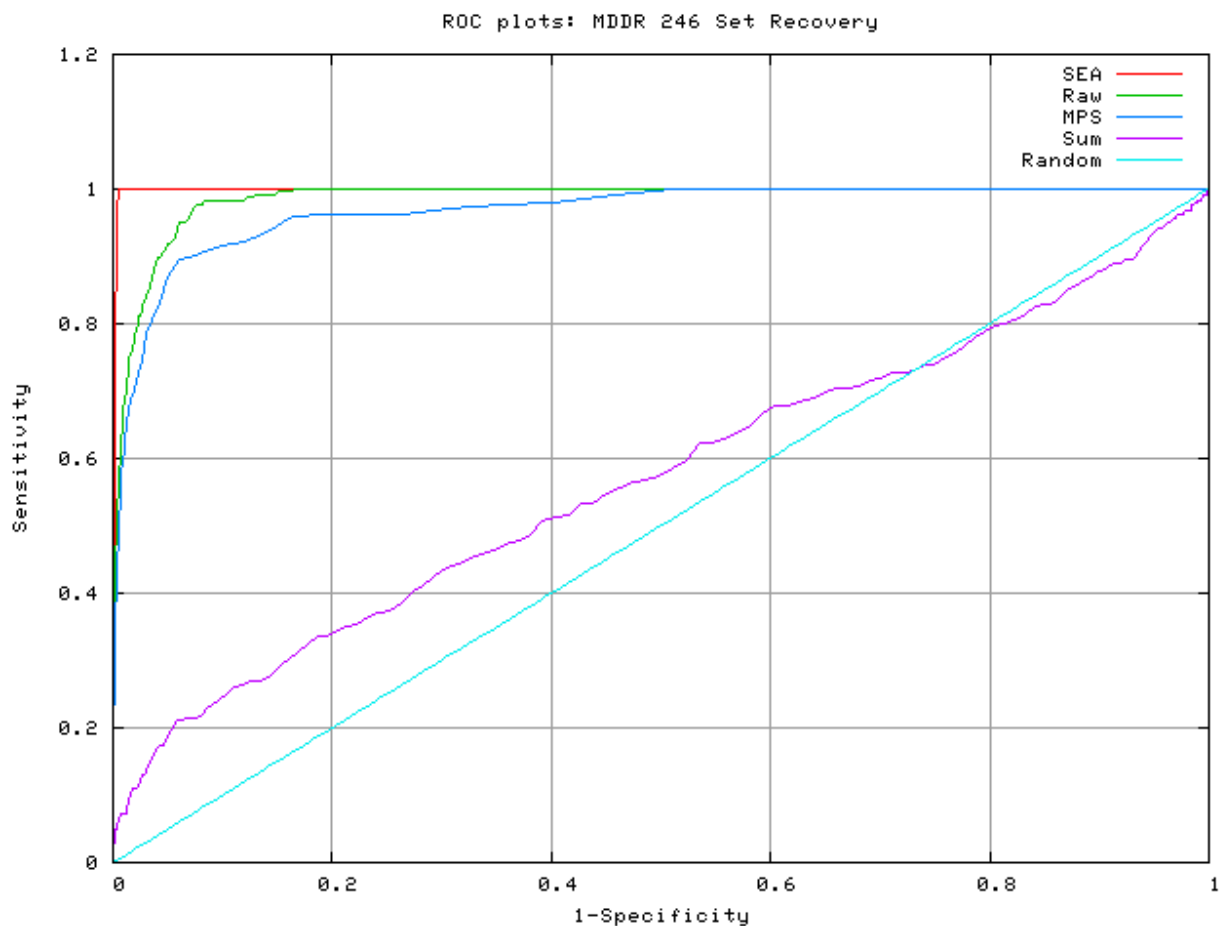
Figure A.1.4 Set recovery in database search over 246 MDDR classes

Each MDDR activity class was queried against all 246 activity classes in the collection, and the results ranked by score for each of four methods: SEA E-values (red), raw SEA scores (green), mean pair-wise similarity (blue), and a simple sum of Tanimoto coefficients between sets (violet). The ROC plot measures the ability of the scoring technique to recover the correct match of the query activity class to itself in the reference collection, over a sliding threshold.

Background distribution fit vs. Tc threshold
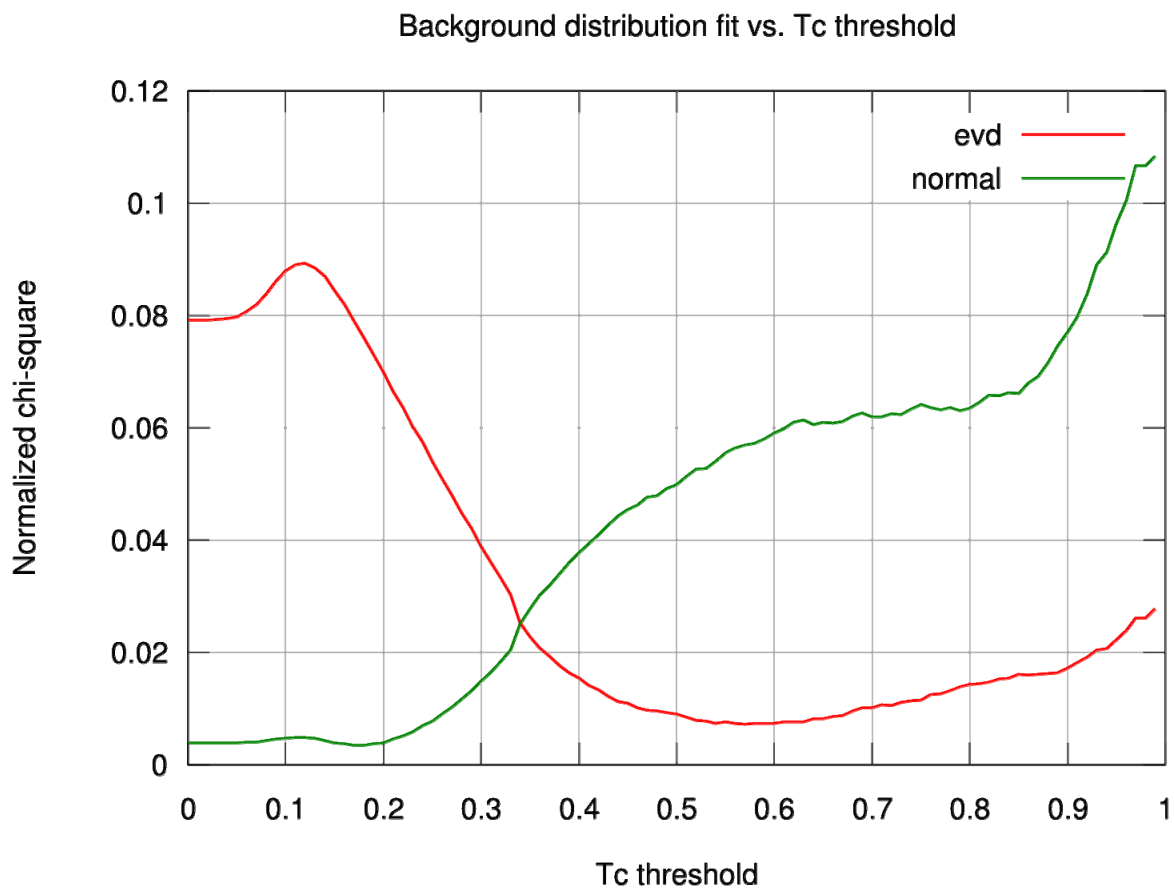
Figure A.1.5 Choice of threshold parameter

For each possible Tc threshold from 0 to 0.99, the best fit of the background Z-score distribution (as in **Supplementary Figure 1c**) to a normal distribution (green) and to an extreme value (red) distribution was calculated. The goodness-of-fit criterion used was a normalized chi-square measure, as described in **Supplementary Table 5**. Any threshold above 0.35 Tc favors an EVD over a normal distribution, and this appears to be an inflection point for both distributions. The 0.55-0.65 Tc threshold region achieves the best EVD fits, with a 0.57 Tc threshold being optimal.
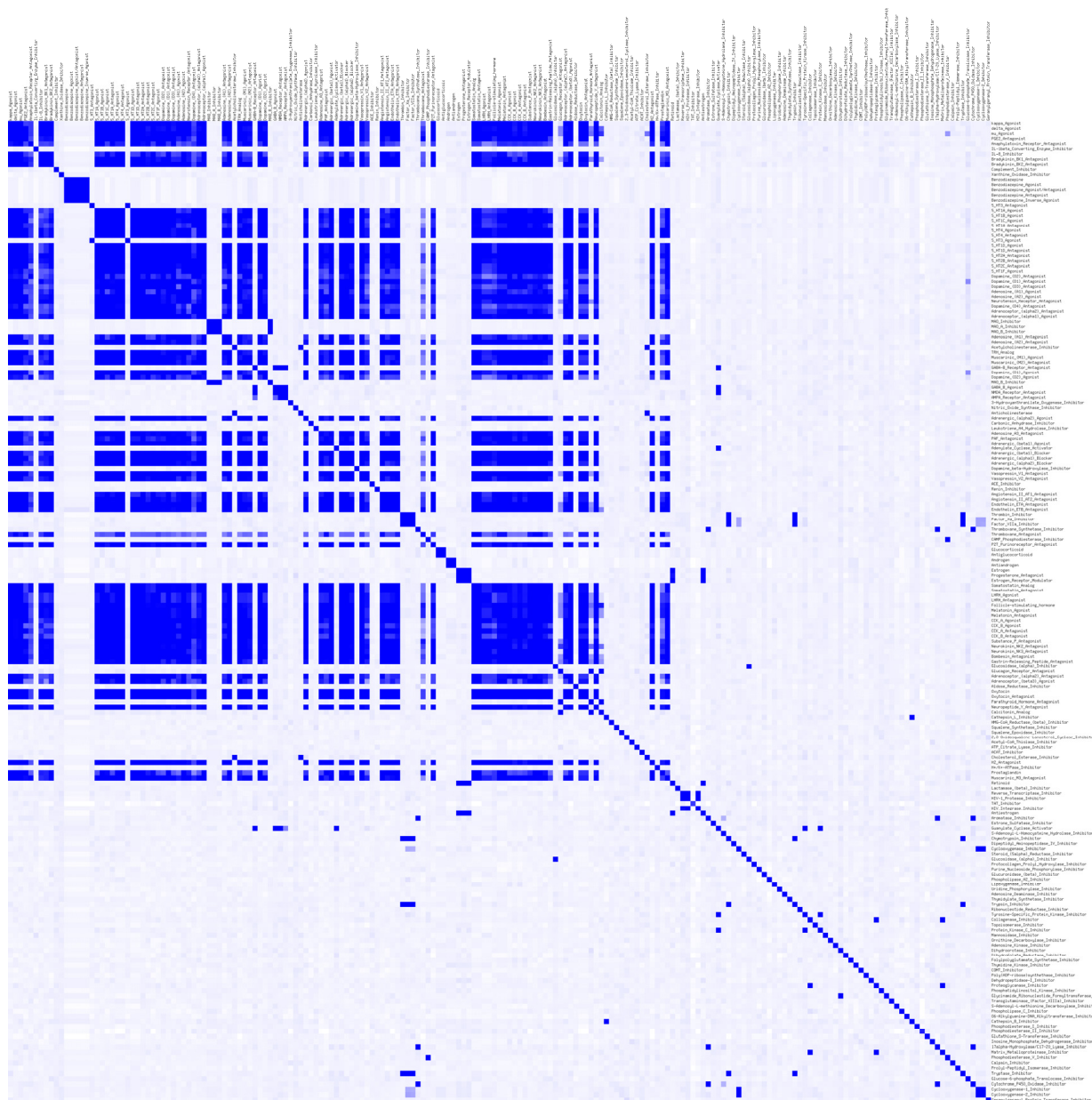
Figure A.1.6 PSI-BLAST heat map of MDDR activity class target protein sequences compared against themselves

Matrix of sequence similarity for 194 MDDR activity classes. The 194 classes form the x- and y-axis, and any given cell is colored by the natural log of the PSI-BLAST E-value of the comparison between the two relevant activity classes. White represents weak or no similarity (E-value of 1×10^5 and above), and dark blue high similarity (E-value of 1×10^{-50} or better).
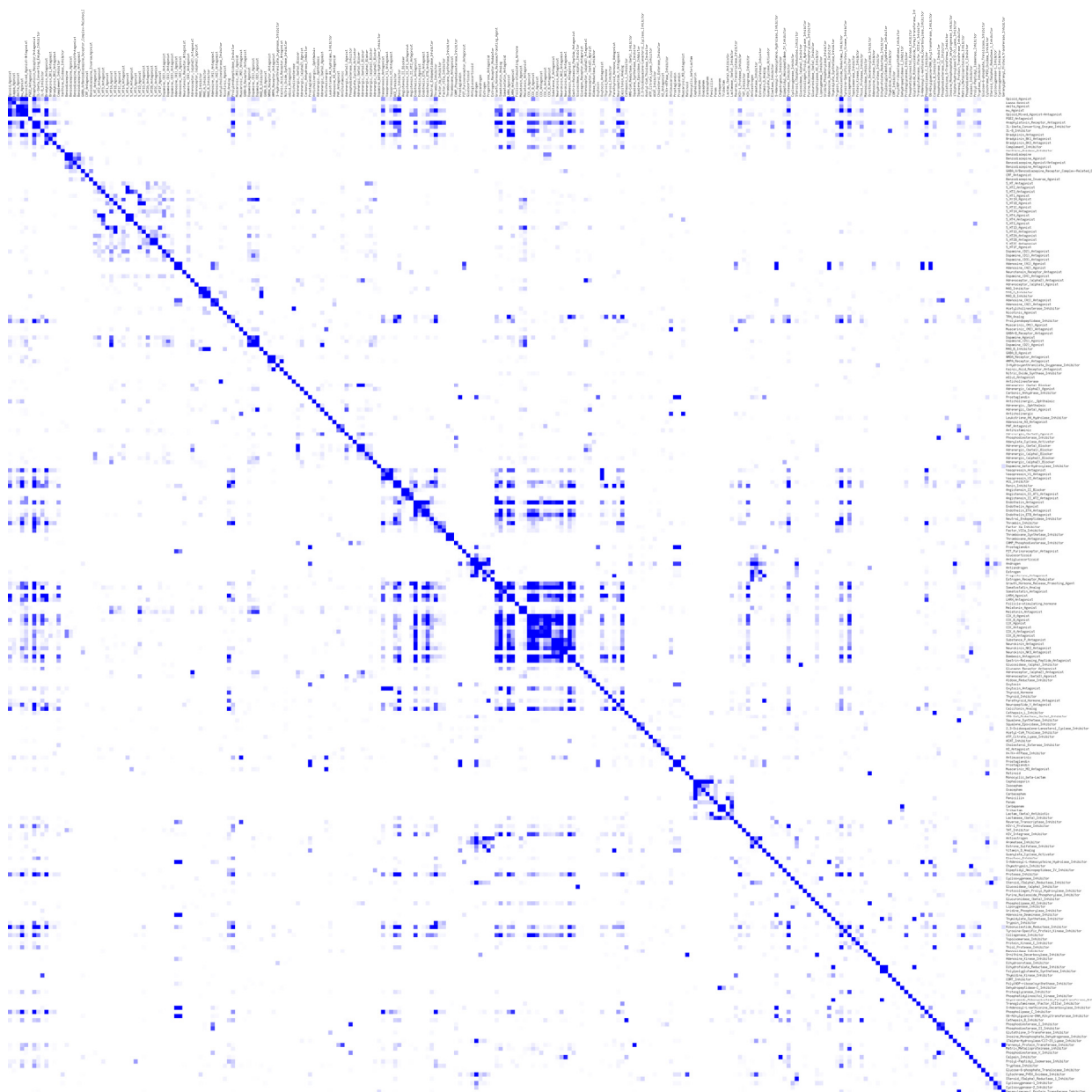
Figure A.1.7 SEA heat map of MDDR activity classes compared against themselves

This heat map is an alternate representation of the same SEA E-value matrix data used to build the naïve network graph and similarity maps of **Figure 2**. The 246 MDDR activity classes under consideration form the x- and y-axis, and any given cell is colored by the natural log of the SEA E-value of the comparison between the two relevant activity classes. White represents weak or no similarity (E-value of 1×10^5 and above), and dark blue high similarity (E-value of 1×10^{-50} or better). This heat map was subtracted from that in **Supplementary Figure 6** to create **Figure 3**, as described in **Methods**.

III. Supplementary tables

Table A.1.1 Expanded statistics for Table 1.1 and Table 1.2

Query	Rank	Size	Similar Activity Classes	E-value	MPS	TC Sum	TC> 0.85	TC 1.0	Min TC	Max TC
AMPA Receptor Antagonist	1	569	AMPA Receptor Antagonist	2.45×10^{-219}	0.236	76429.8	1833	577	0.067	1.00
	2	75	Kainic Acid Receptor Antagonist	5.28×10^{-80}	0.222	9473.3	210	74	0.012	1.00
	3	1485	NMDA Receptor Antagonist	3.08×10^{-63}	0.195	164377.0	729	181	0.004	1.00
	4	22	Anaphylatoxin Receptor Antagonist	3.81×10^{-4}	0.188	2355.4	0	0	0.080	0.70
	5	130	mu Agonist	1.69×10^{-3}	0.190	14075.3	0	0	0.067	0.83
	6	99	Ribonucleotide Reductase Inhibitor	1.00×10^{-1}	0.166	9330.1	0	0	0.017	0.73
Carbacephem	1	98	Carbacephem	0*	0.573	5504.9	518	106	0.260	1.00
	2	1614	Cephalosporin	1.11×10^{-222}	0.390	61624.0	155	14	0.098	1.00
	3	35	Isocephem	2.30×10^{-17}	0.413	1416.2	0	0	0.258	0.64
	4	257	Penem	2.43×10^{-4}	0.286	7192.3	0	0	0.191	0.68
	5	13	Oxacephem	8.38×10^{-3}	0.352	448.5	0	0	0.262	0.69
	6	39	Lactam (beta) Antibiotic	2.62×10^{-2}	0.306	1168.1	0	0	0.159	0.62
	7	223	Lactamase (beta) Inhibitor	6.58×10^{-1}	0.248	5420.7	1	1	0.064	1.00
	8	116	Monocyclic beta-Lactam	3.18×10^{2}	0.370	4203.3	0	0	0.144	0.61
Androgen	1	50	Androgen	0*	0.668	1670.2	514	138	0.083	1.00
	2	577	Aromatase Inhibitor	6.87×10^{-307}	0.256	7375.7	30	0	0.025	0.88
	3	43	Antiglucocorticoid	2.30×10^{-102}	0.255	548.2	11	0	0.036	0.89
	4	6	Cytochrome	4.01×10^{-93}	0.229	68.7	20	0	0.091	0.92

Query	Rank	Size	Similar Activity Classes	E-value	MPS	TC Sum	TC> 0.85	TC 1.0	Min TC	Max TC
			P450 Oxidase Inhibitor							
	5	179	Estrogen	9.97×10^{-89}	0.207	1852.5	24	0	0.025	0.91
	6	86	Antiestrogen	2.18×10^{-76}	0.208	895.5	0	0	0.033	0.84
	7	936	Steroid (5alpha) Reductase Inhibitor	1.58×10^{-72}	0.258	12094.3	0	0	0.024	0.80
	8	103	Antiandrogen	1.14×10^{-70}	0.157	808.4	24	0	0.033	0.99
	9	86	17alpha-Hydroxylase/C17-20 Lyase Inhibitor	7.88×10^{-66}	0.162	697.3	0	0	0.015	0.76
	10	164	Progesterone Antagonist	3.26×10^{-44}	0.321	2634.2	11	0	0.066	0.89
	11	62	Prostaglandin	1.93×10^{-38}	0.310	961.0	0	0	0.078	0.75
5 HT1F Agonist	1	111	5 HT1F Agonist	6.72×10^{-187}	0.376	4627.1	257	113	0.139	1.00
	2	621	5 HT1D Agonist	8.08×10^{-38}	0.317	21841.9	40	0	0.126	0.95
	3	51	5 HT1B Agonist	2.96×10^{-10}	0.310	1756.2	5	0	0.160	0.95
	4	65	5 HT1 Agonist	3.03×10^{-8}	0.301	2175.2	0	0	0.138	0.81
	5	670	Dopamine (D4) Antagonist	1.90×10^{-6}	0.266	19777.1	0	0	0.108	0.79
	6	565	5 HT1A Antagonist	8.64×10^{-1}	0.283	17733.1	0	0	0.086	0.71
	7	33	5 HT2 Antagonist	8.78×10^{-1}	0.259	949.3	0	0	0.136	0.65
	8	705	5 HT2A Antagonist	1.47	0.275	21529.6	0	0	0.089	0.73
Adrenergic (beta1) Agonist	1	8	Adrenergic (beta1) Agonist	3.85×10^{-241}	0.621	39.7	28	10	0.260	1.00
	2	305	Adrenergic (beta) Agonist	9.50×10^{-34}	0.311	759.9	0	0	0.165	0.81
	3	67	Adrenergic (beta1) Blocker	4.99×10^{-32}	0.360	193.0	0	0	0.126	0.64

Qu-ery	Ra-nk	Size	Similar Activity Classes	E-value	MPS	TC Sum	TC> 0.85	TC 1.0	Min TC	Max TC
	4	563	Adrenoceptor (beta3) Agonist	2.98×10^{-24}	0.304	1369.3	0	0	0.125	0.72
	5	212	Adrenergic (beta) Blocker	3.96×10^{-13}	0.159	16.5	0	0	0.113	0.78
	6	13	Adrenergic, Ophthalmic	2.77×10^{-7}	0.247	25.7	0	0	0.134	0.70
	7	518	Adrenergic (alpha1) Blocker	6.84×10^{-5}	0.239	990.0	0	0	0.121	0.73
	8	124	Melatonin Agonist	1.04×10^{-1}	0.322	319.1	0	0	0.184	0.63
	9	76	Dopamine (D1) Agonist	2.18×10^{-1}	0.258	157.2	0	0	0.117	0.71
	10	102	Adrenergic (alpha2) Agonist	4.72×10^{-1}	0.191	156.1	0	0	0.086	0.66
Dihydrofolate Reductase Inhibitor	1	216	Dihydrofolate Reductase Inhibitor	7.07×10^{-182}	0.340	15898.7	736	218	0.104	1.00
	2	53	Glycinamide Ribonucleotide Formyltransferase Inhibitor	3.97×10^{-100}	0.330	3787.1	36	16	0.110	1.00
	3	6	Folylpolyglutamate Synthetase Inhibitor	4.59×10^{-62}	0.372	482.5	36	6	0.140	1.00
	4	253	Thymidylate Synthase Inhibitor	1.11×10^{-61}	0.309	16864.7	108	30	0.089	1.00

Note on E-values vs. MPS:

It is often the case that MPS scores decrease with SEA E-values in the table above. However, MPS does not result in the same ranking as SEA. Consider the top six "AMPA Receptor Antagonist" hits by E-value (above) as opposed to those by MPS (see following example table):

Rank	Size	Activity Class	E-value	MPS	TC Sum	Min TC	TC> 0.85	TC1.0	Max TC

1	6	Folylpolyglutamate Synthetase Inhibitor	**	0.237	808.9	0.131	0	0	0.39
2	569	AMPA Receptor Antagonist	2.45×10^{-219}	0.236	76429.8	0.067	1833	577	1.00
3	25	LHRH Agonist	6.05×10^{4}	0.231	3291.1	0.092	0	0	0.68
4	18	Anticholinergic, Ophthalmic	**	0.227	2328.3	0.133	0	0	0.46
5	33	Somatostatin Analog	3.33×10^{4}	0.227	4256.2	0.114	0	0	0.61
6	253	Thymidylate Synthetase Inhibitor	**	0.227	32627.0	0.077	0	0	0.50

MPS incorrectly ranks the antifolate Folylpolyglutamate Synthetase Inhibitor class above the actual query, and fails to rank the highly-related Kainic Acid Receptor Antagonist and NMDA Receptor Antagonist classes (AMPA, KA, and NMDA are related glutamate receptors) within the top hits. As is evident from this example, MPS scores do not identify critical high-similarity subsets between two sets, and can rank sets that are completely unrelated in biological function to the query more highly than the query itself. MPS scores often decrease with SEA scores, but they can be a poor filter as they do not enrich for similarity of similar sub-clusters across sets.

** No E-value calculated, as raw score was not >0. These sets did not share a single ligand pair above the 0.57 TC raw score threshold.

Table A.1.2 MDDR unrelated orphans

Activity Class	Self E-value
Acetyl-CoA Thiolase Inhibitor	0
ATP Citrate Lyase Inhibitor	3.76×10^{-90}
Carbonic Anhydrase Inhibitor	5.98×10^{-318}
COMT Inhibitor	2.59×10^{-171}
CRF Antagonist	2.93×10^{-121}
GABA B Agonist	6.08×10^{-174}
GABA-B Receptor Antagonist	9.80×10^{-153}
Glucose-6-phosphate Translocase Inhibitor	4.56×10^{-237}
Glucuronidase (beta) Inhibitor	9.29×10^{-95}
Guanylate Cyclase Activator	0
Inosine Monophosphate Dehydrogenase Inhibitor	8.70×10^{-250}
mGlu1 Antagonist	1.22×10^{-129}
Muscarinic (M1) Agonist	1.70×10^{-151}
Nitric Oxide Synthase Inhibitor	4.64×10^{-105}
Phosphatidylinositol Kinase Inhibitor	1.31×10^{-251}
Poly(ADP-ribose)synthethase Inhibitor	3.44×10^{-111}
Thyroid Inhibitor	1.19×10^{-213}
Uridine Phosphorylase Inhibitor	3.31×10^{-199}

The above table lists all specific MDDR activity classes with best E-values > 1.0 to any other class.

Table A.1.3 Rankings of the correct MDDR activity class for each PubChem MeSH pharmacological action set by SEA and by MPS

	MeSH Pharmacological Action	Rank of top matching MDDR activity class	
		SEA	MPS
1	Adrenergic α-Antagonists	1	10
2	Adrenergic β-Agonists	1	1
3	Adrenergic β-Antagonists	1	2
4	Androgen Antagonists	3	13
5	Androgens	1	1
6	Aromatase Inhibitors	2	2
7	Carbonic Anhydrase Inhibitors	1	1
8	Cholinergic Antagonists	1	1
9	Cholinesterase Inhibitors	1	8
10	Cyclooxygenase Inhibitors	2	57
11	Dopamine Agonists	1	5
12	Estrogen Antagonists	1	1
13	Estrogens	1	1
14	Glucocorticoids	1	1
15	Histamine H2 Antagonists	1	1
16	HIV Protease Inhibitors	1	6
17	Lipoxygenase Inhibitors	1	42
18	Muscarinic Antagonists	2	1
19	Nicotinic Agonists	1	6
20	Phosphodiesterase Inhibitors	1	4
21	Protease Inhibitors	4	11
22	Reverse Transcriptase Inhibitors	2	18
23	Trypsin Inhibitors	1	10
	Average	1.4	8.2

Table A.1.4 Loperamide and emetine functional assay data

N	Receptor	Agonist	Predicted antagonist	pEC50±SEM (EC50, nM)		Fold change in EC50	Two-tail T-test (<0.05)
				Vehicle + agonist	Predicted antagonist + agonist		
3	NK2	[β-Ala8]-Neurokinin A Fragment 4-10	Loperamide HCl	9.24±0.05 (0.57)	8.37±0.05 (4.3)	7.5	0.0002
3	alpha2a	clonidine	Emetine	8.16±0.09 (6.9)	7.14±0.11 (73)	10.6	0.0019
3	alpha2c	clonidine	Emetine	7.40±0.09 (40)	5.96±0.11 (1094)	27.5	0.0005

Table A.1.5 SEA statistical model fits

A) Raw score fits at threshold Tc=0.57 ($N = 1,421$).

Raw Score Fits	Value	Pearson r^2
Mean coefficient	4.24×10^{-4}	0.9998
Mean exponent	1 (const.)	
Std. dev. coefficient	4.49×10^{-3}	0.9882
Std. dev. exponent	6.65×10^{-1}	

B) Distribution fits of background Z-scores at threshold Tc=0.57 ($N = 1,421$).

Distribution	Normalized Chi-Square*	Pearson r^2
Extreme value	6.89×10^{-3}	0.9969
Normal	5.77×10^{-2}	0.9830

* Normalized chi-square (X_N) computed as:

$$X_N = SUM \{ (observed - expected)^2 / (observed + expected) \}$$

C) EVD and Normal distribution parameters at threshold 0.57 Tc.

Distribution	Loc	Scale	Height
Extreme value	-2.88×10^{-1}	9.45×10^{-1}	1.69×10^{-2}
Normal	-4.50×10^{-2}	1.00	1.63×10^{-2}

A.2 Supplementary material for Chapter 2

I. Supplementary figures

Radioligand displacement assays

A

5-HT1A Receptors
[³H]8-OH-DPAT (0.5 nM)
Standard Binding Buffer

- Methysergide
  Ki = 60 nM
- 9363
  Ki = 112 nM
- 9414
  Ki = 3.2 nM

B

5-HT1D Receptors
[³H]GR125743 (0.43 nM)
Standard Binding Buffer

- Ergotamine
  Ki = 1.9 nM
- Fluanisone
  Ki = 137 nM

C

Alpha$_{1A}$ Receptor
[^{125}I]HEAT (0.1 nM)
Alpha1 Binding Buffer

- Prazosin
  K_i = 0.25 nM
- Domperidone
  K_i = 71 nM

D

Alpha1B Receptor
[^{125}I]HEAT (0.1 nM)
Alpha1 Binding Buffer

- Corynanthine
  Ki = 612 nM
- 9414
  Ki = 14 nM

E

Alpha1B Receptor
[^{125}I]HEAT (0.1 nM)
Alpha1 Binding Buffer

- Corynanthine
  Ki = 530 nM
- 9363
  Ki = 245 nM

F

Alpha1B Receptor
[^{125}I]HEAT (0.1 nM)
Alpha1 Binding Buffer

- Corynanthine
  Ki = 629 nM
- 9474
  Ki = 535 nM

Radioligand displacement assays

G

Alpha$_{1D}$ Receptor
[^{125}I]HEAT (0.1 nM)
Alpha1 Binding Buffer

Prazosin
K_i = 0.29 nM

Dimetholizine
K_i = 171 nM

Fluanisone
K_i = 7 nM

H

Alpha$_{1D}$ Receptor
[^{125}I]HEAT (0.1 nM)
Alpha1 Binding Buffer

Prazosin
K_i = 0.63 nM

Domperidone
K_i = 705 nM

I

Alpha$_{2A}$ Receptor
[^{125}I]Clonidine (0.1 nM)
Alpha2 Binding Buffer

Oxymetazoline
K_i = 0.41 nM

9415
K_i = 959 nM

J

Alpha$_{2C}$ Receptor
[^{125}I]Clonidine (0.1 nM)
Alpha2 Binding Buffer

Prazosin
K_i = 14 nM

Tetrabenazine
K_i = 1,318 nM

K

Delta Opioid Receptor
[^{3}H]DADLE (0.32nM)
Standard Binding Buffer

Naltrindole
K_i = 0.7 nM

Diphemanil
K_i = 13,610 nM

L

Dopamine$_2$ Receptor
[^{3}H]N-Methylspiperone (0.17 nM)
Dopamine Binding Buffer

Haloperidol
K_i = 2.7 nM

Dimetholizine
K_i = 179 nM

Radioligand displacement assays

M

Dopamine$_4$ Receptor
[^{3}H]N-Methylspiperone (0.31 nM)
Dopamine Binding Buffer

- ■ Chlorpromazine K_i = 26 nM
- ▼ Indoramin K_i = 18 nM

N

Alpha$_{1A}$ Receptor
[^{125}I]HEAT (0.1 nM)
Alpha1 Binding Buffer

- ■ Prazosin K_i = 0.29 nM
- ▲ 9414 K_i = 1.2 nM
- ▼ 9415 K_i = 498 nM

O

Beta$_3$ Receptor
[^{125}I]Pindolol (0.1 nM)
Beta Binding Buffer

- ■ Alprenolol K_i = 62 nM
- ▲ Denopamine K_i = 2,090 nM

P

Beta$_1$ Receptor
[^{125}I]Pindolol (0.1 nM)
Beta Binding Buffer

- ● Alprenolol K_i = 2 nM
- ■ 10711 K_i = 553 nM

Q

Beta$_3$ Receptor
[^{125}I]Pindolol (0.1 nM)
Beta Binding Buffer

- ● Alprenolol K_i = 12 nM
- ■ 10711 K_i = 11,967 nM

R

Alpha1A Receptor
[^{125}I]HEAT (0.092 nM)
Alpha1 Binding Buffer

- ■ Prazosin K_i = 0.15 nM
- ▼ 9363 K_i = 79 nM

Radioligand displacement assays

S

Beta$_1$ Receptor
[^{125}I]Pindolol (0.1 nM)
Beta Binding Buffer

- Alprenolol K_i = 1.6 nM
- Paroxetine K_i = 10,420 nM

T

Beta$_2$ Receptor
[^{125}I]Pindolol (0.1 nM)
Beta Binding Buffer

- Alprenolol K_i = 3.9 nM
- Paroxetine K_i = 77,000 nM

U

Beta$_1$ Receptor
[^{125}I]Pindolol (0.1 nM)
Beta Binding Buffer

- Alprenolol K_i = 2.0 nM
- Fluoxetine K_i = 4,385 nM

V

Beta$_2$ Receptor
[^{125}I]Pindolol (0.1 nM)
Beta Binding Buffer

- Alprenolol K_i = 4.4 nM
- Fluoxetine K_i = 21,000 nM

W

5-HT$_{5A}$ Receptor
[^{3}H]LSD (1 nM)
Standard Binding Buffer

- Chlorpromazine K_i = 121 nM
- Mebhydrolin K_i = 129 nM

X

Histamine H$_4$ Receptor

- Histamine K_i = 28 nM
- Delavirdine K_i = 5334 nM

Y

μ Opioid Receptor

- Naloxone K_i = 9.2 nM
- Ifenprodil K_i = 1,423 nM

Z

SERT

- Amitriptyline K_i = 10 nM
- Ifenprodil K_i = 77 nM

173

Radioligand displacement assays

AA

SERT

- ■ Amitriptyline
  K_i = 10 nM
- ▼ RO-25-6981
  K_i = 1,417 nM

AB

NET

- ■ Desipramine
  K_i = 5.9 nM
- ▲ RO-25-6981
  K_i = 1,248 nM

AC

Dopamine D$_4$ Receptor

- ■ Chlorpromazine
  K_i = 77 nM
- ▼ 9421
  K_i = 117 nM

AD

κ Opioid Receptor

- ■ Salvinorin A
  K_i = 0.9 nM
- ▲ RO-25-6981
  K_i = 3,128 nM

Figure A.2.1 Testing off-target activities

Testing off-target activities via drug-target binding assays. (A) Sedalande and Dimetholizine at 5-HT$_{1A}$, (B) Sedalande at 5-HT$_{1D}$, (C) Motilium at α$_{1A}$, (D) Sedalande at α$_{1B}$, (E) Dimetholizine at α$_{1B}$, (F) Motilium at α$_{1B}$, (G) Sedalande and Dimetholizine at α$_{1D}$, (H), Motilium at α$_{1D}$, (I) Xenazine at α$_{2A}$, (J) Xenazine at α$_{2C}$, (K) Prantal at δ-opioid, (L) Dimetholizine at D$_2$, (M) Doralese at D$_4$, (N) Sedalande and Xenazine at α$_{1A}$, (O) Kalgut at β$_3$, (P) Kalgut at β$_1$, (Q) Kalgut at β$_2$, (R) Dimetholizine at α$_{1A}$, (S) Paxil at β$_1$, and (T) Paxil at β$_2$, (U) Prozac at β$_1$, (V) Prozac at β$_2$, (W) Fabahistin at 5-HT$_{5A}$, (X) Rescriptor at H$_4$, (Y) Vadilex at μ-opioid, (Z) Vadilex at SERT, (AA) RO-25-6981 at SERT, (AB) RO-25-6981 at NET, (AC) RO-25-6981 at D$_4$, (AD) RO-25-6981 at κ-opioid.

Drug name synonyms for Figure A.2.1:

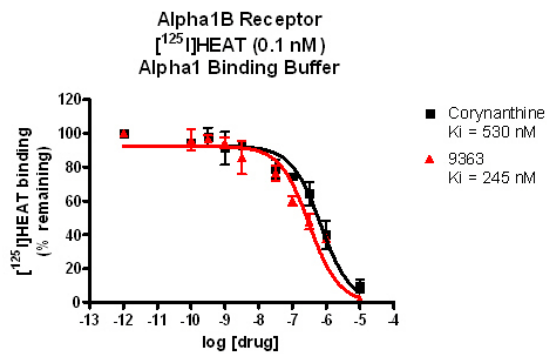| Trade	Generic name	PDSP ID
-	Dimetholizine	9363
DMT	N,N-dimethyltryptamine	-
Doralese	Indoramin	-
Fabahistin	Mebhydrolin	-
Kalgut	Denopamine	10711
Motilium	Domperidone	9474
Paxil	Paroxetine	-
Prantal	Diphemanil	10571
Prozac	Fluoxetine	-
Rescriptor	Delavirdine	-
-	RO-25-6981	9421
Sedalande	Fluanisone	9414
Vadilex	Ifenprodil	-
Xenazine	Tetrabenazine	9415
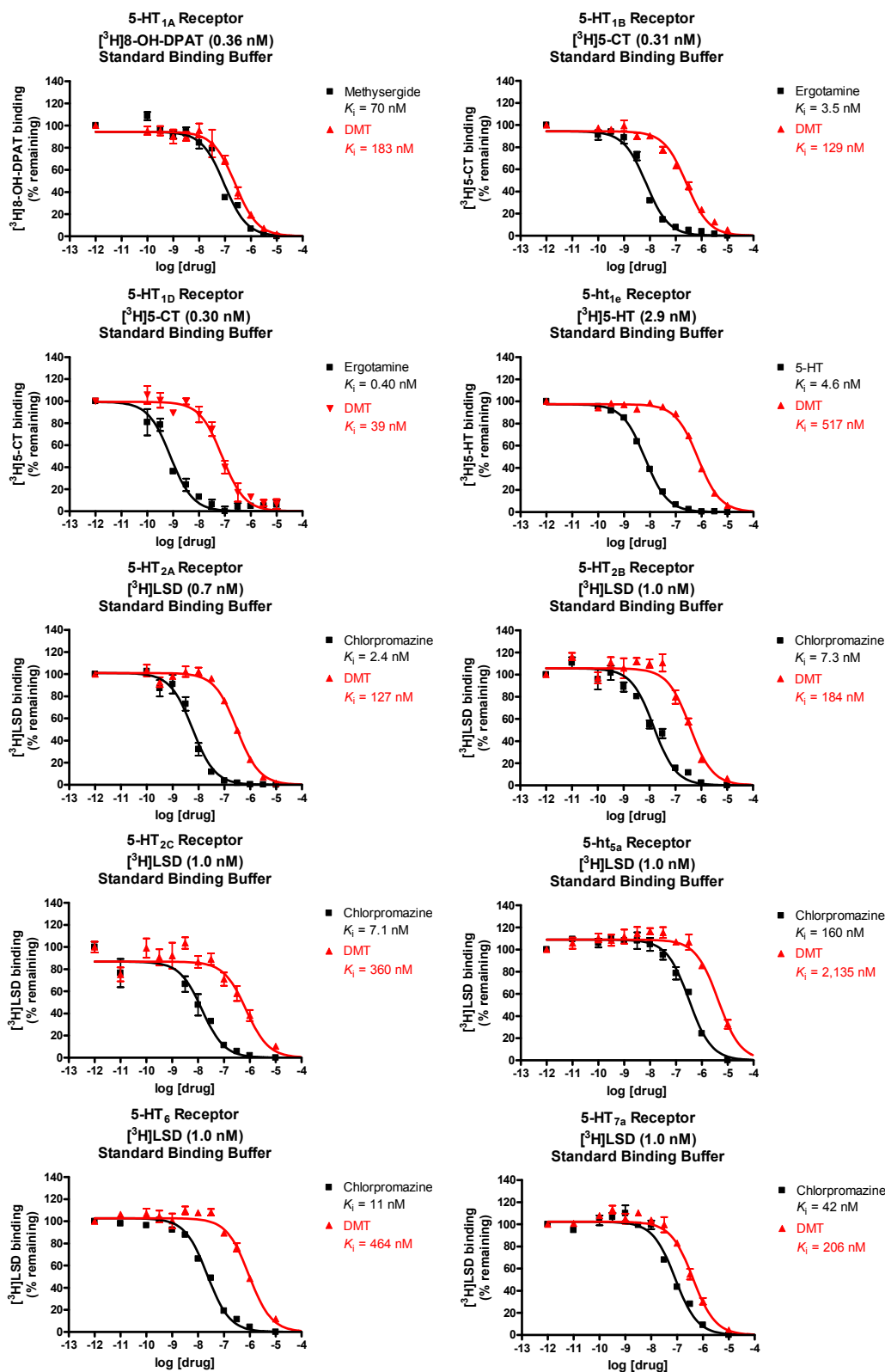
Figure A.2.2 Testing DMT's affinity for serotonergic receptors
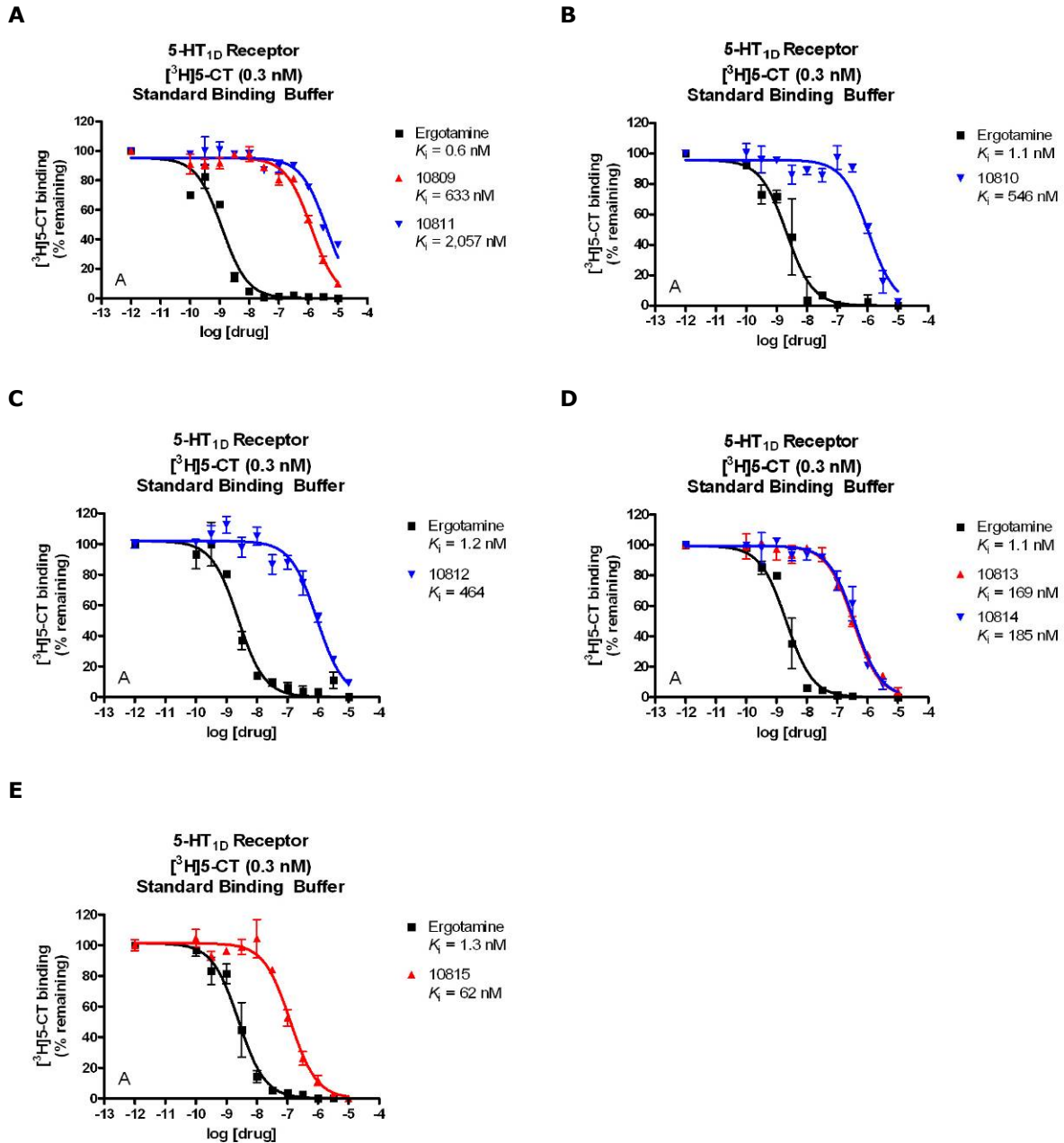
Figure A.2.3 Testing Sedalande-derivative affinities at 5-HT1D

(A-E) All seven Sedalande derivatives assayed against 5-HT$_{1D}$ and identified by their PDSP number. See key for PDSP-numbered compounds in **Table A.2.5**.

II. Supplementary tables

Table A.2.1 Comparison of novel SEA predictions vs. Naïve Bayesian Classifier predictions, on the same dataset

Drug	Bayes Rank	Bayes Score	Off-targets predicted by Naïve Bayesian Classifier
Vadilex	1	6.0114	AgrC
	2	4.6029	Glutamate [NMDA] receptor subunit zeta 1 precursor (N-methyl-D-aspartate receptor subunit NR1)
	3	4.5886	Somatostatin receptor type 1 (SS1R) (SRIF-2)
	4	**4.205**	**Mu-type opioid receptor (MOR-1) (Opioid receptor B) (MUOR1)**
	5	4.1422	Delta-type opioid receptor (DOR-1) (Opioid receptor A)
	6	4.0124	Motilin receptor (G protein-coupled receptor 38)
	7	3.833	Sortilin precursor (Glycoprotein 95) (Gp95) (Neurotensin receptor 3) (NT3) (100 kDa NT receptor)
	8	3.7178	Microbial collagenase precursor (120 kDa collagenase)
	9	3.1589	Neurotensin receptor type 1 (NT-R-1) (High-affinity levocabastine-insensitive neurotensin receptor) (NTRH)
	10	2.9583	Pol polyprotein
	...	...	...
	N/A	**No score > 0**	**Sodium-dependent serotonin transporter (5HT transporter) (5HTT)**
RO-25-6981	1	4.9049	Glutamate [NMDA] receptor subunit zeta 1 precursor (N-methyl-D-aspartate receptor subunit NR1)
	2	4.525	Somatostatin receptor type 1 (SS1R) (SRIF-2)
	3	4.4263	C-C chemokine receptor type 3 (C-C CKR-3) (CC-CKR-3) (CCR-3) (CCR3) (CKR3) (Eosinophil eotaxin receptor)
	4	4.302	Microbial collagenase precursor (120 kDa collagenase)
	5	3.9402	AgrC
	6	3.8675	retropepsin, aspartic peptidase [Human immunodeficiency virus 1]
	7	3.4463	Motilin receptor (G protein-coupled receptor 38)
	8	**3.3128**	**Kappa-type opioid receptor (KOR-1)**
	9	3.2941	Sodium channel protein type II alpha subunit (Voltage-gated sodium channel alpha subunit Nav1.2) (Sodium channel protein, brain II alpha subunit)
	10	3.174	Sortilin precursor (Glycoprotein 95) (Gp95) (Neurotensin receptor 3) (NT3) (100 kDa NT receptor)
	...	...	...
	N/A	**No score > 0**	**5-HTT**

Drug	Bayes Rank	Bayes Score	Off-targets predicted by Naïve Bayesian Classifier
	N/A	**No score > 0**	**Dopamine (D4)**
	N/A	**No score > 0**	**NET**
Paxil	1	3.0678	Endothelin ETA Antagonist
	2	2.3462	Endothelin Antagonist
	3	1.9745	Bradykinin BK1 Antagonist
	4	1.7655	5 HT2A Antagonist
	5	1.5033	Cytochrome P450 Oxidase Inhibitor
	6	1.0956	Endothelin ETB Antagonist
	7	1.0634	Calcitonin Analog
	8	0.9058	5 HT2 Antagonist
	9	0.4644	Topoisomerase Inhibitor
	10	0.37	Follicle-stimulating hormone
	…	…	…
	N/A	**No score > 0**	**Adrenergic (beta) Blocker**
Rescriptor	1	17.515	hepatitis B virus reverse transcriptase [-]
	2	6.7161	tyrosine-protein kinase receptor FLT3 [-]
	3	5.8913	vanilloid receptor subtype 1 [-]
	4	5.2463	RNA-directed DNA polymerase, DNA nucleotidyltransferase, revertase [-]
	5	4.0386	growth hormone-releasing hormone receptor [-]
	6	3.9353	vanilloid receptor subtype 1
	7	3.7865	delta-type opioid receptor [-]
	8	3.7348	calcitonin receptor precursor [-]
	9	3.7348	multidrug resistance-associated protein 1
	10	3.6563	calcitonin gene-related peptide type 1 receptor
	…	…	…
	N/A	**No score > 0**	**histamine H4 receptor [-]**
DMT	1	14.681	5-hydroxytryptamine 1D receptor, serotonergic receptor
	2	**14.149**	**5-hydroxytryptamine 1B receptor, serotonergic receptor**
	3	13.235	5-hydroxytryptamine 1D receptor, serotonergic receptor [-]
	4	12.905	5-hydroxytryptamine 6 receptor, serotonergic receptor [-]
	5	12.033	5-hydroxytryptamine 1B receptor, serotonergic receptor [-]
	6	9.7713	somatostatin receptor type 4 [-]
	7	9.7031	urotensin II receptor
	8	9.5647	GRP-preffering bombesin receptor [-]
	9	9.5587	somatostatin receptor type 2
	10	9.5587	somatostatin receptor type 3
	…	…	…
	112	**0.7865**	**5-hydroxytryptamine 7 receptor, serotonergic receptor**
DMT	1	15.16	5-hydroxytryptamine 1D receptor, serotonergic receptor

Drug	Bayes Rank	Bayes Score	Off-targets predicted by Naïve Bayesian Classifier
	2	14.445	5-hydroxytryptamine 1B receptor, serotonergic receptor
	3	13.341	5-hydroxytryptamine 1D receptor, serotonergic receptor [-]
	4	12.647	5-hydroxytryptamine 6 receptor, serotonergic receptor [-]
	5	12.172	5-hydroxytryptamine 2A receptor, serotonergic receptor
	6	11.585	5-hydroxytryptamine 1B receptor, serotonergic receptor [-]
	7	9.74	somatostatin receptor type 4 [-]
	8	9.7269	urotensin II receptor
	9	9.6387	5-hydroxytryptamine 1-like receptor
	10	9.5849	somatostatin receptor type 2
	…	…	…
	122	**0.9117**	**5-hydroxytryptamine 5A receptor, serotonergic receptor [-]**
Doralese	1	12.392	Neurokinin NK2 Antagonist
	2	9.3756	5 HT1D Agonist
	3	7.735	Neurokinin Antagonist
	4	7.249	Follicle-stimulating hormone
	5	6.5828	LHRH Agonist
	6	5.2603	Substance P Antagonist
	7	4.9672	Bradykinin BK2 Antagonist
	8	4.8699	CCK Agonist
	9	4.742	Somatostatin Analog
	10	4.6065	CCK B Agonist
	…	…	…
	14	**3.228**	**Dopamine (D4) Antagonist**
Sedalande	**1**	**14.448**	**Adrenergic (alpha1) Blocker**
	2	10.676	5 HT1A Antagonist
	3	9.6677	5 HT2A Antagonist
	4	9.0628	Dopamine (D2) Antagonist
	5	6.3506	Dopamine (D3) Antagonist
	6	6.1541	5 HT1A Agonist
	7	5.9937	5 HT2 Antagonist
	8	5.2663	Antihistaminic
	9	4.1768	Neurokinin Antagonist
	10	3.846	Neurokinin NK2 Antagonist
	…	…	…
	12	**2.739**	**5 HT1D Antagonist**
Motilium	1	13.343	Adrenoceptor (beta3) Agonist
	2	12.526	Adrenergic (beta) Agonist
	3	11.512	Adrenergic (beta1) Agonist
	4	9.5046	Adrenergic (beta1) Blocker

Drug	Bayes Rank	Bayes Score	Off-targets predicted by Naïve Bayesian Classifier
	5	7.4653	LHRH Antagonist
	6	6.6349	Adrenergic (beta) Blocker
	7	6.567	Gastrin-Releasing Peptide Antagonist
	8	**2.8824**	**Adrenergic (alpha1) Blocker**
	9	2.0723	Endothelin Antagonist
	10	1.8653	Adrenoceptor (alpha2) Antagonist
Prozac	1	4.4575	Substance P Antagonist
	2	3.7865	Neurokinin Antagonist
	3	3.0703	LHRH Antagonist
	4	**3.0284**	**Adrenergic (beta) Blocker**
	5	1.9373	Bradykinin Antagonist
	6	1.7865	Glucagon Receptor Antagonist
	7	1.677	Glutathione S-Transferase Inhibitor
	8	1.5663	Bradykinin BK1 Antagonist
	9	1.5551	Bradykinin BK2 Antagonist
	10	1.5221	Phospholipase A2 Inhibitor
Dimetholizine	**1**	**14.055**	**Adrenergic (alpha1) Blocker**
	2	11.413	5 HT1A Antagonist
	3	**9.7703**	**Dopamine (D2) Antagonist**
	4	**8.7188**	**5 HT1A Agonist**
	5	6.7689	Dopamine (D3) Antagonist
	6	6.3947	Adrenergic (alpha) Blocker
	7	5.4796	Neurokinin Antagonist
	8	5.4706	5 HT2A Antagonist
	9	5.1264	5 HT1D Antagonist
	10	4.2757	Neurokinin NK2 Antagonist
Xenazine	1	65.506	Opioid Mixed Agonist-Antagonist
	2	**54.313**	**Adrenergic (alpha2) Blocker**
	3	27.638	Dopamine (D1) Agonist
	4	27.333	mu Agonist
	5	24.465	Phosphodiesterase V Inhibitor
	6	22.526	Opioid Agonist
	7	21.351	Angiotensin II AT2 Antagonist
	8	19.727	Oxytocin Antagonist
	9	17.679	Substance P Antagonist
	10	17.643	LHRH Antagonist
Fabahistin	1	11.053	5-hydroxytryptamine 5A receptor, serotonergic receptor
	2	**9.3414**	**5-hydroxytryptamine 5A receptor, serotonergic receptor [-]**
	3	4.413	delta-type opioid receptor [-]

Drug	Bayes Rank	Bayes Score	Off-targets predicted by Naïve Bayesian Classifier
	4	4.3632	delta-type opioid receptor [-]
	5	3.3544	somatostatin receptor
	6	3.2781	calcitonin gene-related peptide type 1 receptor
	7	3.259	insulin receptor [-]
	8	3.1125	secreted aspartic protease 2, candidapepsin 2, aspartate protease 2 [-]
	9	3.0665	mu-type opioid receptor
	10	3.0023	multidrug resistance-associated protein 1
Kalgut	**1**	**13.343**	**Adrenergic (beta3) Agonist**
	2	12.526	Adrenergic (beta) Agonist
	3	11.512	Adrenergic (beta1) Agonist
	4	9.5046	Adrenergic (beta1) Blocker
	5	7.4653	LHRH Antagonist
	6	6.6349	Adrenergic (beta) Blocker
	7	6.567	Gastrin-Releasing Peptide Antagonist
	8	2.8824	Adrenergic (alpha1) Blocker
	9	2.0723	Endothelin Antagonist
	10	1.8653	Adrenoceptor (alpha2) Antagonist
Prantal	**1**	**29.593**	**delta Agonist**
	2	27.51	Antihistaminic
	3	21.941	5 HT2 Antagonist
	4	12.945	Carbapenem
	5	11.189	Dopamine Agonist
	6	11.026	Estrogen Receptor Modulator
	7	9.2016	Muscarinic M3 Antagonist
	8	6.8151	Oxytocin Antagonist
	9	6.673	PAF Antagonist
	10	6.6255	5 HT2A Antagonist

SEA off-target predictions being compared are highlighted in **bold**.

Table A.2.2 MDDR drug binding predictions matching known WOMBAT targets

Table A.2.3 Examples of drug off-target predictions confirmed by literature sources but unknown to the databases

Note: Due to size constraints, Tables A.2.2 and A.2.3 are not reproduced here. They can be found in the online Supplementary Materials for the published paper at http://www.nature.com.

Table A.2.4 N,N-dimethyltryptamine affinities serotonergic receptor panel

Receptor	K_i (nM)	SEA E-value
5ht1a	183	3.60E-17
5ht1b	129	1.04E-28
5ht1d	39	9.16E-81
5ht1e	517	n/a
5ht2a	127	3.63E-14
5ht2b	184	1.55E-15
5ht2c	360	6.87E-16
5ht5a	2135	3.37E-08
5ht6	464	3.91E-30
5ht7	206	7.37E-06

Table A.2.5 Prediction and testing of Sedalande derivatives against 5-HT$_{1D}$

Compound	Name	K$_i$ (nM)	Structure (SMILES)
PDSP10815	1-(4-methoxyphenyl)-3-[4-(2-methoxyphenyl)-1-piperazinyl]-1-propanone	62	COc1ccc(cc1)C(=O)CCN2CCN(CC2)c3ccccc3OC
PDSP10813	1-(4-fluorophenyl)-4-(4-phenyl-1-piperazinyl)-1-butanone dihydrochloride	169	Fc1ccc(cc1)C(=O)CCCN2CCN(CC2)c3ccccc3
PDSP10814	3-[4-(2-fluorophenyl)-1-piperazinyl]-1-(4-methoxyphenyl)-1-propanone	185	COc1ccc(cc1)C(=O)CCN2CCN(CC2)c3ccccc3F
PDSP10812	3-{[4-(2-methoxyphenyl)-1-piperazinyl]methyl}-6-methyl-4H-chromen-4-one	464	COc1ccccc1N2CCN(CC3=COc4ccc(C)cc4C3=O)CC2
PDSP10810	6-chloro-3-{[4-(2,3-dimethylphenyl)-1-piperazinyl]methyl}-4H-chromen-4-one	546	Cc1cccc(N2CCN(CC3=COc4cc(Cl)cc4C3=O)CC2)c1C
PDSP10809	3-{[4-(2-ethoxyphenyl)-1-piperazinyl]methyl}-4H-chromen-4-one	633	CCOc1ccccc1N2CCN(CC3=COc4ccccc4C3=O)CC2
PDSP10811	6,8-dichloro-3-{[4-(2-methoxyphenyl)-1-piperazinyl]methyl}-4H-chromen-4-one	2057	COc1ccccc1N2CCN(CC3=COc4c(Cl)cc(Cl)cc4C3=O)CC2

Table A.2.6 Attempt to recapitulate SEA predictions via target sequence similarities alone

Drug	Known closest target to sequence match		Top sequence similarity matches	PSI-BLAST E-values
Rescriptor	HIV1 Reverse Transcriptase	1	HIV Integrase	1×10^{-89}
		2	Renin	0.11
		3	Factor Xa	0.40
		...	...	...
		167	**Histamine H$_4$**	**350**
RO-25-6981		1	Guanylate cyclase	7×10^{-118}
		2	GABA$_B$ receptor	1×10^{-108}
		3	5-HT$_{1A}$	6×10^{-96}
		...	...	...
	α$_1$ (for κ opioid)	**34**	**κ opioid**	**7×10^{-67}**
		...	...	...
	α$_1$ (for Dopamine D4)	**78**	**Dopamine D$_4$**	**2×10^{-40}**
		...	...	...
	NMDAR (for 5-HTT)	**90**	**5-HTT**	**0.54**
		...	...	...
	α$_1$ (for NET)	**103**	**NET**	**3.8**
Vadilex		1	Muscarinic M2	2×10^{-131}
		2	Muscarinic M3	5×10^{-127}
		3	Adrenergic (α2)	7×10^{-125}
		...	...	...
	5-HT$_{2A}$ (for μ opioid)	**32**	**μ opioid**	1×10^{-85}
		...	...	...
	NMDAR (for 5HTT)	**92**	**5-HTT**	**0.54**
Xenazine	VMAT2	1	5-HT$_{5A}$	1.6
		2	Angiotensin II AT$_1$	2.1
		3	Adenosine (A$_2$)	8.2
		...	...	...
		78	**Adrenergic (α$_2$)**	**125**
Doralese	5-HT$_{2A}$	1	5-HT$_{1C}$	6×10^{-113}
		2	5-HT$_{2C}$	6×10^{-113}
		3	5-HT$_{2B}$	2×10^{-101}
		...	...	...
		76	**Dopamine (D$_4$)**	**3×10^{-44}**
Prantal	Muscarinic M$_3$	1	Muscarinic M$_2$	3×10^{-144}
		2	Muscarinic M$_1$	2×10^{-136}
		3	Histamine H$_1$	2×10^{-134}
		...	...	...
		26	**δ opioid**	**3×10^{-65}**

Drug	Known closest target to sequence match	Top sequence similarity matches		PSI-BLAST E-values
DMT	5-HT$_{1D}$ (for 5-HT$_{1B}$ match)	**1**	**5-HT$_{1B}$**	**3×10^{-126}**
		2	Adrenergic (α_2)	9×10^{-104}
		3	Adrenergic (α_1)	7×10^{-103}
		...	...	...
	5-HT$_{2A}$ (for 5-HT$_7$ match)	**11**	**5-HT$_7$**	**2×10^{-97}**
		...	...	...
	5-HT$_6$ (for 5-HT$_{5A}$ match)	**24**	**5-HT$_{5A}$**	**1×10^{-88}**
Fabahistin	5-HT$_6$	1	Muscarinic M$_2$	2×10^{-131}
		2	Muscarinic M$_3$	5×10^{-127}
		3	Adrenergic (α_2)	7×10^{-125}
		...	...	...
		21	**5-HT$_{5A}$**	**1×10^{-88}**
Paxil	Adrenergic (α_1)	1	Histamine H$_1$	2×10^{-134}
		2	Dopamine (D$_2$)	2×10^{-125}
		3	Adrenergic (α_2)	7×10^{-124}
		4	5-HT$_{1A}$	2×10^{-105}
		...	...	...
		22	**Adrenergic (β_1)**	**2×10^{-81}**
Prozac	5-HT$_{2A}$	1	Muscarinic M$_2$	2×10^{-131}
		2	Muscarinic M$_3$	5×10^{-127}
		3	Adrenergic (α_2)	7×10^{-125}
		4	Dopamine (D$_2$)	1×10^{-123}
		...	...	...
		22	**Adrenergic (β_1)**	**1×10^{-92}**
Sedalande	5-HT$_{2A}$	1	5-HT$_{1C}$	6×10^{-113}
		2	5-HT$_{2C}$	6×10^{-113}
		3	Adrenergic (α_2)	6×10^{-104}
		4	5-HT$_{1B}$	5×10^{-103}
		...	...	...
		7	**Adrenergic (α_1)**	**7×10^{-103}**
		...	...	...
		19	**5-HT$_{1D}$**	**7×10^{-94}**
Dimetholizine	Histamine H$_1$	1	Muscarinic M$_2$	2×10^{-131}
		2	Muscarinic M$_3$	5×10^{-127}
		3	Adrenergic (α_2)	7×10^{-125}
		...	...	...
		6	**Dopamine (D$_2$)**	**1×10^{-123}**
		...	...	...
		8	**5-HT$_{1A}$**	**4×10^{-118}**
		...	...	...
		18	**Adrenergic (α_1)**	**4×10^{-63}**

Drug	Known closest target to sequence match	Top sequence similarity matches		PSI-BLAST E-values
Motilium	$5\text{-}HT_{2A}$	1	$5\text{-}HT_{1C}$	6×10^{-113}
		2	$5\text{-}HT_{2C}$	6×10^{-113}
		3	**Adrenergic (α_1)**	**7×10^{-103}**
Kalgut	Adrenergic (β_1)	**1**	**Adrenergic (β_3)**	**2×10^{-96}**
		2	$5\text{-}HT_4$	3×10^{-95}
		3	Dopamine D_1	1×10^{-94}

Novel off-targets predicted by SEA are highlighted in **bold**.

Table A.2.7 Off-target predictions with observed binding affinities > 10 μM

Drug	Existing MDL Annotation	SEA E-value	Predicted target
Emilace	Antipsychotic Dopamine D2 Antagonist	7.50E-103	5 HT4 Agonist
Centrax	Anti-Anxiety Agents GABA Modulators	3.80E-59	CCK B Antagonist
Valium	Anticonvulsant Anxiolytic Benzodiazepine Agonist	3.10E-47	CCK B Antagonist
Dromoran	Analgesics, Opioid Narcotics	6.90E-46	Dopamine (D2) Agonist
Zatebradine	Bradycardic	2.80E-30	Dopamine (D1) Antagonist
Doralese	Adrenergic (alpha1) Blocker Antihypertensive Antimigraine Prostate Disorders, Agent for	1.45E-29	Neurokinin Antagonist (by functional assays vs. NK1 and NK2)
Duocaine	Anesthetic	6.40E-21	kappa agonist

Table A.2.8 Datasets and descriptors used for each novel SEA prediction

	Drug	SEA Off-target predictions	Database Used	Descriptor Used
1	Dimetholizine	5-HT1A Antagonist	MDDR	ECFP4
2	Dimetholizine	Adrenergic alpha1 blocker	MDDR	ECFP4
3	Dimetholizine	Dopamine (D2) Antagonist	MDDR	ECFP4
4	DMT	5-HT1B Agonist	WOMBAT 1uM	ECFP4
5	DMT	5-HT5A Antagonist	WOMBAT 10uM	ECFP4
6	DMT	5-HT7 Modulator	WOMBAT 1uM	ECFP4
7	Doralese	Dopamine (D4) Antagonist	MDDR	ECFP4
8	Fabahistin	5-HT5A Antagonist	WOMBAT 10uM	ECFP4
9	Kalgut	Adrenoceptor (beta3) Agonist	MDDR	ECFP4
10	Motilium	Adrenergic (alpha1) Blocker	MDDR	ECFP4
11	Paxil	Adrenergic (beta) blocker	MDDR	ECFP4
12	Prantal	delta Agonist	MDDR	DAYLIGHT
13	Prozac	Adrenergic (beta) blocker	MDDR	ECFP4
14	Rescriptor	Histamine H4 Antagonist	WOMBAT 1uM	ECFP4
15	RO-25-6981	5-HTT	STARLITE	ECFP4
16	RO-25-6981	Dopamine (D4)	STARLITE	ECFP4
17	RO-25-6981	kappa Opioid	STARLITE	ECFP4
18	RO-25-6981	NET	STARLITE	ECFP4
19	Sedalande	5-HT1D Antagonist	MDDR	ECFP4
20	Sedalande	Adrenergic alpha1 blocker	MDDR	ECFP4
21	Vadilex	5-HTT	STARLITE	ECFP4
22	Vadilex	mu Opioid Receptor	STARLITE	ECFP4
23	Xenazine	Adrenergic (alpha2) Blocker	MDDR	DAYLIGHT

This table lists the particular reference database and descriptor used in each SEA off-target prediction from **Table 2.1** and **Table 2.2** of the main text. Over the course of this work, multiple reference databases became available at different times: First the MDDR, then WOMBAT, and finally StARlite.

Whereas many of these predictions may be recapitulated regardless of database choice, some databases contain targets that others lack—for instance, the MDDR lacks ligands for Histamine H_4. For our 5-HT_{5A} ligand set, we accepted all ligands with $\leq$ 10 μM affinity (instead of our 1 μM default WOMBAT cutoff) because we have found this cutoff to yield better 5-HT_{5A} SEA predictions, in unpublished work.

Table A.2.9 MDDR to WOMBAT mapping

Note: Due to size constraints, Table A.2.9 is not reproduced here. It can be found in the online Supplementary Materials for the published paper at http://www.nature.com.

Table A.2.10 Related phrases used in novelty filtering

Phrase	IsRelatedTo
5ht	serotonin;5-hydroxytryptamine
adenosine	metabolite
adrenergic	calcium;adrenoceptor
androgen	sterone;sterol;enol;estone;estrogen;estrel;contraceptive;cortico;cholest;inflammatory;anabolic;aromatase;sterone;progest;steroid
bacterial	bactam;lactam;illin;ceph
biotic	penem;ceph;bactam
bradykinin	inflammatory
cephalosporin	bacterial;ceph;penicillin
delta	narcotic;analgesic
dihydrofolate	folic;neoplastic
folyl	folic
folylpolyglutamate	folic;neoplastic
insulin	tose
kappa	narcotic;analgesic
mu	narcotic;analgesic
muscarinic	parasympatholytic;cholinergic
opioid	narcotic;analgesic
penicillin	illin;bacterial
prostaglandin	anabolic;androgen;cortico;inflammatory;sterone
serotonin	5ht1;5ht2;5ht3;5ht4;5ht5;5ht6;5ht7
steroid	contraceptive;cortico;inflammatory;sterone;estrogen;androgen;aromatase
thymidine	metabolite;antiviral
thymidylate	folic;neoplastic

A.3 Supplementary material for Chapter 3

I. Supplementary datasets

This chapter references three supplementary datasets, which provide the original data used for calculation of the target-target similarities, as well as the full matrix of SEA similarities calculated among then. These datasets are too large to include here, but are freely available online at http://www.ploscompbiol.org.

Appendix B:

Off-target networks derived from ligand set similarity

Michael J Keiser[11] and Jérôme Hert

B.1 Abstract

Chemically similar drugs often bind biologically diverse protein targets, and proteins with similar sequences or structures do not always recognize the same ligands. How can we uncover the pharmacological relationships among proteins, when drugs may bind them in defiance of bioinformatic criteria? Here we consider a technique that quantitatively relates proteins based on the chemical similarity of their ligands. Starting with tens of thousands of ligands organized into sets for hundreds of drug targets, we calculated the similarity among sets using ligand topology. We developed a statistical model to rank the resulting scores, which were then expressed in minimum spanning trees. We have shown that biologically sensible groups of targets emerged from these maps, as well as experimentally-validated predictions of drug off-target effects.

Key Words: SEA, expectation value, target network, polypharmacology, off-targets

B.2 Introduction

How similar are two proteins? Typically, proteins are compared using bioinformatics approaches based on sequence or structure. While these methods quantify historical protein divergence,

---

[11] Corresponding author: michael.james.keiser@ucsf.edu; Department of Pharmaceutical Chemistry, University of California San Francisco, 1700 4th St, San Francisco California 94143-2550, USA.

drugs and other small molecules often bind to targets that are unrelated from an evolutionary standpoint (**1, 2**). For example, the enzymes thymidylate synthase, dihydrofolate reductase and glycinamide ribonucleotide formyltransferase have no substantial sequence identity or structural similarity but they all recognize folic acid derivatives and are inhibited by antifolates. Similarly, the drug methadone binds both the μ-opioid receptor, a GPCR, and the structurally-unrelated *N*-methyl-D-aspartate receptor, an ion channel. Polypharmacology, the ability of chemically similar drugs to bind biologically diverse proteins, has inspired recent efforts to find protein relationships by means other than their sequence or structure (**3-5**).

The Similarity Ensemble Approach (SEA) considers proteins from a chemo-centric point of view, relating them through the chemical similarity of their ligands (**6**). The idea is that similar molecules have similar biological profiles (**7**) and bind similar targets (**8, 9**). This technique links hundreds of ligand-sets—and correspondingly their protein targets—together in minimal spanning trees where biologically related proteins cluster together as an emergent property (*see* **Figure 1**). These networks are robust (**10**) and may be used to predict off-target effects (**6**). The similarities among ligand-sets may reveal the pharmacological relationships of the targets whose actions they modulate.
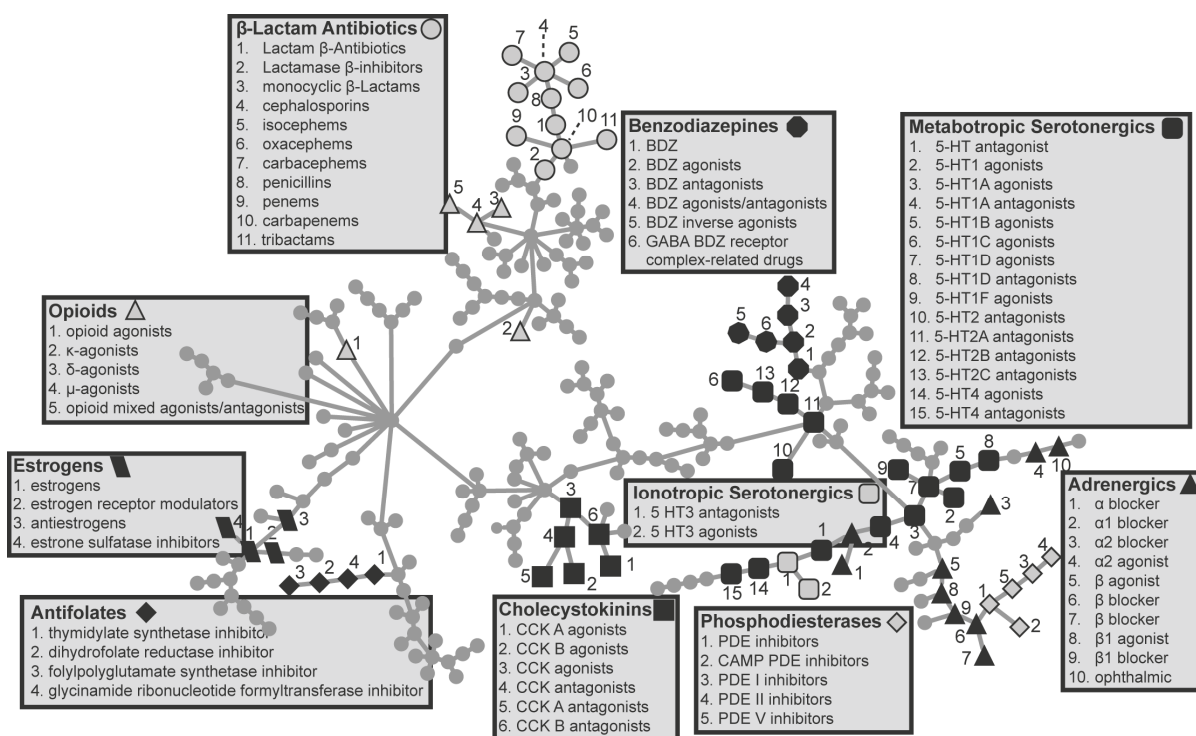
Figure B.1 Pharmacological network of the MDDR drug targets

Each vertex represents a ligand-set, and hence a protein target. The vertices are linked together by their SEA *E*-values (edges) and organized into a minimum spanning tree. Several protein families are highlighted to emphasis the natural clustering that emerges.

How does SEA work? An overview of the different stages is available in **Figure 2**. The similarity between two ligand-sets is first approximated by summing the similarity scores of molecule pairs across the sets (*see* **Figure 2B**). In itself, the resulting *raw score* is not a good estimate of the overall similarity of the sets, as it does not discriminate relevant similarities from random and depends on the number of ligands in each set. SEA corrects for these shortcomings via a statistically-determined threshold—pairs of molecules that score below it are discarded and do not contribute to the overall set similarity. We then convert the raw score to a size-bias-free z-score using the mean and standard deviation of raw scores modeled from sets of random molecules. Finally, we express the similarity score between two sets as an *E*-value, *i.e.*, the

probability of a given z-score that high or better to be observed from random data. Small *E*-values, then, reflect relationships between ligand-sets that are stronger than would be expected by random chance alone.
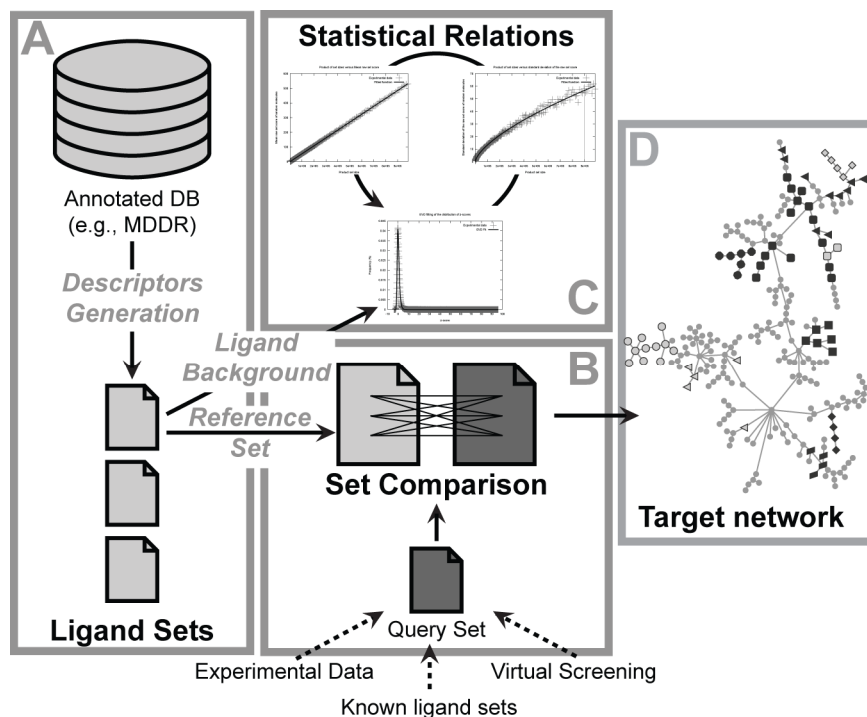

Figure B.2 Method overview

Ligand-sets derived from existing databases [A] are used in set-wise comparisons [B] against a query set, the result of which is quantified by the statistical model inferred from that reference database [C]. The generated probabilistic data can be used to construct chemical mappings of the ligand-sets and correspondingly the biological targets [D].

B.3 Materials

3. A reference database of chemical structures, annotated by therapeutic indication or mechanism of action. For the purpose of illustration, we used the *MDL Drug Data Report* (MDDR) (**11**) which contains 65,367 molecules organized in 249 sets (*see* **Note 1**).

4. A molecular descriptor generator to encode the structural information of the compounds. We obtained the best results with 2-dimensional fingerprints based on topology of the molecules such as the 2048-bit default Daylight or 1024-bit folded Scitegic ECFP_4 descriptors (*see* **Note 2**).

5. A similarity coefficient, such as the Tanimoto coefficient (*see* **Note 3**).

B.3.1 Calculating the parameters of the reference database

1. A fitter program to calculate nonlinear regressions (*see* **Note 4**).

B.3.2 Calculating set-wise similarity ensembles

1. No additional materials are necessary.

B.3.3 Building a similarity network

1. A graph visualization program, such as Cytoscape (**12**).

B.4 Methods

SEA quantifies the similarity among sets of compounds which may be organized by the targets they modulate, the therapeutic indications they address, their activity in a high-throughput screening campaign, or a variety of other criteria. So far, we have focused on sets organized by targets, but SEA can be used with other annotations.

Before comparing any sets with SEA, the parameters of the background database—generally the one containing the sets one wishes to compare to—need to be calculated. While this step is computationally intensive, it is only required once for a given database, molecular descriptor and similarity coefficient (s*ee* **Section B.4.1**). Once the optimal threshold t_i and the formulae of the mean y_μ and standard deviation y_σ as a function of the product of the sets' sizes

$(|a| \times |b|)$ have been determined, SEA can be applied to quantify set similarity (*see* **Section B.4.2**).

B.4.1 Calculating the parameters of the reference database

In this section, we generate thousands of randomly-populated pairs of ligand sets, and determine the uncorrected similarity among them. We use these "random" similarities to build an empirical model of background chemical similarity. The particular choice of chemical database will determine the type of background: KEGG molecules will yield a metabolic background, whereas ZINC molecules will produce drug- or lead-like backgrounds (depending on the exact subset used). It is preferable to choose as large a database as possible; those in excess of 100,000 molecules are often ideal.

2. Choose minimum and maximum set sizes s_{min} and s_{max} for sampling, such that they will be representative of molecule sets annotated in the database (*see* **Note 5**).

3. Sample at least 1,000 integers s_i from the range $(s_{min} \times s_{min})$ to $(s_{max} \times s_{max})$ (*see* **Note 6**).

4. For each product of sets' sizes s_i, calculate all of its integer factors f_i, such that $s_{min} \leq f_i \leq s_{max}$.

5. For each s_i, choose 30 of its f_i at random and construct two sets a and b, consisting of f_i and s_i/f_i molecules respectively, randomly selected from the background molecule database (*see* **Note 7**).

6. For each pair of sets a and b, calculate standard chemical similarities $c_{a,b}$ for each pair of ligands across the sets using your previously chosen chemical similarity descriptor and coefficient.

7. For t_i, where $0 \leq t_i < 1$ with step size 0.01, calculate a "raw score" $r_{a,b}(t_i)$ equal to the sum of all $c_{a,b}$ where $c_{a,b} > t_i$. Store all calculated $r_{a,b}(t_i)$, along with the sizes of sets a and b (*see* **Note 8**).

8. For each t_i, plot all $r_{a,b}(t_i)$ scores vs. the product of set sizes a and b, e.g., plot all points $(|a| \times |b|, r_{a,b})$. There should be 100 plots (*see* **Note 9**), each corresponding to a particular choice of t_i.

9. For each plot, use the nonlinear fitter to determine the mean expected random chemical similarity (*see* **Figure 2C** and **Figure 3A**). Typically, an equation of the formula $y_\mu = mx^n + p$ will be appropriate (*see* **Note 10**).

10. For each plot, bin the data by the x-axis values, such that each bin ideally has no fewer than five data points. Given the previously fitted y_μ, calculate the standard deviation of each bin with Laplacian correction, and fit the resulting standard deviation points nonlinearly (*see* **Figure 2C** and **Figure 3B**). Again, $y_\sigma = qx^r + s$ will typically be appropriate.

11. For each plot, use the fitted y_μ and y_σ to transform all original points $(|a| \times |b|, r_{a,b})$ to their z-scores $z_{a,b} = (r_{a,b} - y_\mu(|a| \times |b|)) / y_\sigma(|a| \times |b|)$ (*see* **Note 11**). Construct a histogram of these z-scores.

12. For each histogram, nonlinearly fit the data to Gaussian and extreme value type I (EVD) distributions (*see* **Note 12**, **Figure 2C**, and **Figure 3C**).

13. Based on goodness-of-fit, such as each fit's observed-vs.-expected χ^2 value, select the threshold choice t_i, such that the histogram best fits an EVD instead of a Gaussian distribution (*see* **Note 13**).

14. Record the chosen t_i and that t_i's formulae for y_μ and y_σ. These values comprise the random background model. All other plots, histograms, and formulae may be discarded at this point.
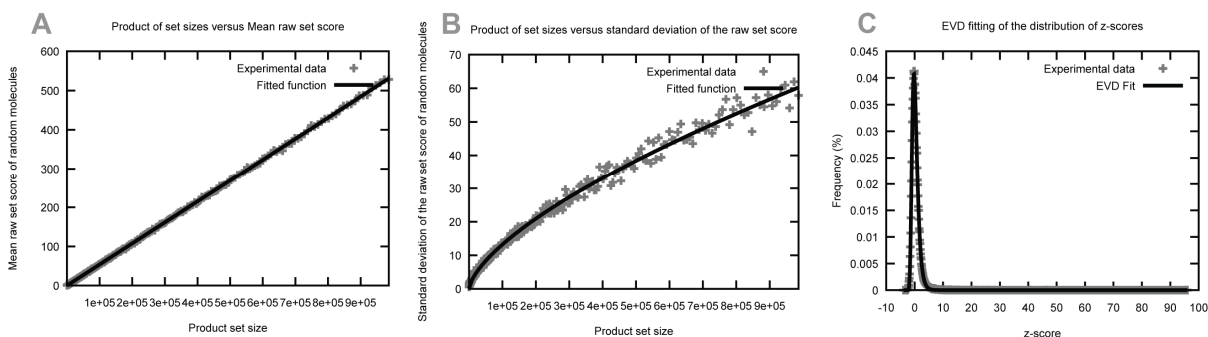

Figure B.3 Statistical models

[A] Correlation between the product of sets' sizes and the mean of the raw score. The fitted function typically corresponds to an equation of the formula $y_\mu = mx^n + p$ with $n = 1$. [B] Correlation between the product of sets' sizes and the standard deviation of the raw score. The fitted function typically corresponds to an equation of the formula $y_\sigma = qx^r + s$, with $0.6 < r < 0.7$. [C] Distribution of the z-scores obtained from random data using ECFP_4 fingerprints, with a similarity score threshold (t_i) of 0.57 and fitted to an extreme value distribution.

B.4.2 Calculating set-wise similarity ensembles

To calculate the set-wise similarity among sets of ligands, we reuse much of the machinery developed to calculate background models, and extend it to calculate *E*-values. By exhaustively comparing all pairs of sets across two collections (databases), we can then rank the top hits for any particular ligand set.

In practice, a ligand set should not comprise fewer than ten ligands, unless you intend to compare it against large sets only. For instance, it would not be statistically reliable to compare two sets of five ligands each, but a set of five ligands compared against a set of thirty should be

acceptable. Although the particular choice of set size should depend on the diversity of ligands within a set, a good rule of thumb is to build sets such that the product of the set sizes will be no less than 100 (e.g., the product of set sizes is 25 for the five-by-five case, and 150 for the five-by-thirty case mentioned above).

1. To calculate similarity ensembles, choose two collections of sets C_a and C_b to compare (*see* **Note 14**).

2. For each set a and b from collections C_a and C_b, respectively, calculate $r_{a,b}(t_i)$ as previously described using only the optimal threshold t_i from the background model. Be sure to use the actual molecule structures annotated for each set.

3. Transform each $r_{a,b}(t_i)$ to z-score $z_{a,b}$ as described in **Section B.4.1.10**.

4. Transform each $z = z_{a,b}$ to p-value $P(Z > z) = 1 - \exp(-e^{-z\pi/\mathrm{sqrt}(6)-\Gamma'(1)})$, where $\Gamma'(1)$ is the Euler-Mascheroni constant (≈ 0.577215665) (*see* **Note 15**).

5. Optionally, the p-value may be transformed to a BLAST-like E-value by calculating $E(z) = P(Z > z) \times n_{db}$, where $n_{db} =$ the number of set-vs.-set comparisons made when comparing all sets from collection C_a against all sets from collection C_b. Typically, $n_{db} = |C_a| \times |C_b|$.

6. For each set a, rank all sets b_i from C_b by their E-value, where values approaching zero are the best scores (*see* **Note 16**).

B.4.3 Building a similarity network

A similarity network is a graphical view of the E-value relationships among all ligand sets in a particular database (*see* **Note 17**). If these ligand sets represent particular drug targets, for instance, it is a visualization of the significant chemical similarity present among these targets (*see* **Figure 1**).

1. Calculate the similarity ensemble E-values between all sets a_i and a_j from C_a versus itself (*see* **Note 18**), as previously described.

2. The resulting matrix of E-values defines a strongly connected graph, where each node corresponds to a molecule set and each edge to the E-value between two sets (*see* **Note 19**).

3. We use Kruskal's algorithm (**13**) to construct a minimum spanning tree (MST):

    a. Create a set S_{tree} that initially contains all individual nodes, unconnected. We refer to elements of S_{tree} as "trees."

    b. Create a set S_e that contains all possible edges e_i (E-values).

    c. While S_e is not empty

        i. Remove the minimum-weighted (best) edge e_{min} from S_e

        ii. If e_{min} connects two existing trees t_a and t_b in S_{tree}

            1. Remove t_a and t_b from S_{tree}, connect them into a single new tree t_{ab} using e_{min}, and add t_{ab} back into S_{tree}.

        iii. Else, discard e_{min}.

    d. When the algorithm finishes, S_{tree} will contain only one tree, which is the graph's MST.

B.5 Notes

1. Examples of other freely or commercially available annotated chemogenomics databases include *WOMBAT*, *KEGG*, and *DrugBank*. Note, however, that SEA can be used with any kind of annotation and is not limited to ligand—target association.

2. For efficiency, the steps in **Methods** will be faster if fingerprints are pre-calculated and stored for each molecule.

3. While it is not technically necessary, we assume that the similarity coefficient is normalized from 0.0 to 1.0. If not, choose appropriate bounds for the range of t_i thresholds discussed in **Section B.4.1.6**.

4. The open-source Scientific Python (SciPy) package (**14**) provides a least-squares optimizer that can be used for fitting nonlinear regressions.

5. If you are unsure of appropriate values, use $s_{min} = 10$ and $s_{max} = 300$.

6. More than 1,000 points may be sampled, but in our experience this does not yield a substantial difference in the final model.

7. If there are fewer than 30 distinct factors f_i for a particular integer s_i, randomly sample from the available f_i 30 times. Sampling more than 30 points is also acceptable, depending on the diversity of the background database and computational resources.

8. These raw scores are the "random" similarities that form the background model. Besides the choice of similarity descriptor and coefficient, the threshold t_i is the only settable major SEA parameter. By sampling across the range of t_i choices, we will be able to determine an optimal choice of t_i in later steps.

9. For the steps plotting these data (and later, the histograms), you need not actually draw out the full plots. All that is strictly necessary is that your data is formatted appropriately for input into your chosen fitter. Using SciPy, for instance, it is enough to store these data points in internal arrays.

10. In our experience, the mean raw score fit y_μ has always been linear.

11. The z-score is the number of standard deviations by which a particular raw score exceeds the expected mean.

12. You may use the "norm" and "gumbel_r" SciPy data-types for Gaussian and extreme value type I distributions, respectively.

13. There is currently no formal justification for choosing the t_i threshold, but this approach is consistent and enriches for a BLAST-like background probability distribution. Some experiments also suggest that this choice is reasonable, as thresholds derived from retrospective cross-fold analysis are identical or close to the threshold t_i (unpublished).

14. One such collection may be built from the annotated molecular structure database. The second may be the exact same collection (for symmetric comparisons), or derived from a different database of annotated molecules.

15. This formula converts EVD z-scores to their p-values, where the p-value expresses the probability of finding a z-score that strong or better, by random chance alone.

16. An E-value of 1 or higher is not statistically significant. The similarity between two sets becomes significant when it is at least one order of magnitude smaller than random chance alone, *i.e.*, 10^{-1}. Sets that are highly similar have E-values $<< 10^{-50}$, although there is no single cutoff for E-value significance. The SEA Search tool at http://sea.docking.org may also be used check the accuracy of the z-scores and E-values calculated in **Section B.4.2**.

17. While there are many appropriate graph-theoretic approaches, we have chosen a minimum spanning tree (MST). A MST is a selection over all graph edges (E-values) such that the resulting tree links all nodes (ligand sets) at lowest "cost" to the network as a whole. For example, an edge with an E-value approaching zero has a lower cost to the tree than one with an E-value of 1. The resulting MST will preferentially include only those edges with the smallest E-values. It may be interpreted as a simplified view of higher-dimensional chemical similarity space.

18. These instructions apply only to symmetric collection comparisons, e.g., $C_a = C_b$.

19. You may either (a) use Cytoscape to filter out all edges above an E-value threshold of your choice, or (b) construct a global minimum spanning tree.

B.6 Acknowledgements

B.7 References

1. Roth B, Sheffler D, Kroeze W. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nature Reviews Drug Discovery 2004;3:353-9.

2. Paolini G, Shapland R, van Hoorn W, Mason J, Hopkins A. Global mapping of pharmacological space. Nature Biotechnology 2006;24:805-15.

3. Nidhi, Glick M, Davies JW, Jenkins JL. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model 2006;46:1124-33.

4. Izrailev S, Farnum MA. Enzyme classification by ligand binding. Proteins 2004;57:711-24.

5. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science 2008;321:263-6.

6. Keiser M, Roth B, Armbruster B, Ernsberger P, Irwin J, Shoichet B. Relating protein pharmacology by ligand chemistry. Nature Biotechnology 2007;25:197-206.

7. Johnson MA, Maggiora GM, eds. Concepts and Applications of Molecular Similarity. New York: John Wiley; 1990.

8. Frye SV. Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. Chemistry & Biology 1999;6:R3-R7.

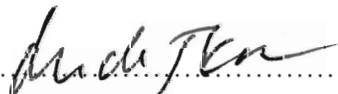9. Jacoby E, Schuffenhauer A, Floersheim P. Chemogenomics knowledge-based

strategies in drug discovery. Drug News & Perspectives 2003;16:93-9102.

10. Hert J, Keiser M, Irwin J, Oprea T, Shoichet B. Quantifying the Relationships among Drug Classes. Journal of Chemical Information and Modeling 2008;48:755-65.

11. The MDL Drug Data Report Database is available from MDL Information Systems, Inc. (Accessed at http://mdl.com.)

12. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 2003;13:2498-504.

13. Kruskal J. On the shortest spanning subtree and the traveling salesman problem. Proceedings of the American Mathematical Society 1956;7:48-50.

14. SciPy: Open Source Scientific Tools for Python. 2001. (Accessed at http://www.scipy.org.)
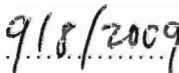
Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations 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 or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

_____   _____9/8/2009_____

Michael James Keiser   Date