Title
Rationalizing Drug Pharmacology based on Computational Methods

Permalink
https://escholarship.org/uc/item/3mf2z296

Author
Yera, Emmanuel Ramon

Publication Date
2013

Peer reviewed|Thesis/dissertation
Rationalizing Drug Pharmacology based on Computational Methods

by

Emmanuel Ramón Yera

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
Copyright 2013
by
Emmanuel Ramón Yera
Acknowledgments

Successfully (and happily) completing a graduate program requires the help and support of many. First, I would like to thank my research adviser, Ajay Jain. He has been an amazing mentor and I am extremely grateful for his input throughout the years. If it were not for his constant help, encouragement, intellectual insight, and feedback, who knows where I would be today. Also, I would like to thank the members of the Jain lab. I am extremely thankful for the constant guidance and mentorship that Ann Cleves provided. My graduate student experience would have been very different if it were not for her. I would also like to thank the fellow graduate students in the lab: Rocco Varela and Russell Spitzer. Graduate student life is hard, but having them around to discuss algorithms and joke around helped get past the difficult times.

Next, I would like to thank the members of my dissertation committee: Thomas Ferrin and Michael Fischbach. Getting concrete and insightful feedback about my dissertation work has been critical in being productive and moving forward with my work. I would also like to thank the professors that I was fortunate enough to work with during rotations through their labs: Patricia Babbitt, Michael McManus, Esteban Gonzalez Burchard, and Joseph DeRisi. Investigating and learning about a wide range of interesting and difficult biological problems has been truly important in my professional growth. Also, I would like to thank Herbert Silber and Frank Bayliss for providing support and encouragement in my pursuit of a doctorate degree.

Without the support of my family throughout, my academic career success would have been very difficult if not impossible. They have been a source of inspiration and joy outside of the lab. I am very thankful for all the support of my loving parents, Maria Julia and Ricardo Yera. They have provided endless
support for me all these years which I feel very blessed to have had. They taught me so much and always pushed me to work hard knowing that if I put my mind on something that I could achieve it. Also, I would like to thank my sisters, Leticia Yera-Aceves and Judith Yera-Juarez, and their families for their love and support.

Last but not least, I would like to thank my beautiful wife, Maria Jose Moraga-Yera. She has been my rock throughout my graduate career, providing encouraging words during the lows and celebrating with me during the highs. I would like to thank her for listening to my presentations at odd hours of the day and for providing insightful comments when reviewing my research papers. Most importantly, I am thankful for her infinite love and patience.
Abstract

Rationalizing Drug Pharmacology based on Computational Methods
Emmanuel Ramón Yera

Large-scale experimental determination of the protein targets of small molecules is both time-consuming and costly. Computational methods can be used to predict interactions between small molecule and targets, which can help experimentalists find new therapeutic targets or off-targets responsible for undesired side-effects. A data fusion framework for combining multiple similarity computations and a novel method for drawing relationships between drugs based on their clinical effect was developed. Small molecules may be quantitatively compared based on 2D topological structural considerations, based on 3D characteristics directly related to binding, and based on their clinical effects. Given a new molecule along with a set of molecules sharing some biological effect, a single score based on comparison to the known set is produced, reflecting either 2D similarity, 3D similarity, clinical effects similarity or their combination. The methods were systematically applied to a large set of FDA approved drugs (nearly two-thirds). For prediction of off-target effects, the performance of 3D-similarity over either 2D or clinical effects similarity alone was substantial, and there was added benefit from combining all of the methods. In addition to assessing predictive accuracy of the different similarity methods, the relationship between chemical similarity and pharmacological novelty was studied with regards to protein target modulation and clinical effects. Drug pairs that shared high 3D similarity but low 2D similarity (i.e. having different underlying scaffolds) were shown to be much more likely to exhibit pharmacologically relevant differences in terms of specific target modulation and differences in clinical effects.
# Table of Contents

1 Introduction ................................................................. 1

2 Statistical Framework for Target Prediction ........................ 4
  2.1 Abstract ............................................................... 4
  2.2 Introduction .......................................................... 4
  2.3 Methods and Data ...................................................... 9
    2.3.1 Molecular Data Sets ........................................... 9
    2.3.2 Computational Methods ....................................... 12
    2.3.3 Computational Procedures .................................. 20
  2.4 Results and Discussion ............................................... 20
    2.4.1 Predicting On- and Off-Target Effects ....................... 21
    2.4.2 Excess Targets: False Positive Predictions ................ 25
    2.4.3 Relationship to Other Methods ............................... 30
    2.4.4 Off-Target Prediction: Detection of Surprising Effects ... 35
    2.4.5 Recent Off-Target Predictions ................................ 38
    2.4.6 Relationship of Structural Novelty to Pharmacology ....... 42
  2.5 Conclusions ............................................................ 47

3 Computationally Efficient Similarity Score Calculation and Normalization ........................................... 50
  3.1 Introduction ............................................................ 50
  3.2 3D Molecular Similarity with Surflex-Sim ......................... 51
3.2.1 Similarity Metrics: Volume Based and Surface Based . . . 53
3.2.2 Electrostatic Similarity ........................................ 60
3.2.3 Other Methods .................................................... 62
3.2.4 Conformation and Alignment Optimization ................. 63
3.2.5 Conformational Optimization .................................. 64
3.2.6 Alignment Optimization ......................................... 66
3.2.7 Tradeoffs Between Speed and Thoroughness ............... 69
3.2.8 Approximations for 3D Similarity: Extreme Speed ........ 73
3.2.9 Summary .......................................................... 74
3.3 Imprint-based p-value normalization .............................. 74
3.4 Conclusion ............................................................ 77

4 Prediction of Off-Target Biological Effects through Data Fusion 82
4.1 Abstract .................................................................. 82
4.2 Introduction ............................................................ 83
4.3 Methods and Data .................................................... 87
   4.3.1 Molecular Data Sets ............................................. 87
   4.3.2 Patient Package Insert Similarity ......................... 90
   4.3.3 Target Prediction using Patient Package Insert Similarity . 94
   4.3.4 3D Similarity and \( p \)-value Computation with Surflex-Sim . 96
   4.3.5 GSIM-2D .......................................................... 100
4.4 Computational Procedures .......................................... 101
4.5 Results and Discussion ............................................ 101
4.5.1 Relationship between Structural Novelty and Clinical Effects 101
4.5.2 Internal SPDB Validation: On and Off Target Effects . . 106
4.5.3 Prediction of New Drug-Target Pairs within ChEMBL . . 115
4.5.4 Predicting Novel Drug/Target Interactions . . . . . . . . 121
4.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 128

5 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 130

References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133
List of Tables

3.1 Parameters for the mean linear model. ................................. 80
3.2 Parameters for the standard deviation linear model. ................. 81

4.1 Proportions correctly predicted for primary and secondary targets
at various log-odds thresholds using different similarity methods. . 110
4.2 Proportions correctly predicted for primary and secondary targets
at various log-odds thresholds using different combinations of sim-
ilarity methods. ........................................................................ 111
4.3 Proportions correctly predicted for new drug target interactions
from ChEMBL at various log-odds thresholds using different sim-
ilarity methods. ........................................................................ 115
4.4 Proportions correctly predicted for new drug target interactions
from ChEMBL at various log-odds thresholds using different com-
binations of similarity methods. ................................................. 116
List of Figures

2.1 Example of structural novelty leading to novelty in pharmacological action. ........................................ 6
2.2 Normalizing raw similarity scores to $p$-values. ................. 15
2.3 Relationship between raw 2D and 3D similarity scores and corresponding $p$-values. ................................. 17
2.4 Computing the likelihood that a test molecule shares annotation with a set of molecules with a known biological effect. .......................... 18
2.5 3D and 2D similarity log-odds calculation example. ............. 19
2.6 Proportion of drug targets correctly predicted for primary and secondary targets. .................................. 22
2.7 2D similarity can sometimes correctly predict a target. ........... 24
2.8 3D similarity method more accurately predicts off-targets. ...... 26
2.9 Combining 2D and 3D similarity methods makes a small improvement over 3D alone. ............................... 27
2.10 The design intention and surprising effects of some older drugs. . 37
2.11 Example of drugs identified by SEA where the off-target potencies are weaker than on-target drug effects. ............... 41
2.12 Structurally similar drugs are more likely to modulate identical targets relative to structurally novel drug pairs. ............... 44
2.13 Examples of structural novelty leading to novel pharmacology. . 45
3.1 Example of 3D molecular similarity using Surflex-Sim. ............ 51
3.2 Volumetric and surface-based molecular density functions for benzamidine. ........................................... 55
3.3 Volume-based and surface-based alignment of aminomethylcyclohexane to benzamidine. ............................ 59
3.4 Relationship of molecular alignments to underlying similarity functions. .................................................... 60
3.5 Optimal alignment of two muscarinic antagonists using surface shape and polarity. .................................... 62
3.6 Conformational sampling independent of molecular alignment. ................................................................. 65
3.7 Canonicalization of alignment and sampling thoroughness. ......................................................................... 67
3.8 Pose optimization tradeoff of speed versus quality. ....................................................................................... 69
3.9 Volume and surface shape optimization with variations of search depth. ..................................................... 71
3.10 Molecular imprint-based p-value normalization. ......................................................................................... 75
3.11 Relationship between empirical and estimated p-values and Log Odds scores. ......................................... 78

4.1 Relationship between small molecules based on molecular similarity, protein target modulation, and clinical effects. .................................................................................................................. 85
4.2 Modeling patient package inserts in the vector space model. ...................................................................... 91
4.3 Patient package insert similarity between a pair of drugs. ........................................................................... 95
4.4 Patient package insert similarity log-odds calculation example. ................................................................ 97
4.5 Relationship between empirical and estimated p-values and log-odds scores. ............................................... 99
4.6 Relationship between me-too drugs and clinical effects based on patient package insert similarity. .......................... 103
4.7 Examples of the relationship between me-too drugs and clinical effects. ...................................................... 104
4.8 Proportion of drug target annotations correctly predicted for primary and secondary targets using individual similarity methods. . 107
4.9 Proportion of secondary drug target annotations correctly predicted by combining methods. .......................... 109
4.10 3D similarity alone can predict important off-target effects. . . . . 113
4.11 Combining similarity methods can make off-target predictions stronger. 114
4.12 Proportion of ChEMBL drug target annotations correctly predicted. 116
4.13 ChEMBL example: combining similarity effectively predicts a drug target interaction. ................................. 118
4.14 Novelty of ChEMBL drug target annotations. .................... 120
4.15 Predicted interaction between levetiracetam and voltage-gated potassium channels. ........................................ 125
4.16 Predicted interaction between selegiline and the 5-HT and norepinephrine reuptake transporters. ....................... 127
CHAPTER 1

Introduction

Much of drug discovery requires the design of small molecules that modulate the activity of specific biological targets (typically protein receptors or enzymes) with minimal effects on other targets. These small molecules, despite best efforts, may have interactions with undesired targets and may cause side effects of a magnitude that exceeds their therapeutic benefits. Such interactions are best discovered prior to human clinical trials. Historically, many undesirable interactions have not been discovered early in the drug development process, and in some cases, have lead to serious adverse outcomes. In the United States, it is estimated that 2 million patients suffer from adverse drug reactions that result in 200,000 deaths annually [25]. Additionally, unforeseen side effects are often responsible for the failure of small molecules in clinical trials, contributing 30% to overall failures [50].

A noteworthy example of an approved drug that caused serious adverse effects is terfenadine, an antihistamine indicated to treat allergic conditions. The therapeutic benefit of terfenadine is obtained by its metabolite, fexofenadine, which antagonizes the histamine H1 receptor. Terfenadine is mostly metabolized by CYP3A4 to the active metabolite fexofenadine, but a small concentration of the pro-drug is still present in the blood [88]. Inhibition of CYP3A4 (which commonly occurs by ingestion of certain drugs and foods) leads to a higher concentration of terfenadine in the blood, which resulted in severe and sometimes fatal cardiac
arrhythmia. Terfenadine turned out to be an antagonist of the cardiac potassium channel hERG [88]. In 1997 the FDA recommended that therapeutics containing terfenadine be removed from the market [16]. Five months after terfenadine was withdrawn, fexofenadine was approved to supersede terfenadine because it did not interact with hERG. Early detection of the unexpected hERG interaction would have been extremely beneficial.

One can identify potential off-target effects through a combination of in vitro and in vivo testing [10]. However, comprehensive testing of this kind is both time consuming and expensive. Alternatively, one can make use of computational methods to identify a small number of potential problem targets for a molecule of interest. Very early in the lead optimization process, one might reject a molecule or scaffold based purely on a computational prediction. Later in the process, closer to the identification of a clinical candidate, follow-up assays would need to be done.

The goal of computational methods for pharmacological profiling is to identify the smallest number of potential off-targets while maintaining a high detection rate of candidates. In principle, such methods may be protein structure based. Given a small molecule of interest and a set of experimentally determined protein structures with specified ligand binding sites, will the molecule bind the site or not. Molecular docking methods can then be used that fit the molecule into the active site and if the appropriate interactions are made will be predicted to interact with the protein target in question. Ligand-based approaches that consider molecular similarity offer a means to profile small molecules without the need for protein structures. The idea is to make inferences about the potential activities of a particular molecule based on the known biological activities of other molecules that have been well-studied. The Jain lab established the feasibility of
using 3D ligand similarity for making such inferences [13] and the Shoichet lab
has identified unexpected targets based on 2D topological approaches [48].

This work establishes a framework for making inferences of biological activity
based on either 3D molecular similarity, 2D similarity, and clinical effects similarity
and examine the degree to which the three approaches can accurately identify
off-target effects. Further, the relationship between novelty in ligand structure
(i.e. different 2D topology compared with ligands for a target of interest) and
pharmacological novelty was explored. The framework is capable of incorporating
any computation involving a data source that relates chemicals to one-another,
relates targets to one-another, or relates chemicals to targets directly.
CHAPTER 2

Statistical Framework for Target Prediction

2.1 Abstract

Drug structures may be quantitatively compared based on 2D topological structural considerations and based on 3D characteristics directly related to binding. A framework for combining multiple similarity computations is presented along with its systematic application to 358 drugs with overlapping pharmacology. Given a new molecule along with a set of molecules sharing some biological effect, a single score based on comparison to the known set is produced, reflecting either 2D similarity, 3D similarity, or their combination. For prediction of primary targets, the benefit of 3D over 2D was relatively small, but for prediction of off-targets, the added benefit was large. In addition to assessing prediction, the relationship between chemical similarity and pharmacological novelty was studied. Drug pairs that shared high 3D similarity but low 2D similarity (i.e., a novel scaffold) were shown to be much more likely to exhibit pharmacologically relevant differences in terms of specific protein target modulation.

2.2 Introduction

We have previously examined the relationship between drug pharmacology and structural similarity in the context of demonstrating that the human design pro-
cess has a strong 2D reasoning bias [14]. Using a deeply annotated database of drug structures linked with their primary (desired) targets and secondary ones (“off-targets” generally responsible for side-effects) [13], we identified drug pairs that shared primary targets (primary target pairs) and those where the primary target of one drug was a secondary target of another (side-effect pairs). Among side-effect pairs, 2D similarity was extremely low when compared with primary target pairs. That is, when making an intentional design (developing a new drug for a particular indication where other drugs exist), we observed much higher 2D structural bias than when a pharmacological effect was unintentional. Apart from quantifying the 2D bias in human design, the study provided ample support for the proposition that molecules that appear to share little structural similarity by eye often share pharmacologically important effects [14].

The economic incentives underlying the discovery process appeared to be a key driver of incremental design strategies. Among drug pairs patented close together in time, 2D structural similarity was much higher than for drug pairs patented at distant times. In cases where on-patent therapeutics exist for an indication, introduction of a patentable close analogue can still be profitable. However, novelty in pharmacological action is clearly more important when competing against cheaply priced off-patent drugs. We speculated that structural novelty, measured by lower 2D similarity, leads to increasing novelty of pharmacologic action in the whole human organism. Figure 2.1 shows a typical example that illustrates these points. The 2D structures show imipramine (the first serotonin reuptake inhibitor), amitriptyline (a fast follow-on drug), and citalopram (a much more selective serotonin reuptake inhibitor). The 3D overlay shows that, while citalopram exhibits significant structural novelty at the 2D level, consideration of its similarity to imipramine in 3D shows high congruence.
Figure 2.1: Shown are three 5HT reuptake transporter ligands: imipramine (the first in its class), amitriptyline (a fast follow-on compound), and citalopram (a later generation SSRI). (A) Minor structural differences between imipramine and amitriptyline (highlighted in red). (B) Surflex-Sim’s 3D similarity overlay of citalopram (green carbons) and imipramine (atom colors). The significant regions of similarity within the molecule pair are illustrated with sticks, green (hydrophobic), blue and red (polar).
Our previous work considered what was true about the similarities of drug pairs given that one knew about the pharmacology of both of the drugs for the pairs in question. The present study asks the converse question. What is true about the molecular pharmacology of a new molecule given its similarity to a molecule or sets of molecules with known pharmacology? The question takes two forms. One is formulated as the task of prediction of primary and secondary targets of an as yet uncharacterized molecule. This is an important operational issue: identifying potential off-targets early in drug discovery. The other question asks how much novelty in pharmacological action is expected to arise from structural novelty in a new drug. This question revolves around me-too drugs. It is primarily a strategic issue for pharmaceutical development and a public policy issue for regulatory bodies. Both the prediction and novelty questions hinge on differences between 2D and 3D molecular similarity approaches because their underlying biases are different.

The present study establishes a framework in which 2D and 3D similarity computations can be directly compared and also combined. Given this framework, we studied the similarity patterns exhibited by 358 marketed small molecule drugs linked through partially shared molecular pharmacology and addressed two broad questions. First, we quantified the degree to which primary and secondary targets could be predicted using 2D similarity, 3D similarity, or a combination of both by making use of sets of drugs whose targets were known. Second, we quantified the likelihood, based purely on molecular similarity, that drug pairs would exhibit different levels of overlap in terms of their detailed molecular pharmacology. The specific methods used were Surflex-Sim, and the 2D GSIM computation implemented within the Surflex platform [14].

With respect to the first question, the results were expected, but striking as
to degree. The performance of the methods for predicting target annotations was $2D < 3D < 2D + 3D$. For primary target prediction, at a conservative threshold at which to assign annotations, true positive rates were 37%, 46%, and 59%, respectively. Consistent with our prior observations, 3D similarity did not yield a dramatic gain over 2D for primary targets due to the historical design bias problem: molecules designed to hit target X are often made specifically to look (in 2D) very much like molecules that are already marketed to modulate X. However, for off-target predictions, 2D yielded a 37% true positive rate, but 3D yielded 61%, with the combination yielding 68%. So, for off-targets, we observed a dramatic improvement over 2D in the ability of 3D molecular similarity to identify relevant pharmacological effects.

With respect to the second question, the primary broad observation is that a drug that shares high 2D and 3D structural similarity with another drug is likely to have indistinguishable pharmacological effects at the level of biochemically characterized modulation of protein targets. If, on the other hand, a drug shares little 2D similarity to existing drugs for the same cognate target but has high 3D similarity, there is greater likelihood to obtain a novel pharmacological effect. Specifically, drug pairs with high 3D and high 2D similarity showed identical biochemical targets four times more frequently than did pairs with high 3D similarity but low 2D similarity. This second case is particularly important from a design perspective because it represents a tractable one: where computational approaches making use of knowledge of existing therapeutics can guide design by mimicking 3D surface shape and electrostatics. It is exemplified in Figure 2.1 with imipramine and citalopram, the latter bringing both structural novelty and greater target specificity. Shared targets occurred very rarely when no structural similarity existed between pairs of molecules: 94% of the time there was no overlap at all in biological targets, with overlap in primary targets occurring just 2%
of the time.

The work reported here introduces a new methodological approach for data fusion, demonstrated with 2D and 3D molecular similarity. Given other recent reports of methods for data fusion and off-target prediction [47, 48, 64, 84], the differentiating features of what can be concluded based upon 2D and 3D molecular similarity is important to understand. This analysis of roughly one-third of small molecule drugs with extensively overlapping pharmacology has implications for both practical and strategic aspects of molecular design for therapeutic intervention.

2.3 Methods and Data

The following describes the molecular data sets, computational methods, and specific computational procedures. Data and protocols are available for download (see http://www.jainlab.org for details).

2.3.1 Molecular Data Sets

Beginning with a set of nearly 1000 drugs for which we have curated primary and secondary target information, we identified a set of drugs with overlapping pharmacology. This was done by considering the drugs as nodes of a graph. Edges existed between pairs of nodes when the drugs representing each node shared either a primary or secondary target. The connected subgraph that contains the molecules shown in Figure 2.1 consisted of 358 small molecule drugs (approximately one-third of those marketed in North America), which formed a connected web through their established biological targets.

The database of annotations of primary and secondary drug targets has been
described previously [14, 13], and only the specific salient aspects will be described here. In particular, all target annotations for each drug have been established well enough to identify a particular binding site on a particular protein assembly. So, benzodiazepines and barbiturates, both of which target the GABA\textsubscript{A} receptor, are distinguished because they bind different sites. This is a key distinction between our database and resources such as DrugBank [49] or commercial databases such as MDDR. DrugBank records annotations at the level of formal names (as we do, using HUGO [21] gene nomenclature conventions), but the distinction between different sites within an assembly of proteins is not made. Databases such as MDDR are even less specific in many cases, with biological activity defined by broad activity classes. This distinction is crucial because inferences may be drawn based on the comparison of a molecule to a set of drugs that share an annotation. If the set binds a single receptor assembly but contains two groups that bind different sites, inferential power of a similarity-based scheme will be diminished because one expects a new molecule might be similar to one set but not both. Machine-learning approaches can rely upon such data because they can form disjunctive models relating different chemical features to the same activity class, as was done by Nidhi et al. [67] with multicategory Bayesian models. Others have used data in this form to analyze aspects of the network topology of drugs in the context of biological pathways, which is also appropriate because the focus is on the relationship of target modulation to biological effect but does not depend upon site of action [92].

The second important difference between our data set and other resources is that a distinction is made between \textit{primary} targets, thought to modulate the therapeutically beneficial effects of each drug, and \textit{secondary} targets, which mediate pharmacologically relevant (but undesirable) effects. Drugs may have multiple primary and multiple secondary targets. Annotation of this distinction varies
among public resources, with DrugBank explicitly indicating whether a target is responsible for desirable pharmacological action, but with resources such as BindingDB [11] not making such a distinction (instead focusing on careful curation of biochemical assay results). One important aspect of our database is that a missing annotation between a drug and a target cannot be interpreted to mean that the drug does not hit the target. It is simply a lamentable fact that comprehensive biochemical profiling of marketed drugs has not been carried out and made public.

While detailed information regarding a drug’s effects, side-effects, dosage, and administration is available, systematized data systems that accurately capture this natural-language information in a form suitable for query and analysis are only beginning to become available. The SIDER [51] resource, for example, links drugs to a formal library of side-effect terms, but automated extraction of information from free-form documents is extremely challenging and can limit the utility of such resources for quantitative performance assessments. In our curation effort, we make use of resources such as DrugBank and SIDER, but annotations within our data set are always linked to primary literature, texts that themselves cite primary literature, or drug package-insert information. Annotations of primary and secondary targets or of specific side-effects are only made after direct examination of the evidence by an expert.

The set of 358 drugs identified through the primary and secondary target graph analysis was especially rich in targeting aminergic GPCRs (194 drugs) and ion channels and transporters (90) because drugs within these classes have long been known to have promiscuous and overlapping activity profiles [54, 79, 12]. The remaining drugs were roughly evenly split among pathogenic and human enzyme inhibitors. A total of 67 different biochemical targets were represented
among the 358 drugs. Of the 67 targets, 44 were annotated to be modulated by 5 or more drugs.

2.3.2 Computational Methods

The core computational methods for computation of molecular similarity have been published and will be briefly summarized here.

2.3.2.1 3D Similarity Computation with Surflex-Sim

The Surflex-Sim 3D molecular similarity method and its use for virtual screening has been described at length in multiple publications [14, 13, 41, 42, 66]. Briefly, the method uses a molecular similarity function that computes, given two molecules in specific poses, a value from 0 to 1 that reflects the degree to which their molecular surfaces are congruent with respect to both shape and polarity. The function is computed based on the differences in distances from observer points surrounding the molecules to the closest points on their surfaces, including both the closest hydrophobic surface points and the closest polar surface points. So, two molecules that may have very different underlying scaffolds may exhibit nearly identical surfaces to the observer points, which are intended to be analogous to a protein binding pocket, which also "observes" ligands from the outside.

A command has been added to Surflex-Sim for the specific computation of "agnostic" similarities between two molecules, called "simcover". For each molecule pair A and B, each is sampled to yield a diverse set of conformers (default 20). Molecule B optimized in terms of both conformation and alignment to each of the conformations of A, and scores are recorded for each such optimization. The same is done for A matched to the conformations of B. Both the average bidi-
rectional similarity and maximum are reported for each pair. For this work, the maximum similarity was used, but little difference is apparent making use of the average instead. Scores range from 0.0 to 10.0. Figure 2.1 illustrates the optimal mutual superimposition of imipramine and citalopram, which exhibits excellent concordance in terms of overall surface shape and correspondence of charge and directionality of the amines. The corresponding similarity score was 8.2.

2.3.2.2 GSIM-2D

The GSIM method was implemented as a strawman approach to identify ligand-retrieval problems that are relatively unchallenging [14]. However, due to the presence of structural analogues within many activity classes, it is frequently effective. Given two molecules A and B as input, it identifies all subgraphs of molecule A of depth 1, 2, and 3 at each heavy atom. For each subgraph containing 3 or more heavy atoms, existence of the subgraph is checked in molecule B. If it exists in molecule B, the score is incremented based on the number of subgraph atoms and whether the root atom is a carbon or not (non-carbon rooted subgraphs are weighted 5.0, else 1.0). We repeat the procedure for molecule B looking for its subgraphs within molecule A. The two scores are normalized to the interval [0,1] based on the maximum possible score in each direction. The minimum of the ratio of number of heavy atoms in molecule A to molecule B and vice versa is computed, and the overall similarity is the minimum of the two scores multiplied by the minimum heavy atom ratio. The overall effect is that to yield high similarity, molecules A and B must be roughly the same size and have contain subgraphs, especially those rooted at heteroatoms.
2.3.2.3 Statistical Framework for Unifying 2D and 3D Comparisons

Both of the similarity approaches just described yield scores in arbitrary units that have no fundamental physical meaning except at the level of perfect identity. To compare the approaches directly, or to combine the computed similarities for a single molecule against multiple molecules, some normalization is required. Figure 2.2 shows the inverse cumulative histogram of similarity scores for imipramine and citalopram computed against a random set of 1000 screening molecules from the ZINC database, the same set we have used previously as a decoy set in assessing docking and similarity virtual screening performance [14, 69, 70]. Both range from approximately 4.0 to 8.0 overall. The distributions are nearly perfectly normal, with respective means of 6.6 and 6.1 and respective standard deviations of 0.58 and 0.43.

Given that the background molecule set was chosen randomly, the chances that any particular molecule within the 1000 is related to a given ligand are very small. Consequently, we can make use of these distributions to estimate $p$-values for particular levels of similarity for a given compound. In the comparison shown in Figure 2.2, the more pessimistic $p$-value based on the distributions came from citalopram because it had a larger number of similarities to random molecules that met or exceeded 8.2. The corresponding $p$-value was 0.008. More generally, given any molecular similarity method that yields some score $S$ when comparing molecules A and B, a $p$-value can be computed by assessing the proportion of random molecules that have equal or greater similarity to $S$ than to each of A and B and then taking the larger of the two proportions.

Figure 2.3 shows the relationship between raw similarity scores and $p$-values for both the 2D and 3D computations. Note that even within narrow bands of numerical similarity, both for 2D and 3D, very different $p$-values obtain. Be-
Figure 2.2: To normalize the similarity score $s$ for a pair of molecules, $x$ and $y$, we compare the magnitude of $s$ to the empirically computed distributions of scores for $x$ and $y$ against a background molecule set. Here, the molecules are imipramine and citalopram, with similarity score 8.2, and similarity distributions are shown in red and green, respectively. The blue line marks the similarity score between the two. For both molecules, all background scores are greater than 4.0, but scores significantly greater than 6.0 become rare quickly. Imipramine had the higher proportion of background similarities $>8.2$, resulting in a $p$-value of 0.008.
tween 3D similarity of 7.4 and 7.6, we observed $p$-values between 0.00 and 0.45, corresponding to highly improbable and clearly significant all the way to clearly random. The relationship of molecular structure to changes in distributional character will be discussed later, but the salient feature from a computational perspective is that while there is a global relationship between a similarity score and $p$-value irrespective of the molecule under consideration, there is enough variability to warrant normalization on a per-molecule basis. Because the specific $p$-values will depend on the exact composition of the decoy set used, we have chosen to bin the values, which also leads to a simple means to combine multiple $p$-values into a single value for the purpose of data fusion.

Given some new molecule and a collection of molecules known to share some activity, we wish to be able to combine similarities to the set of knowns, potentially using multiple similarity methods, into a single value that reflects the combined information. We make use of the multinomial distribution for this purpose, as shown in Figure 2.4. The basic idea is simple. Given some set of $k$ different outcomes, each with associated prior probability $p_i$, and counts of each outcome $x_i$, $M$ gives the probability of observing the set of outcomes that gave rise to the counts observed. For application to the molecular data fusion problem, because we have converted each similarity score into a probability, we simply count the number of occurrences within each $p$-value bin, and the prior probability of each is simply the bin size itself. We make the computation using the computed probabilities for a set of similarities and using the converse probabilities, which yields symmetrical treatment for high similarity and for high dissimilarity. The final step in computing the score $S$ is taking the log of the ratio of the two probabilities $M$ and $M^*$ and inverting the sign. So, in a case where the similarities of a particular molecule to a collection that share some annotation are low, $S$ will be high. In a case where they are evenly spread from
Figure 2.3: The top plots show the relationship between 2D and 3D similarity scores and corresponding $p$-values for all 63903 drug pairs. Higher 2D and 3D similarity scores generally lead to low $p$-values (i.e., are more significant). However, even within narrow ranges of raw similarity scores, the $p$-values varied substantially. The bottom plots show histograms of the $p$-values for 2D similarity scores between 0.09 and 0.13, and 3D similarity scores between 7.4 and 7.6. Numerically close similarity scores can give rise to very different $p$-values depending on the molecules in question. The labels $x_{1-6}$ identify different qualitative bins of significance at $p$-value boundaries of 0.01, 0.05, 0.1, 0.2, and 0.5 (see Figure 2.4 for their use in the multinomial distribution computation).
Figure 2.4: We compute a single score $S$ to reflect the likelihood that a test molecule shares an annotation with a set of molecules known to have a particular activity. The similarity $p$-values are determined for the test molecule the set of known ligands. We compute the likelihood that the observed set of similarity $p$-values is extreme using the multinomial distribution. The $p$-values for the set of similarity comparisons are binned (top), and the bin counts are computed ($x_{1−6}$). $M$ is the likelihood of having observed such a set of $p$-values, and $M^*$ is the same computation using the converse probabilities. The Log Odds score $S$ combines the two. Positive $S$ indicates that it is more likely that the molecule in question shares the annotation with the ligand set than that it does not.

low to high, $S$ will be zero. And if similarity values skew low, $S$ will be negative.

The interpretation of a log-odds score of 2.0 is that it is 100 times more likely that the molecule in question shares high similarity with the annotated set of molecules than that it does not. To the extent that the similarity approaches that underpin the probability computations are related to biological effects, the log-odds score $S$ may be used as a predictor of such effects (see Figure 2.5 for an example).
Figure 2.5: Log Odds calculation example. To test if a new molecule, in this case azatadine, is a ligand of a target of interest such as the histamine H\textsubscript{1} receptor, we compute 2D and 3D similarity \(p\)-values for azatadine to known ligands of the target. Shown above are the 2D structures, 3D similarity overlays, and corresponding \(p\)-values of azatadine with three ligands. To calculate the Log Odds, similarity \(p\)-values were determined for the complete list of 30 histamine receptor ligands shown on the right. Both the 2D and 3D Log Odds scores are high and indicate a much higher probability that azatadine binds the histamine H\textsubscript{1} receptor rather than it does not.
2.3.3 Computational Procedures

Detailed scripts for generating the results presented here are available in the data archive associated with this paper. Briefly, all 3D molecular similarity computations between drugs were made using default parameters for Surflex-Sim: sf-sim simcover drug-list drug-list log-file-3d. Similarly, all 2D molecular similarity computations between drugs were made as follows: sf-sim gsimcover drug-list drug-list log-file-2d. To compute the background distributions to normalize the similarity values, the analogous computations were done using the 1000 molecule ZINC decoy set: sf-sim [g]simcover druglist zinc-list log-file-[3d][2d]-norm.

2.4 Results and Discussion

The primary results of the study fall into two basic categories. The first relate to what can be predicted about a compound’s potential biological effects based purely on molecular similarity and whether the use of 2D or 3D similarity methods influences the types of inferences that can be made. The second involve a census of the pharmacological congruence between pairs of drugs, where the pairs have been defined based upon the characteristics of their 2D and 3D similarities. The principal observation from the first category is that both 2D and 3D similarity (and their combination) are able to predict biological targets, but 3D similarity is more likely to identify effects that are not obvious from knowledge of pre-existing molecular pharmacology. The principal observation from the second category is that molecules sharing high similarity in both a 2D and 3D sense are much more likely to exhibit highly similar target profiles than those molecules that exhibit topological variation but retain high 3D similarity.
2.4.1 Predicting On- and Off-Target Effects

For each of the 358 drugs (see Methods and Data), we asked what their computed log-odds score was for each of the 44 targets that had 5 or more annotated drugs as either primary or secondary modulators to serve as positive examples. Figure 2.6 shows the results of the computation for both primary targets (top) and secondary targets (bottom). Nearly all computed log-odds scores were positive (about 90% for all methods), indicating greater similarity than dissimilarity to example sets of compounds for either 2D, 3D, or 2D + 3D log-odds computations. This was the desired result because the definitions of targets differentiated between different binding sites on the same protein assemblies, so ligands within a given set of modulators were known to bind competitively.

For primary target predictions, performance of the methods was 2D < 3D < 2D + 3D, but the degree of improvement in moving beyond 2D ranged from 9 percentage points at a logodds threshold of 6.0–15 percentage points at a log-odds threshold of 20.0. The added value of combining the two similarity approaches yielded typical gains of 10 percentage points over a broad range of log-odds values. At a threshold of 6.0, the combination of 2D + 3D similarity methods was able to identify a majority (59%) of all primary target annotations. As mentioned earlier, and as we have previously reported, the relatively limited gains of 3D over 2D are explained directly by human design bias [14]. The new observation here is that the effect holds in the forward predictive direction: when one has a set of ligands with known activity, 2D similarity works quite well in assigning primary targets to new molecules. For secondary target prediction, the same qualitative performance was observed, but the performance gains for 3D over 2D were 24 percentage points (log-odds threshold of 6.0) to 30 points or greater for log-odds thresholds of 10.0 or more. The combination of methods yielded only a
Figure 2.6: Proportion of drug targets correctly predicted. These plots indicate the proportion of drug targets correctly predicted by the three similarity methods at various Log Odds thresholds. Using 2D (red line), 3D (green line), or a combination of the similarity methods (blue line), we determined the Log Odds for both primary and secondary targets of the 358 drug set against 67 targets. At a Log Odds threshold of 6.0 (black vertical line), prediction success rates for primary targets for 2D, 3D, and combination were 37%, 46%, and 59%, respectively. For secondary targets, success rates were 37%, 61%, and 68%. Examples will be shown of predictions highlighted by the gray circles in the bottom plot.
marginal gain, as with primary targets, of typically 10 percentage points or less, identifying 68% of the secondary targets at a log-odds threshold of 6.0.

The highlighted circles from Figure 2.6 (bottom) provide examples of specific secondary target predictions shown in Figures 2.7, 2.8, and 2.9. Figure 2.7 shows the case of promethazine, whose primary target is the H1 receptor, and whose off-targets include multiple dopamine receptor subtypes. The drugs promazine and trifluoperazine are examples of the degree of structural concordance that can occur, allowing for predictions of targets by essentially any method for computing molecular similarity. For these two molecules compared to promethazine, both 2D and 3D approaches produced p-values less than 0.01 (the most extreme bin from Figure 2.4), and combined with p-values from 33 other dopamine D2 receptor drug comparisons, yielded log-odds scores of 10, 15, and 22 for 2D, 3D, and 2D + 3D, respectively. These drugs were all synthesized as part of the medicinal investigation of what were then termed “anti-histaminic phenothiazines”, many of which had antipsychotic properties [80]. These properties were due to a host of effects on different brain receptors but are thought to primarily derive from modulation of dopamine receptors of multiple subtypes. Relatively subtle changes in structure (e.g., from promazine to promethazine) yield sufficient different in target potencies to shift primary indication from antipsychotic for promazine to antihistamine for promethazine. However, the shifts in potency are not so dramatic as to abrogate the multiple target effects entirely.

Figure 2.8 shows another example of a phenothiazine antipsychotic whose primary effects derive from dopamine receptor modulation. Here, however, some of the more significant side-effects are those modulated by muscarinic antagonism, including dry-mouth and blurred vision. In this case, the 2D log-odds was just 3. Whereas 2D similarity did not produce a low p-value when comparing thioridazine
Figure 2.7: A 2D similarity method can sometimes correctly predict a target. Shown above are the 2D structures and 3D similarity overlays for the drug promethazine (a histamine $H_1$ receptor antagonist) compared to two dopamine receptor antagonists, promazine and trifluoperazine. Although promethazine is prescribed as an antihistamine (Phenergan), it is a branched derivative of the antipsychotic phenothiazines and is known to have about $1/10^*$ the dopamine receptor antagonistic activity. Our 2D method alone was sufficient to identify the dopamine receptors as an off-target of promethazine. As shown above, the Log Odds predictions that promethazine is a dopamine $D_2$ receptor ligand were $2D = 10$, $3D = 15$, and the combination of $2D$ and $3D = 22$. Note that at a log-odds score of 10, just 20% of secondary targets are identified by 2D alone, with 50% and 60% identified by 3D and the combination of 2D + 3D, respectively.
to either oxybutynin or diphenidol (two potent antimuscarinics), 3D similarity yielded much lower $p$-values. Coupled with those derived from comparisons to 61 other $M_1$ drugs, the 3D log-odds score was 39, allowing very confident assignment of muscarinic targeting to thioridazine. In this case, the addition of 2D similarity to 3D produced a slight reduction in computed log-odds score. Log-odds scores are not additive; additional observations affect the combinatorics such that a collection of $p$-values which alone yield a marginally positive log-odds score may diminish the score derived from a collection of $p$-values that produced a high score.

Figure 2.9 shows a case where the combination of 2D and 3D similarity produced a log-odds score greater than 6.0 where neither alone met that threshold. Nefazodone yields its primary effects through modulation of multiple reuptake transporters, but it has a significant side-effect of postural hypotension deriving from modulation of $\alpha$-adrenergic receptors. In this case, for $\alpha_{1A}$ receptor, 2D alone yielded log-odds of 2.4, with 3D yielding 5.5. Comparisons to dapiprazole and doxazosin produced no extreme $p$-values using either 2D or 3D, but all four scores leaned in favor of similarity to nefazodone. Along with 50 other drug comparisons, the combined log-odds for adrenergic effects was 6.5.

2.4.2 Excess Targets: False Positive Predictions

The framework we have developed allows for the combination of multiple sources of information to yield a single scalar value associated with a class prediction. In such a situation, it is both customary and desirable to make an estimate not only of true positive success rates but also of the corresponding false positive rates (e.g., with a receiver-operator characteristic (ROC) analysis). Here, we were able to identify primary and secondary targets about 60–70% of the time at
Figure 2.8: Our 3D similarity method more accurately predicts off-targets. Shown above are the 2D structures and 3D similarity overlays for the drug thioridazine (a dopamine receptor antagonist) compared to two muscarinic receptor antagonists, oxybutinin and diphenidol. Although thioridazine is prescribed as an antipsychotic (Mellaril), it is well-known to have antimuscarinic effects such as dry-mouth and blurred vision. Our 2D method did not identify the muscarinic receptors as an off-target of thioridazine (Log Odds for $M_1 = 3$). In contrast, our 3D method strongly predicted the off-target, with an $M_1$ Log Odds score approaching 40. Whereas over 10% of secondary target annotations are captured at this level by 3D similarity, none are captured with such a high level of confidence by 2D.
Figure 2.9: A combination of 2D and 3D similarity methods makes a small improvement over 3D alone. Shown above are the 2D structures and 3D similarity overlays for the drug nefazodone (a 5HT reuptake transporter inhibitor) compared to two $\alpha_1$ adrenergic receptor antagonists, dapiprazole and doxazosin. Although nefazodone is prescribed as an antidepressant (Serzone), it causes postural hypotension, a known side effect of $\alpha_1$ blockers. Neither the 2D or 3D methods alone identified the $\alpha_1$ adrenergic receptors as an off-target of nefazodone at the conservative threshold of 6.0 (Log Odds for $\alpha_1A = 2.4$ and 5.5, respectively). In contrast, a combination of the 2D and 3D methods predicted the off-target. Close to 70% of all off-targets were correctly identified at Log Odds = 6.0 using the combination approach.
a combination log-odds score threshold of 6.0. However, at that threshold, there are targets suggested for drugs for which no annotation is known. At a log-odds threshold yielding a true positive rate of 60%, the typical ratio of excess predicted targets relative to the total number of known primary and secondary targets was roughly 2–3, depending on the class of drugs involved. Larger numbers of excess targets were observed for drugs whose primary targets were among the aminergic GPCRs. The difficulty in interpreting this observation is that public data do not exist that systematically profile small molecule drugs in biochemical assays.

As a surrogate for biochemical data in unknown drug-target relationships, we manually assessed package insert and related information to make a determination of whether muscarinic side-effects were both present and drug-related. These included dry-mouth, urinary retention, blurred vision, drowsiness, mydriasis, and other effects. For the 358 drugs where we had no formal annotations of muscarinic target effects, which totaled 294 compounds, we surveyed a random subset of slightly more than half of them (180 drugs total), resulting in 84 with muscarinic side-effects and 96 without. We also surveyed 29 of the 64 drugs that we had previously annotated as binding muscarinic receptors. All 29 of the previously annotated muscarinic modulators showed clear, drug-related side-effects (90% exhibited dry mouth effects, 69% drowsiness, and a majority also showed urinary retention, blurred vision, and dizziness). For the 64 drugs with annotated muscarinic target effects, the mean log-odds score for muscarinic receptors was 25.8, with 92% scoring higher than 6.0.

Overall, using the side-effect assessments as a binary class label for the 180 surveyed drugs that had not been annotated as muscarinic modulators, the log-odds score produced an ROC area of 0.88 (95% confidence interval of 0.83–0.93). The enrichment for drugs with muscarinic side-effects among the top 1% log-odds
scores was 19-fold. Of the surveyed drugs, 90% of those with a log-odds score of 6.0 or greater showed classic muscarinic side-effects (38/42 surveyed drugs). Even at a threshold of just 2.0, 85% were positive (55/65). Above a threshold of 26.0, all surveyed drugs (16 total) showed such side-effects. Conversely, below a log-odds threshold of -6.0, just 6% (3/47 surveyed) showed potentially muscarinic effects. Below a threshold of -16.0, no drugs showed such effects (26 total).

Three examples of drugs that had lacked muscarinic annotations are particularly informative. Amoxapine, an antidepressant working primarily through the norepinephrine reuptake transporter, received a log-odds score of 4.8. Prescribing information indicated that the most frequent side-effects included dry mouth, constipation, and blurred vision. It has also been shown biochemically to bind muscarinic receptors [36, 15, 57]. Orphenadrine, an antihistamine prescribed to relieve muscular pain, received a score of 42.9. Prescribing information indicates that “dryness of the mouth is usually the first adverse effect to appear.” The drug has been shown to antagonize muscarinic receptors with a $K_i$ of 100 nM [83]. Mesoridazine received a score of 37.1, had clear muscarinic side-effects, and also has a $K_i$ of 69 nM against the $M_1$ receptor [36]. Notably, it received a log-odds score of 6.4 against the HERG potassium channel, although it had not been annotated for such activity. It was withdrawn from the U.S. market in 2004 due to HERG-mediated cardiac side-effects [81].

This survey of muscarinic side-effects among previously unannotated drugs makes three points. First, the empty cells of the annotation matrix of drug to target interactions cannot be thought of as indicating no effect. Second, the log-odds scores were both sensitive and specific with respect to muscarinic target annotation. Third, the lack of systematic profiling of drugs for which ample human data exist represents a large gap in our knowledge. Manual curation
of this depth requires on the order of 30 min to 1 h per drug per side-effect, after establishing the relationship between a particular target and the relevant human pharmacology down to specific terms and variations. We are exploring automated means to consider databases of side-effect terms and their relationship to predicted on- and off-targets in order to carry out a more comprehensive study.

Approaches for semiautomatic curation such as SIDER [51] are challenged by variations in language such as “dryness of the mouth” instead of “dry mouth.” The MedDRA dictionary [8], for example, lists the latter as a defined medical term (but not the former), and relatively sophisticated language parsing is required to relate the two together. In the case of orphenadrine, one of the 84 drugs with clear muscarinic side-effects, SIDER misses the dry mouth effect, which is clinically the most prominent. Even with much more extensive synonym mapping, cases exist where side-effects are listed as not being present or are listed as being present but then dispensed with as not different from placebo, which is challenging to assess without expert manual curation.

2.4.3 Relationship to Other Methods

Two relatively recent approaches to data fusion involving molecular similarity are particularly relevant to our log-odds scoring approach. Muchmore and Hadjuk’s belief theory approach [64] and the similarity ensemble approach (SEA) [47] introduced by Shoichet’s group both offer the means to make predictions about a given molecule’s activity based upon its relationships to other molecules.

The belief theory approach makes use of Hooper’s Rule, which was devised in the late 1600s by George Hooper, predating the Bayesian belief approaches later popularized by Laplace and his adherents [78]. The rule was devised to address the credibility of a report of some fact when simultaneously attested
by $N$ reporters, each with credibility $p$ (high $p$ implying high credibility). This rule formalizes the notion that multiple partially credible sources strengthen one another’s credibility. In the original report applying this rule to predictions of molecular activity based on similarity, the definitions of positive pairs of molecules and negative pairs differed from the current work, with molecule pairs considered as positive sharing not only a target but similar potency against the target. Similarity descriptors were converted into probability functions by considering a large set of positive and negative pairs and counting the number of times that a pair with some level of similarity was a positive example. Evidence from multiple similarity methods concerning pairs of molecules was combined using Hooper’s Rule. A key distinction with our log-odds approach is that the belief theory formalism always increases belief, no matter how marginal an additional source’s belief may be. In the log-odds approach, $N$ very low $p$-values coupled with $N$ symmetrically high $p$-values yield a log-odds of associating a target to a ligand of zero. The belief theory approach treats the cases of very low similarity as attesting in favor of the proposition that the query molecule will hit the target in question, but with low belief. One might argue that the interpretation of such a value is more akin to a reporter attesting against a fact rather than giving it marginal support, making the application of Hooper’s Rule a matter of empirical choice rather than purely logical.

The SEA method [47] uses a framework for estimating probabilities that is similar to that used for comparing sequence similarity of proteins, with likelihoods represented as $E$-values (a $p$-value multiplied by a large, arbitrary constant representing a database size). SEA makes use of 2D topological similarity to compute pairwise similarities between sets of molecules. By choosing a threshold below which to ignore similarity values, the pairwise sum of all similarities between two sets of unrelated molecules was shown to fit an extreme value distribution.
So, to compare one (or several) molecules against a set with known activity, the magnitude of the raw similarity set comparison score is compared with that expected from unrelated sets, a probability is derived, and an $E$-value is produced. In contrast with the belief theory approach, in this formulation, the presence of poor similarity values yields poorer $E$-values.

Both the belief theory and SEA approaches treat raw similarity values as being equivalent regardless of the specific molecules or molecule sets in question. Our observation of both the GSIM and Surflex-Sim methods, which we believe will also hold for other methods such as ROCS [73](3D) and Daylight fingerprint-based similarity [22](2D), is that the probability of observing some raw value varies depending on the particular structure involved. As seen in Figure 2.3, $p$-values associated with narrow similarity ranges included extremely significant values as well as clearly random ones. For similar molecules, the distributions of observed similarities to the background set tend to be close (see Figure 2.2 for an example). However, for a small and simple molecule, such as acetaminophen, the required similarity score to reach a $p$-value of 0.01 is higher (8.3) than that for more complex molecules, such as azithromycin, where the required similarity is lower (5.8). Clearly, the particular values depend on the composition of the background molecule set, but we do not believe it is possible to construct a nondegenerate background set against which all molecules will exhibit congruent similarity distributions. By assessing differences in likelihood of observing different similarity levels within the context of each specific molecule pair, it is likely that the associated log-odds scores better reflect the underlying similarity relationships than approaches that take a coarser-grained approach. Of course, it is also possible to make use of global similarity distributions with the log-odds approach, but it is difficult to justify doing so.
2.4.3.1 Quantitative Comparison to Other Approaches

The muscarinic side-effect prediction task offers the opportunity for direct comparison of our approach to belief theory and to SEA. We computed joint beliefs regarding muscarinic activity for the 180 drugs and used these beliefs to assess ROC area. Recall that the 180 drug set consisted of 86 positives and 94 negatives based on the presence of side-effects, with similarities for each computed against 64 known muscarinic modulating drugs. For belief theory, the formula for combining evidence is given by $B = 1 - (1 - B_1) \times (1 - B_2) \times \ldots \times (1 - B_N)$, where $B_1 \ldots B_N$ are the separate beliefs associated with the assertion that a given molecule has a particular activity. The most direct comparison to Muchmore and Hadjuk’s formulation is made by setting each $B_i = (1 - p_i)$, with each $p_i$ derived from the 3D similarity computations used above for the log-odds approach. Using a single global distribution to obtain $p$-values from the similarities, we observed an ROC area of $0.61 \pm 0.05$ (95% confidence interval), which was significantly worse than for the log-odds approach ($0.88 \pm 0.05$). Using empirically determined $p$-values for each molecular comparison (as the log-odds approach does), the performance improved to $0.72 \pm 0.05$ but was still significantly worse than the log-odds result.

Note, however, that the ROC area comparisons are somewhat misleading due to the degeneracy in the belief theory evidence rule. If a single belief is 1.0 ($p$-value of 0.0), the overall joint belief will be 1.0 no matter what the other belief values may be. For the muscarinic side-effect prediction task, this results in a large proportion of joint beliefs for the 180 drugs to be exactly 1.0. This degeneracy stems from the definition of Hooper’s Rule, but its effect can be ameliorated by scaling down all beliefs by a constant factor. The best result we were able to obtain for belief theory was an ROC area of $0.85 \pm 0.05$ (nominally indistinguishable from log-odds), using $B = 0.5(1 - p)$, with empirically determined $p$-values.
for each pairwise molecular comparison. Even with this augmentation, there were a significant number of tied values of high belief, covering nearly 10% of the 180 ligands. The maximal enrichment for belief theory, in this most favorable (and artificial) formulation, was 6.4, corresponding to a true-positive (TP) rate of 55% and false-positive (FP) rate of 9%. Much better early enrichment was possible with the log-odds approach because there is no multiplicative degeneracy involving strict interpretation of \( p \)-values. We obtained maximal enrichment of 20-fold at a false-positive rate of just 1% using 3D log-odds.

For the SEA approach, a direct performance comparison (with the same set of 64 annotated muscarinic ligands used here) was not possible using the web-based SEA interface (sea.bkslab.org). However, the annotations underpinning SEA predictions are far more extensive than those used here, with over 1000 ligands having muscarinic target activity (including exact matches for 50% of the 64 used here, and close analogues for over 85%). We queried the 180 drugs for SEA predictions, which were reported for target predictions with \( E \)-values < 10.0 (recall that such \( E \)-values are generally thought to be significant when less than \( 10^{-10.0} \)). For each drug, we recorded the most extreme \( E \)-values against any muscarinic subtype. Those molecules with no predicted muscarinic targets were assigned an \( E \)-value of 100.0. The corresponding ROC area was 0.57 ± 0.05, significantly worse than the log-odds approach. As with the belief theory approach, interpretation of ROC areas is problematic due to tied values. With SEA, the tied values were at the low end of the ranking because the majority of drugs received no muscarinic target predictions at all. Maximal enrichment for the SEA approach occurred within the nontied value range at an \( E \)-value cutoff of \( 10^{-1.2} \), allowing for a direct comparison. Maximal enrichment was 3.8-fold. This corresponded to an FP rate of 4% and a TP rate of 15%. Three direct comparisons between the log-odds approach and SEA are particularly meaningful:
the maximal enrichment, which was 20-fold for log-odds vs 4-fold for SEA,
corresponding TP rates at the same 4% FP rate, 48% vs 15%, respectively,
and corresponding FP rates at the same 15% TP rate, 0% vs 4%.

We believe that the inherent degeneracy in Hooper’s rule favoring high beliefs makes it inappropriate to use in a situation where belief values cannot be fully trusted. Given a single spurious annotation or a single similarity method yielding an inappropriately high confidence in a single molecular comparison, belief theory will produce incorrectly high belief in a prediction. In the case of SEA, we believe that the fundamental divergence of 2D similarity methods from the direct biophysical underpinnings of molecular activity limit the degree to which one can identify surprising off-target effects with high specificity.

2.4.4 Off-Target Prediction: Detection of Surprising Effects

The distinctions among different methods for data fusion, while clearly important, are not as critical as the distinctions among similarity methods that provide information to the data fusion computations themselves. Those similarity approaches whose scores are derived from directly relevant biophysical features (like surface shape and electrostatics) will yield different inferences than those that are less directly related to physical characteristics but which may be closely related to design ancestry. Two particularly telling examples of the distinction involve methadone and imipramine, compounds whose long history allows us to understand not only what the compounds do pharmacologically but also why they were synthesized and tested to begin with.

Figure 2.10 illustrates the historic context of the synthesis and testing of methadone and imipramine. Methadone was synthesized during WWII as part of an effort to develop anticholinergics for use as nerve gas antidotes [80] due
to the limited availability of the natural product atropine (nerve gas results in
an accumulation of acetylcholine by inhibition of acetylcholinesterase, leading
to spasm and death). On testing in animals, the surprising finding was that
methadone (and demerol as well) produced the Straub tail effect, indicative of
opioid analgesic activity. In a similar serendipitous story [33], the compound
G-22,355, which became known as imipramine, was selected for testing as an
antipsychotic. Roland Kuhn, a psychiatrist at the Cantonal Mental Hospital of
Münsterlingen, and Robert Domenjoz, a medicinal chemist at Geigy Pharmaceu-
ticals, identified it as being structurally similar to chlorpromazine. Kuhn tested
the compound with no success on psychotic patients, but prior to returning the
supply, it was tested on a small number of depressive patients. The effects were
sufficiently dramatic after just three patients to suggest the compound had unique
properties and warranted further testing. Imipramine established a new class of
drugs[52], which ultimately came to be understood as acting primarily through
the serotonin reuptake transporter.

Methadone’s surprising on-target activity could have been predicted by the
3D log-odds approach based on the structures of morphinan-based opioids such
as hydrocodone and codeine that had been identified well before methadone’s
synthesis. These had very low $p$-values using the Surflex-Sim approach. For
hydrocodone, codeine, morphine, and oxycodone, the 3D $p$-values were, respec-
tively, 0.007, 0.048, 0.057, and 0.060. The 2D GSIM $p$-values were, respectively,
0.35, 0.35, 0.35, and 0.63. Clearly, to predict the opioid activity, 3D structural
comparisons would be required. The case of imipramine cannot be considered in
this pseudoprospective fashion because its synthesis and testing led subsequently
to the identification of both its primary biological mechanism of action as well as
to the line of chemical inquiry that produced selective agents such as citalopram.
However, if we consider citalopram’s relationship to the eight serotonin-reuptake
Figure 2.10: The design intention and surprising effects of some older drugs are known. Methadone and demerol were synthesized in an effort to make synthetically scalable anticholinergics as nerve gas antidotes by the Nazis in WWII. Their opioid effects were discovered serendipitously in a live cat assay. Imipramine was synthesized as an analogue of other tricyclic antipsychotics. Its surprising antidepressive effects were discovered through direct human experimentation, later identified to result from inhibition of the serotonin reuptake transporter. Using 3D similarity, one can correctly associate the surprising target effects with the drugs, but using 2D similarity, one cannot.
inhibitors that predated it from our set of 358 (imipramine, clomipramine, trimipramine, amitriptyline, trazodone, paroxetine, fluvoxamine, and fluoxetine), we see that 3D similarity yielded \( p \)-values \( \leq 0.05 \) for all eight, but 2D similarity yielded \( p \)-values \( \leq 0.05 \) for only three.

Overall, for known secondary targets (most of which can be considered surprises to some degree), the 3D log-odds scores were, on average, 9.3 log units higher than the 2D scores. For known primary targets (where relatively fewer can be considered surprises), the difference was 4.0 in favor of 3D log-odds over 2D. Relationships that can be deduced through 3D molecular similarity include those that genuinely are surprising, not just those that would be obvious to someone knowledgeable of molecular pharmacology in a particular area.

### 2.4.5 Recent Off-Target Predictions

Given these anecdotes, there clearly can be differences between the types of inferences that can be drawn from 2D and 3D molecular similarity methods. The supporting information behind predictions such as these is important because the natural application of computational methods for predicting off-target effects is to identify those that someone intimately involved in a particular pharmacological area could not reasonably guess. What we have seen is that in cases where we are able to understand both the reasoning behind molecular design and the serendipitous discoveries about activity, it is the province of 2D methods to uncover effects related to historical reasoning that anticipated the effects but 3D methods to also find the surprises.

In 2006, we observed that methadone, based on 3D molecular similarity, cosegregated with muscarinic and histamine receptor antagonists, echoing its genesis more than 60 years earlier [13]. We did not show biochemically that methadone
was a muscarinic antagonist, but we pointed out that its side-effects included those associated with muscarinic antagonism: dry mouth, urinary retention, sweating, and reduced bowel motility. Subsequently, a biochemical assay showed that methadone has a $K_i$ of 1.0 $\mu$M for the $M_3$ receptor [47]. Using Surflex-Sim 3D similarity, methadone could be properly associated with the $\mu$ opioid receptor. What our study lacked was the perspective that 2D similarity provides as to what should have been considered obvious in this case: the basic reasoning behind synthesis of methadone was topological analogy to atropine and its analogues. Keiser et al. [47] directly showed that the SEA 2D similarity approach could reveal the off-target muscarinic effect of methadone (but not the on-target opioid effect). The 2D SEA approach successfully detected the association between methadone and the muscarinic receptor because attempts to create antimuscarinics from 2D analogy to atropine eventually succeeded, resulting in compounds such as adiphenine, diphenidol, tolterodine, oxybutynin, dicyclomine, and many others with a clear 2D similarity to methadone.

We observed this same pattern involving scaffold ancestry in a recently published application of the SEA approach [48]. In it, a set of predictions were correctly made for four drugs, where each of the predicted off-targets was unrelated by sequence or structure to the primary targets of the drugs. Figure 2.11 shows two of the drugs, primary canonical targets, predicted off-targets, and an example of a previously published [29, 37, 19, 90, 85, 93] high-affinity ligand of each off-target protein that shares a scaffold with each drug. In each case, the scaffold in question had been actively probed in medicinal chemistry exercises for the predicted off-target effect. The specificity of the highlighted scaffold for the off-target in question among CHEMBL39 annotations was over 40-fold for tetrabenazine, and the highlighted scaffold for the delavirdine prediction was over 1000-fold greater for $H_4$ compared with any other target. Two other sets
of predictions were made on drugs which target the NMDA receptor: ifenprodil and a simple analogue thereof. The predicted and verified activities included reuptake transporters (5HTT and NET), opioid receptors (µ and κ), and the D₄ receptor. These activities shared the same pattern as those in Figure 2.11 with respect to the presence of previously published high-affinity analogues against the predicted targets (data not shown). The more general point relates to experimental molecular pharmacology. In 1991, ifenprodil was investigated for activity in addition to the NMDA and adrenergic ones already known [12], and potent activity was reported for the σ and 5HT1α receptors. Established pharmacological crosstalk among ligands of σ receptors and the opioid µ, δ, and κ subtypes [54, 87] anticipated weak opioid activities for ifenprodil and its analogue. Crosstalk between ligands of the adrenergic and 5HT receptors and reuptake transporters [79] anticipated these activities as well. Complex specificity patterns across multiple reuptake transporters and multiple receptor subtypes of σ, NMDA, opioid, adrenergic, and serotonin have been probed for many years.

The presence of many published ligand/target relationships provides data for computational inferences that parallel pharmacological knowledge. For predictions to have high practical utility, they must identify off-target effects for drugs automatically, reliably, and with high specificity, and ideally they must identify effects that are truly surprising. Evaluating computational methods is challenging because even nominally prospective predictions can be driven by the evolutionary history of drugs. One can “predict” an activity for a ligand based on the fact that someone thought of the activity in connection with the ligand’s scaffold before, causing analogues to be developed and probed for that activity. In such cases, tools that ferret out such information will be useful only to the extent that they are either more effective than someone knowledgeable in molecular pharmacology or that they are facile to apply automatically and have a low rate of false
Figure 2.11: At left, two drugs are shown for which off-targets were identified through application of the SEA 2D similarity approach. Potencies for off-target effects were much weaker than for the on-target drug effects (shown below the drug names). The off-target effects were also much weaker than for primary modulators of the predicted targets. Molecular series containing common core substructures had been actively investigated as desirable scaffolds for the predicted targets.
predictions. Developers of predictive methods should disclose the reasons why a method made a particular prediction. Usually this requires only the provision of typical molecular structures that underpinned an inference. Special care must be taken in the case of methods for predicting off-target effects because the goal is to identify those effects that might otherwise derail a clinical candidate, and it is reasonable to believe that the more obvious potential effects would have been extensively investigated.

2.4.6 Relationship of Structural Novelty to Pharmacology

From the foregoing discussion and our previous work [14], it is clear that the drug design process shows a clear component of design relating directly to topological reasoning about the biological activity expected from a particular molecular structure. It is also clear that clinically relevant surprises occur both with respect to primary targets as well as secondary ones. To assess the degree to which chemical structural novelty was directly related to novelty in pharmacological effect, we computed the pairwise similarity of all 358 drugs and split them into four groups: pairs with high 2D and high 3D similarity, low 2D but high 3D, low 3D but high 2D, and low 2D and low 3D. Figure 2.12 shows the proportions of molecule pairs within each group that had identical annotated targets (blue bars), overlapping primary targets but including some differences as to overall target effects (orange), non-overlapping primary targets but some overlap among secondary targets (green), and completely non-overlapping targets (purple). It is important to understand that the annotation of target effects include only those where sufficient experimentation exists in order to localize an effect to a specific binding site on a particular protein assembly. So, as we saw above with the analysis of muscarinic side-effects, many unannotated drug-target relationships may
well exist.

In the case of high 3D and high 2D similarity (upper right), nearly 80% of drug pairs show some degree of target overlap, with nearly 40% having identical targets and nearly 70% sharing primary targets. With the same level of 3D similarity but with low 2D similarity (upper left), slightly less than half of the drugs share targets, and just 10% have identical targets. The converse case (high 2D, low 3D, bottom right) produces somewhat similar proportions but with 70% having no common targets. As expected, molecules sharing no molecular similarity shared no targets nearly 95% of the time. Figure 2.13 shows examples from each quadrant. The case of imipramine and its fast follow-on compound amitriptyline fell into the identical target set; indeed, they have very little to differentiate them in terms of pharmacology even beyond specific targets [34]. However, the structural creativity shown by citalopram relative to imipramine (high 3D, low 2D) produced much more specificity with respect to the serotonin reuptake transporter, and citalopram along with other SSRIs came to dominate antidepressant therapy. A typical case for low 3D but high 2D similarity is the pair bupropion and ketorolac, which share no targets. The overwhelmingly common case for low 2D and low 3D similarity is exemplified by albuterol and imipramine, again sharing no common targets. About 2% of the time, drugs with some overlapping targets share no similarity at all. The case of sildenafil and tadalafil are a particularly striking example, both binding PDE5 within the same volume, but exhibiting no molecular similarity, either by eye or through computational means [14]. Note, however, that while the annotated targets were identical for the pair, their detailed pharmacology is significantly different, particularly with respect to half-life.

These findings are not surprising in a qualitative sense. It should be the case
Figure 2.12: Drug pairs were segregated based on 2D and 3D similarity p-values into the 4 quadrants shown above (number of pairs per quadrant shown in parentheses). Conservative structural modifications are much more likely to yield highly similar pharmacology. Drug pairs with high2D/high3D similarity are about 4 times more likely to exhibit identical pharmacology relative to drug pairs with low2D/high3D similarity. Also, drug pairs with overlapping primary targets are 2 times more likely to have high2D/high3D similarity compared to those with low2D/high3D similarity. As expected, most drug pairs with high2D/low3D or low2D/low3D similarity have no target overlap.
Figure 2.13: Imipramine and amitriptyline differ by only 1 atom, share identical on- and off-targets, and were approved as antidepressants in 1959 and 1961, respectively. In contrast, citalopram has significantly lower 2D similarity to imipramine, has fewer off-targets, and thus a more favorable pharmacological profile. Citalopram was approved as an antidepressant in 1998. For bupropion and ketorolac, the high 2D similarity is apparent on inspection, leading to a very low 2D $p$-value, but the 3D similarity correctly suggests that the compounds are unrelated in effects. Albuterol and imipramine are aminergic GPCR ligands with no target overlap and very low similarity. Sildenafil and tadalafil are a rare drug pair within the low 2D/3D quadrant that share a primary target (PDE5).
that nearly identical molecules will more frequently share very similar biological effects than those that start to differ. We believe that the degree of deviation in effects is striking. There is a 4-fold difference in the a priori likelihood that two drugs will share identical pharmacological targets when shifting from high 2D and high 3D similarity to a case that shares only high 3D congruence. Consider the case of designing a new drug with knowledge of the structures of existing drugs within a therapeutic category. In the modern research environment, it is likely that one will be able to guarantee that the desired target be among those that will exhibit pharmacological effects, but one cannot know that these effects will be the dominating ones. From Figure 2.12, we will consider the molecule pairs that share some targets in this analysis. By designing a me-too analogue (high 2D and 3D similarity to existing drugs), one has about a 47% chance of showing the same target profile as the incumbent compound versus showing either a difference in secondary targets or overlapping targets with different primary effects. By designing a structurally novel compound (high 3D but low 2D), one has a 23% chance of showing the same target profile. In the me-too case, chances are even (53%:47%) in terms of seeing novelty at the level of target specificity, but in the case of a structurally novel drug scaffold, the chances are 3:1 in favor of novelty (77%:23%). Two things are worth noting: First, with modern 3D molecular similarity and 3D QSAR methods, design of such compounds is tractable. Second, the development risks associated with novelty are almost certainly higher because one cannot know a priori with full confidence what the precise biological effect differences might be, only that one is more likely to encounter them.
2.5 Conclusions

We have reported a new method for combining information from molecular similarity computations in order to effectively make inferences based on the known activity of sets of molecules. The approach is general and can be applied to any similarity method, offering a unified means to fuse the output from multiple sources to produce a single log-odds score. Two aspects are particularly important: (1) mapping of similarity values to \( p \)-values in a context-specific way because raw similarity values have different significance depending on the complexity of the molecules being compared, and (2) offering a means of data fusion that balances evidence in favor of an assertion with evidence against the assertion while avoiding degeneracies that arise from literal interpretation of empirically estimated \( p \)-values. By comprehensively applying the method to a large set of drugs with overlapping pharmacological effects, we were able to identify differences in the predictive ability of 2D similarity methods compared with 3D ones. In assessing the predictive value of the approach, the most comprehensive analyses were done considering recovery of known primary and secondary targets. Particularly for the prediction of unanticipated off-target effects, 3D performed much better than 2D, although the combination of both was beneficial. Comprehensive analysis of false positive predictions was impossible due to the lack of systematic profiling of drugs against large panels of targets. However, using the well-known muscarinic side effects as a surrogate for direct muscarinic modulation, we assessed the behavior of drugs lacking explicit annotation. Within this set, we established excellent separation of drugs with muscarinic side-effects from those with none apparent. Alternative approaches such as belief theory and SEA performed less well.

We also considered the relationship between chemical structural novelty and
pharmacological novelty. The key finding was that for a me-too drug pair (exhibiting both high 2D and 3D similarity), the chances were essentially even of observing modulation of identical sets of biological targets compared with non-identical sets. However, in the case where the drugs show high 3D similarity but show differences at the topological level, these odds shifted to roughly three-to-one in favor of observing novel effects at the gross level of target modulation. Clearly, even very subtle changes in chemical structure can yield sufficient differences in potency, selectivity, or ADME/toxicity characteristics to make for novelty in pharmacological action that can provide benefits to patient populations over existing therapies. This is especially true for therapeutics that target rapidly evolving pathogens, where minor structural modifications can overcome resistance. But it is clear that introduction of a novel scaffold brings significantly higher risk and potentially higher reward in terms of novelty that might be beneficial for patients.

While we believe that more detailed study is warranted, consideration should be given to regulatory changes that better balance the risk/reward equation for drug discovery. One possibility would be to always require a head-to-head clinical comparison against an approved close analogue in cases of me-too drug candidates seeking regulatory approval (excepting those that target pathogens). For drug candidates exhibiting novel structures, such requirements might still be imposed as they are currently on an ad hoc basis but would not be presumptively required. We believe that a disproportionate amount of research and clinical development effort is currently spent on drugs that have relatively little chance of providing significant new benefits to patients. Additional study will be required, however, in order to quantify this to the extent desirable for a regulatory modification.

An area for future research will involve systematic projection of phenotypic
side-effects shared by drugs onto potential biological targets. As shown by Scheiber et al. [77], integration of such projections with information about biological network structures can help to identify the specific molecular basis for clinically important adverse drug reactions. Use of more sophisticated methods for the underlying chemical to target inferences, as presented here, should serve to make such exercises more effective in identifying the causative factors.
CHAPTER 3

Computationally Efficient Similarity Score Calculation and Normalization

3.1 Introduction

In the previous chapter, log-odds scores resulted from a set of thorough similarity computations, where each underlying similarity comparison of one molecule onto a fixed conformation of another typically required several seconds. For a single log-odds score against a set of 20 ligands with a known biological activity, the primary similarity computation would take 1-2 hours. Further, similarity scores were empirically normalized to $p$-values by assessing their significance against a background set of similarity scores derived from a random set of 1000 screening molecules from the ZINC database [91]. Because the normalization is molecule specific, it required 1000 additional computations, further reducing the feasibility of the approach for large-scale molecular similarity computations. To address these issues, in order to make the approach broadly applicable for predictive modeling of potential activities involving many putative biological targets, we made use of a major speed improvement within Surflex-Sim (the underlying similarity computation engine), and we have developed an efficient $p$-value normalization scheme that requires just 20 similarity computations. The combined effect of these enhancements allows a computation of a log-odds score in a few minutes.
Figure 3.1: Example of 3D molecular similarity using Surflex-Sim. Carbamazepine (green carbon atoms) and levetiracetam (atom color) are two anticonvulsants that do not share any protein target overlap which is surprising given their significant 3D similarity. Green sticks correspond to regions of significant hydrophobic similarity and blue/red sticks correspond to regions of significant polar similarity.

rather than several hours.

3.2 3D Molecular Similarity with Surflex-Sim

The Surflex-Sim 3D molecular similarity method and its use for virtual screening and off-target prediction has been extensively described in multiple publications [13, 14, 41, 42, 66]. To provide additional context for the method, because it makes a fundamental contribution to the predictive pharmacology application here, the theory and algorithmic basis will be described here. Aspects of this description have been paraphrased and excerpted from a forthcoming book, used with permission [43]. Surflex-Sim addresses molecular shape comparisons from a surface-based perspective, considering both the overall molecular envelope as well as the electrostatic properties of the molecular surface. For application in the predictive pharmacology context, given a molecule pair A and B, the
conformations of molecule A are aligned to sampled conformations of molecule B, and the same procedure is repeated in the opposite direction while maintaining a list of the similarity scores that are computed. Both the average bidirectional and maximum similarity score is reported. In the present work, the maximum similarity score is used. The scores produced range from 0 to 10. Figure 3.1 shows the highest scoring alignment of the pre-searched conformations of carbamazepine and levetiracetam. The overlay shows excellent concordance with regards to overall surface shape and correspondence of charge, illustrated by the overlay of the carboxamide group. The similarity score between these molecules was 8.8.

One of the earliest descriptions of the use of molecular shape in relation to the biological activity of small molecules is due to Hopfinger [35]. The conceptualization of shape comparison was based on volume overlap of molecules that were modeled as collections of spheres. The notion of spherical volume overlap is the foundational concept of a family of molecular similarity approaches, best exemplified in current practice by the ROCS approach [73]. A separate line of thought characterizes molecular similarity by surface overlap, and one of the earliest descriptions of this notion is due to Masek et al. [60]. In that approach, molecules were characterized as having “skins” of a particular thickness, and the volume of the surface was described by the difference between a collection of spheres with standard atomic radii and one of radii made larger by the skin thickness. Similarity was measured based on the shared skin overlap between two molecules, offering some advantages over volumetric approaches, for example, when comparing molecules of very different overall sizes. The notion of molecular surface comparison is best exemplified by the Surflex-Sim approach [41], which owes its genesis to the Compass 3D-QSAR approach [39].

In what follows, volume- and surface-based molecular similarity computation
will be described, beginning with the use of Gaussian functions to transform computations about hard spheres into more tractable algorithms, followed by the representation of polarity, and strategies for conformation and alignment (pose) optimization.

3.2.1 Similarity Metrics: Volume Based and Surface Based

The fundamental underpinning of widely used 3D molecular similarity approaches makes use of the approximation that a molecule in a particular pose can be thought of as a collection of spheres with radii that depend on atomic type and also may have different chemical properties such as charge. The fundamental distinctions among metrics revolve around a basic choice between volume comparison or surface comparison.

3.2.1.1 Shape Similarity

Grant and Pickup [27] showed that the volume of a molecule could be closely approximated by using smooth Gaussians to represent molecular volume instead of using a binary exclusion function and generalized this notion to encompass molecular shape similarity [26]. Conceptually, the idea is quite simple and creates a single density function from the contributions of all individual atoms. Some complexity arises from the need to avoid “double-counting” of atom-atom self-intersections, as shown in the following equations.

\[ Q_i^g(r_i) = p_i e^{-\alpha_i r_i^2} \]  

\[ Q^g(r) = \sum_i Q_i^g - \sum_{i<j} Q_i^g Q_j^g + \sum_{i<j<k} Q_i^g Q_j^g Q_k^g - \sum_{i<j<k<l} Q_i^g Q_j^g Q_k^g Q_l^g + ... \]  

Eq. 3.1 defines an atom-centered Gaussian function in radial coordinates, and Eq. 3.2 defines an overall density function that accounts for atomic overlaps.
By choosing $p = 2.7$ and setting the $\alpha$ values based on the relationship between integral Gaussian volumes and sphere volumes with particular radii, it was shown that the integrals over this density function corresponded closely to the volume of the union of spheres representing a molecule. This formulation of volume as a density function leads to a natural means to compare molecular shapes, where similar shapes are conceptualized as having proportionately high volume overlaps. The attractive mathematical properties of Gaussians that allow for the collapsing of complex sums of products into much simpler forms whose integrals and gradients are tractable allows for rapid computation. This volumetric density formulation is used by ROCS [73], following the work of Grant et al. [26].

Rather than using volumes, the appeal of using surfaces has two aspects. First, interactions between small molecules and proteins occur between surfaces, and there is a direct relationship between binding free energy and encapsulated hydrophobic surface area of a ligand. Second, as pointed out by Masek et al. with their molecular skins approach [60], comparison of molecules with different sizes based on shared volume maximization can produce odd results (i.e. embedding a small molecule in the middle of a larger one). While conceptually attractive, the molecular skins approach was computationally burdensome. A different approach to capturing molecular surfaces was proposed during the development of the Compass 3D QSAR technique [39]. A collection of observation points (conceptually a virtual protein) was used from which to measure the minimum distance to a molecule’s surface, and this distance was compared to a learned ideal distance. This basic concept was quickly generalized to define a similarity measure that used a Gaussian reward function [40]. Similarity functions of this type correspond very closely to a surface density function formulation of molecular shape, as follows.

$$M_i(r_i) = e^{-(r_i - \mu_i)^2/\gamma} \quad (3.3)$$
Figure 3.2: Volumetric and surface-based molecular density functions for benzamidine.

\[ E_k^P(r_k) = e^{-(r_k - d_k)^2} / \gamma \]  \hspace{1cm} (3.4)

\[ R(r) = \left( \sum_i M_i \right) \left( \sum_k E_k^P \right) \]  \hspace{1cm} (3.5)

In the volume-oriented density function, Gaussians are atom-centered. In the surface-oriented formulation, the \( M_i \) of Eq. 3.3 are Gaussians with peaks at the atomic surface (set by the atomic radii, denoted \( \mu_i \)). By itself, the sum over the \( M_i \) produces internal molecular surfaces in addition to external ones. The \( E_k^P \) of Eq. 3.4 define Gaussians on local radial coordinates around each observer point from set \( P \), with peaks at the molecular surface (set by the minimum distances from the observers to the molecule, denoted \( d_k \)). When \( \gamma \) is chosen carefully, the integral of the product of two molecules’ surface density functions \( R \) (defined in Eq. 3.5) is very closely approximated by the morphological similarity function used by Surflex-Sim [41].

Figure 3.2 depicts the volumetric and surface density functions for benzamidine. The molecule benzamidine was placed in a coordinate frame such that the
XY plane bisected the aromatic ring. On the left, the volume density function $Q^g$ is depicted (computed to sixth order for intersections), with the red shaded area indicating the points on the XZ plane with significant computed density. On the right, the surface density function $R$ is depicted, again with the significant density shown with red shading. The green curves indicate the relative value of the density functions along the X axis, penetrating two hydrogen atoms and three aromatic carbons. The area of significant volume density closely covers the collection of atomic spheres, with some smoothing at saddle points. The green curve exhibits five maxima, corresponding to each of the atomic centers upon which the Gaussians were centered. The surface density function leaves the interior of the molecule with extremely low values, creating a peaked zone that also shows smoothing at saddle points.

These density functions form the underlying physical basis of two families of similarity computations, one which equates shape similarity with volume overlap and the other with surface congruence. In each case, molecular similarity can be computed based upon the integral of the product of two molecules’ density functions. In moving from a direct approximation to direct volume or surface overlap to a definition of molecular similarity, physical similitude gives way to heuristics.

The ROCS approach begins with a metric that considers the differences in density of two molecules, as follows.

$$V^{A,B} = \left( \int (Q^g_A(r) - Q^g_B(r))^2 \, dr \right)^{1/2} \quad (3.6)$$

If two molecules $A$ and $B$ have identical volume density functions, the integral above will yield zero. Multiplied out, the equation produces the following.

$$\left( V^{A,B} \right)^2 = \int Q^g_A(r)^2 \, dr + \int Q^g_B(r)^2 \, dr + \int Q^g_A(r)Q^g_B(r)^2 \, dr \quad (3.7)$$
The first two terms are the self-overlaps (squared) of the individual molecules, and the last term is the overlap between the volume density functions of the two molecules. The first two terms are independent of the relative alignment of the two molecules, but the last is not. Rewriting the equation slightly, one can see the relationship between this fundamental shape comparison formula and the Tanimoto and Tversky similarity functions.

\[ S^{A,B} = I^A + I^B - 2O^{A,B} \] (3.8)

\[ Tanimoto^{A,B} = \frac{O^{A,B}}{I^A + I^B - O^{A,B}} \] (3.9)

\[ Tversky^{A,B} = \frac{O^{A,B}}{\alpha I^A + (1 - \alpha)I^B} \] (3.10)

Typical usage of ROCS embraces the Tversky similarity function (Eq. 3.10) with \( \alpha = 0.95 \), considering molecule A to be the “query” to which others are to be matched. In practice, values of the function are computed and reported using both A and B as the object with the high weight factor. This matters very little in comparing molecules of similar overall volume, but can matter significantly when comparing molecules of very different sizes.

Returning now to the surface density approach (Eq. 3.5), the density function was the product of two sums, one “lighting up” the surfaces of each atom of a molecule and the other lighting up the surfaces of spheres packed around the molecule. The product produces a function that has significantly non-zero values only at points close to the overall molecular surface, as shown in Figure 3.2. Consider two molecules A and B and one set of “observation” points \( P \), giving rise two the following two density functions.

\[ R^A(r) = \left( \sum_i M_i^A \right) \left( \sum_k E_k^{P,A} \right) \] (3.11)

\[ R^B(r) = \left( \sum_i M_i^B \right) \left( \sum_k E_k^{P,B} \right) \] (3.12)
Here, the two surface density functions are defined with respect to a single set of observation points $P$. The spheres that “pack” around each of the two molecules $A$ and $B$ share the same centers, but they have different radii, depending on the minimum distance to each molecular surface. As with the ROCS approach, one can define a similarity metric in terms of the overlap integral of the product of the two surface density functions. This function is very closely approximated by the function computed by Surflex-Sim, simplified slightly in what follows.

$$S_k^P (d_k^A, d_k^B) = e^{-((d_k^A - d_k^B)^2)/\sigma}$$  \hspace{1cm} (3.13)$$

$$S_{A,B}^P = \sum_k S_k^P (d_k^A, d_k^B)$$  \hspace{1cm} (3.14)$$

Here, the Gaussian terms are soft reward functions for concordance of the distances from the observer points $P$ measured to the molecule surfaces of $A$ and $B$ (denoted $d_k^A, d_k^B$). When $\sigma$ is roughly twice $\gamma$ from Eqs. 3.3 and 3.4, the equivalence to the surface overlap integral holds. The intuition behind the metric is simple: when the minimum distances from each observer to each molecule are similar, the molecules must exhibit the same surface shape. The initial generalization of the Compass conceptualization of molecular surfaces to a similarity function [40] made use of two concentric spheres of observation points with radii of 6.0 and 9.0Å, but this definition was somewhat restrictive. The morphological similarity function used by Surflex-Sim [41, 42] defines an infinite grid of observer points, with weights set such that a shell of observer points around each molecule subject to a comparison contribute. In practice, finite observer sets having significant weight are selected, and alignment optimization is done using the set constructed with respect to the query ligand. Similarity scores are reported using that set, another set constructed with respect to the final aligned new ligand, and a merger of the two. As with the issue of $\alpha$ for the ROCS Tversky function, it is when molecules are of different sizes that the similarity values derived from
Figure 3.3: Volume-based and surface-based alignment of aminomethylcyclohexane to benzamidine.

differences in these three observer sets varies.

Figure 3.3 shows the optimal overlay between aminomethylcyclohexane (AMC) and benzamidine according to both the volume and surface oriented formulations just discussed using pure shape. Benzamidine is shown in green, with AMC shown in magenta (left) and cyan (middle). At right, the relationship between the ligands derived from their experimentally bound poses to trypsin are shown (PDB codes 1TNG and 3PTB). The two similarity alignments differ only by 0.3 Å rmsd, but qualitatively, there is clearly a difference. The volume-based alignment exhibits tighter congruence of atoms (and thus bonds as well), but the surface-based approach produces a slightly tilted orientation of AMC. The bottom portion of the molecules can be seen to be favoring a centered alignment based on surface considerations instead of leftward.

Figure 3.4 shows how the underlying functions that drive molecular similarity are related to the alignments of AMC to benzamidine. At left, the contour lines of the volume overlap function are shown (from the XY plane), which highlights the underlying structure of the atom-centered Gaussian functions, showing very close correspondence between the coincidence of AMC’s atomic centers with those of
Figure 3.4: Relationship of molecular alignments to underlying similarity functions.

benzamidine. At right, the respective surfaces are shown, with the observer points from the similarity computation in gray, and the differences between the distance to benzamidine’s surface and that of AMC shown in yellow rods. Where possible, rods on one or the other side of the molecules tend to have similar length reflecting a balance in surface discrepancies that is averaged over all of the observer points. In the front of the display, because AMC is thicker than benzamidine, the rods point outward from the observers, reflecting closer distances from the observers to AMC than to benzamidine. Points at which the surfaces are concordant show no yellow rods. The corresponding surface density overlap contours are also shown, exhibiting a clear relationship to the underlying surface density functions as seen in Figure 3.2.

3.2.2 Electrostatic Similarity

The foregoing has addressed only the molecular shape aspect of 3D molecular similarity, but, obviously, the degree to which the polar moieties of two molecules are congruent is important as well. The volumetric formulation used by ROCS
addresses this issue by defining atomic “colors” according to atom types, with steeper atom-centered Gaussians than those for shape and with flexibility as to weighting. The density overlap integral consists of the shape terms above along with the Gaussians of pairs of same-colored atoms from the two molecules being compared. Atom types such as those defined by Mills and Dean [62] (e.g. donor, acceptor, anion, cation, etc.) are typically used. Directional preferences of, for example, hydrogen bonds are not modeled.

The surface-based similarity approach of Surflex-Sim explicitly models hydrogen bond donors and acceptors, formal charges, and the directionality of polar interactions. All molecule atoms are labeled as being hydrophobic, hydrogen bond donors, hydrogen bond acceptors, or formally charged atoms (charge is automatically delocalized where needed, as in carboxylates). From each observer point, in addition to computing the distance to the closest atom of any type, which gives the pure shape of the molecule, distances are also computed to nearest polar positive atom (this includes donors and atoms ascribed positive charge) and polar negative atom (acceptors and atoms ascribed negative charge). For a particular observer point, directionality is treated by comparing the preferred interaction direction of a polar atom to the vector from that atom’s surface to the observer. The coincidence of these directions is combined with formal charge to yield a strength value. Similarity from each observer point’s perspective is maximized when both molecules produce the same distances and strengths, with partial similarity resulting from, for example, a hydrogen bond donor being aligned with the hydrogen of a charged nitrogen.

Figure 3.5 shows the optimal alignment of two competitive muscarinic antagonists using the full Surflex-Sim function, including both shape and electrostatics. The 2D structures are shown above the optimal mutual alignment, with the quin-
Figure 3.5: Optimal alignment of two muscarinic antagonists using surface shape and polarity.

The uclidinene derivative shown in cyan carbons. At right, the individual molecules are shown in the same pose with atomic surfaces and rods that indicate surface areas that have high similarity between the two. Green rods indicate high shape similarity, red indicate high similarity in the hydrogen-bond acceptor position and directional preference, and blue indicate concordance of the charged amine hydrogen atoms. This is an example where very high 3D similarity (0.82 on a scale of 0 to 1) obtains from molecules having different underlying scaffolds.

3.2.3 Other Methods

The foregoing has described the theoretical basis for ROCS and Surflex-Sim, two of the most mature and widely used 3D similarity approaches, representing the
volume-based and surface-based conceptual formulations of the shape question, respectively. This is not intended to be a comprehensive review of such methods, but some other approaches merit mention. The FlexS approach [56] takes an atom-centered Gaussian approach and augments it with dummy atoms to help represent directionality considerations of polar moieties. The recently published Phase Shape method [76] is closely related to the ROCS formulation, but it reverts to using spherical overlap volumes instead of using the Gaussian approach. The method uses a less rigorous (but more computationally expedient) approximation to true molecular volume overlap, avoiding the computation of the nested intersection terms seen in Eq. 3.2.

3.2.4 Conformation and Alignment Optimization

All 3D similarity methods are dependent on the conformations of the molecules to be compared, and all in common use are dependent on the particular alignment of the molecules as well. This property derives from their direct relationship to what physically makes molecules similar to one another in biological systems. Because of this, in order to compare a molecule to a single pose of another, a 3D similarity approach must identify both the conformation and alignment that yields a maximal similarity value.

Typical drugs and drug-like molecules may have a few rotatable bonds (e.g. aspirin, a COX-1/2 inhibitor) but can have more than ten (e.g. saquinavir, an HIV protease inhibitor, has thirteen). Methotrexate, an old drug and inhibitor of dihydrofolate reductase, has nine rotatable bonds, and it has been used frequently to test search strategies involving conformation and alignment optimization. Even a very coarse sampling of conformational space of three rotamers per torsion produces roughly 20,000 conformations. A more generous sampling of six rotamers
produces over ten million conformations. The issue of alignment generates a multiplicative increase in complexity, because the conformations and alignments must be considered together. Assuming translational uncertainty of ± 5 Å and a sampling requirement of 1 Å, sampling of translational and rotational space for a molecule such as methotrexate requires over ten million rigid alignments. Clearly, brute-force enumeration of all energetically reasonable conformations (with alignments sampled to roughly 1 Å) is not feasible even with a very efficiently computed similarity function. There are two basic strategies that have been taken for the conformational problem and two for the alignment problem.

3.2.5 Conformational Optimization

One way to address the conformation question is to search each ligand of interest independently of any other consideration and to retain some maximal number of individual conformers. A typical value for the number of retained conformers is 200, allowing for quite complete sampling of molecules with up to five or six rotatable bonds. This does not provide dense sampling for drug molecules such as methotrexate or saquinavir, but in applications such as virtual screening of very large libraries, the speed requirements may necessitate some tradeoffs.

Figure 3.6 illustrates results for agnostic conformation generation. At left, biotin is shown. It has five rotatable bonds, with two ring conformations of reasonable energy. The lowest-energy conformation is shown in a “canonical” alignment to the Cartesian coordinate system, with the molecular centroid at the origin, the largest radial excursion parallel to the Y axis, and the largest excursion in the XZ plane rotated to be within the XY plane. Conformational expansion of biotin, with a maximal sampling of 200 conformers, contains a conformer that is 0.35 Å RMSD different from the conformation of biotin bound to
Figure 3.6: Conformational sampling independent of molecular alignment.

streptavidin (PDB code 1STP). For methotrexate, the conformational variation is clearly much more significant, particularly in terms of the different conformers that the tail of the molecule can exhibit. In this case, the minimum RMSD from a 200 conformer sample is 1.17 Å when compared with the configuration bound to DHFR (PDB code 4DFR). Conformational enumeration of this type typically takes seconds per molecule, and it need be done only once, offering the resulting sampled molecular state for further processing for the negligible cost of retrieving the conformations.

The other approach is to treat the conformational search as part of the overall similarity optimization procedure, with strategies to identify only those conformations that are likely to yield good matches to the reference molecule. Such strategies include divide and conquer approaches that fragment molecules into significantly less flexible pieces, search those relatively thoroughly, and either incrementally reconstruct the partial solution or make use of some type of crossover procedure. The idea is that a molecule with, for example, 11 rotatable bonds may be broken into three fragments by severing two torsions. If the torsion-
breaks are chosen carefully, the three resulting fragments will each have three rotatable bonds. Coarse sampling of three such fragments yields less than 100 total conformations \((3 \times (3^3))\), and more thorough sampling produces about 600 \((3 \times (6^3))\). This compares to roughly 200,000 for coarse sampling of the un-fragmented molecule. If it is possible to produce reasonable alignments for the conformations that will form part of a close to optimal solution \textit{independently} of one another, then such a strategy can be very effective in identifying high similarity poses. The broken torsions receive their configurations during a reconstruction process.

The question of whether independence is a good assumption or not is essentially an empirical one. However, certain molecular types present known difficulties, such as a molecule with a small central scaffold from which emanate multiple “arms” that can clash with one another. In such a case, independent alignment of the arms may often lead to incompatible geometries for successful reconstruction of a high similarity final pose. Crossover procedures that recombine full solutions instead of incrementally constructing partial solutions can be effective even in these cases. Such procedures were developed initially for docking [44, 38], and they have been adapted for similarity optimization as well.

### 3.2.6 Alignment Optimization

Similar to the question of conformational optimization, the problem of alignment falls into fast sampling approaches and approaches that attempt to make more clever choices based on the context of the particular molecules in question.

Figure 3.7 shows a strategy for generating alignments of a conformation of one molecule A onto a query conformation B. In the example, the query conformation of biotin (as bound to streptavidin in PDB code 1STP) has been canonically
aligned within the Cartesian frame. In a case where one has a highly similar molecule (here biotin is shown as the ligand to be aligned) and the conformational sampling is adequate, simple canonicalization of A can yield a result very close to optimal. Here, beginning with a randomized initial biotin conformation, sampled agnostically as described above, each of the conformations was placed in the canonical alignment to the Cartesian frame. The conformations included the one shown in Figure 3.7, which is clearly very close to the identity alignment. The alignments were generated without any computation of or consideration of molecular similarity. In cases of molecules with highly similar shapes, very limited alignment sampling of a reasonably thorough conformational sample can produce excellent results quickly. This requires only limited evaluation of the similarity function and local optimization in order to produce close to optimal molecular overlays.

As molecules begin to differ, the major axes may align well, but minor axes may not line up correctly, in which case generating a flip or a sampled spin around the axis will help identify high-similarity alignments. Of course, the same principle applies for the major axis, and vertical flips may also be required. Figure 3.7 shows the systematic spinning of biotin around its major axis. It may also be
necessary to consider the “flip” of each conformation along its major axis. But as molecules become less similar, finding close to optimal alignments becomes increasingly challenging. The problem can be treated generally by choosing some minimal spatial sampling interval and ensuring that alignments will be generated to cover the chosen density. This can be done efficiently by treating a conformation as an elliptical body to be spun around its long axis and rotated such that its poles mark out a spherical tessellation.

At the right of Figure 3.7, the uniform sampling of major axis rotation and of axis-direction to sphere tessellation is shown for a single conformation of biotin. In such a sampling, the “nose” of each conformation of the molecule “sees” each point on a uniformly sampled sphere and also spins around its own axis evenly. One further complication arises with molecules of very different size. An assumption of centroid correspondence may be very poor, and this can be avoided by additional sampling (e.g. along the major axis of the molecule to be aligned). Obviously, very coarse sampling can be done within this type of scheme (e.g. the major and minor axis flips of the canonical alignment). The ROCS approach makes an aggressive choice of coarse sampling (typically four starting alignments) and relies upon the smooth behavior of the Gaussian overlap function to aid in local optimization to produce final alignments.

The other approach for addressing alignment optimization is to make specific choices of alignments on the basis of the particular conformation or conformational fragment to be aligned. For example, one can seek to identify matching triplets of points between a conformation of the molecule to be aligned and the query conformation. By ensuring that the triangle edge lengths are similar and possibly that the characteristics of the corresponding points themselves are similar, alignments can be produced quickly by identifying the rigid-body
transform that minimizes the least-squares distance differences between the corresponding triangle vertices. The Surflex-Sim approach offers both an aggressive “blind” alignment enumeration as well as a procedure that generates alignments by making correspondences between observer points of each molecule based on conformation-specific information about the molecule being aligned.

### 3.2.7 Tradeoffs Between Speed and Thoroughness

![Diagram](image)

**Figure 3.8**: Pose optimization tradeoff of speed versus quality.

Depending on the specific application, identifying a molecular pose that is extremely close to the global maximum similarity may or may not be important. In cases such as virtual screening, identification of a pose that produces a similarity score within 5% of the maximal value greater than 90% of the time may be sufficient to produce enrichment of active molecules nearly equivalent to results obtained from more exhaustive search. This is true particularly if the heuristic
sampling scheme works preferentially better on high-similarity molecular alignments. On the other hand, similarity optimization to produce hypotheses about the relative alignment of molecules with different scaffolds, especially if those alignments serve as starting points for quantitative procedures (e.g. 3D-QSAR), might be better served by more exhaustive search.

There is a clear tradeoff in speed versus search thoroughness, as illustrated in Figure 3.8. The top alignments show the combination of pre-sampled conformations combined with brute-force alignment sampling for Surflex-Sim (the approach illustrated at right in Figure 3.7). The queries were the bound poses of biotin and methotrexate, and the input poses had been randomized and minimized. For biotin, a near optimal pose was identified. But for methotrexate, owing to its additional conformational freedom, the pose that was produced exhibited deviation from optimality in the very flexible tail region. Making use of dynamic conformational choice and conformation-specific optimization, it is possible to obtain poses that are closer to optimal, as seen in the bottom panel. In the cases of biotin and methotrexate, the only real differences between the aligned and query poses exist in the linker elements.

On a modern CPU (e.g. an Intel Core i7-2600), using a single computing core, the Surflex-Sim fast screening protocol (conformational pre-search and agnostic alignments), on drug-like screening molecules against a typical query ligand (as in Figure 3.5) yields 20-50 flexible alignments per second (the middle three quintiles of speed). Small benzamidine-sized molecules are processed at more than 100 flexible alignments per second, biotin-sized molecules at 35 per second, and large molecules such as methotrexate at 15 per second. ROCS operates at roughly 20-40 molecules per second in its default mode, which is optimized for very rapid overlays. The Phase Shape method operates at about 5-10 ligands
Figure 3.9: Volume and surface shape optimization with variations of search depth.

per second. FleXS, which uses an incremental construction approach processes roughly 1 ligand per minute. GPU-based implementations can be much faster than traditional CPU targeted implementations.

In the search mode that combines molecular fragmentation, dynamic conformational search, conformation-specific similarity-guided alignments, and all-atom Cartesian coordinate final pose optimization, Surflex-Sim operates roughly 30-fold more slowly (up to a few seconds per molecule). Many gradations of speed versus thoroughness can be obtained by combining different methods and granularities for conformation sampling and alignment generation.

The foregoing discussion of pose optimization made use of examples of self/self comparison, because optimality is easy to assess by eye in such cases. When the molecules are different, particularly if they differ in size and scaffold, the problem of pose optimization is more challenging. Overlay of one molecule onto another depends on both the underlying similarity function and the effectiveness of the search and optimization protocol. Figure 3.9 shows the case of AMC being flexibly aligned to the NAPAP (a potent thrombin inhibitor, PDB code 1DWD). At left, the optimal volume-based and surface-based pure shape overlays are shown. The
volume-based overlay places AMC onto the piperadine moiety of NAPAP, being
driven by the very close atom-atom overlap that can be obtained in that manner.
The surface-based overlay places AMC onto the benzamidine moiety of NAPAP.
The molecular surface shape of NAPAP yields only one sensible maximum for
the surface-based approach because the self-packing between the naphthalene and
piperadine “hides” the internal piperadine surface from the observer points. The
resulting alignment is very close to that observed for AMC aligned to benzamidine
(see Figure 3.3), both of which bind competitively in the S1 pocket of trypsin-like
serine proteases. Note that the ROCS approach, making use of the correct color
force-field choices and optimization strategy would yield a similar alignment.

One the right of Figure 3.9, different alignments are shown using both shape
and polar surface considerations. The optimal alignment again places AMC onto
the benzamidine portion of NAPAP, with similarity highlighted by green (pure
surface) and blue (positive surface) lines. Using a fast screening search mode,
a reasonable alignment is produced (with a score within 5% of optimal), but
making use of a procedure that restricts initial alignments to the centroids of the
molecules yields a poor result (with a score of just 50% of optimal).

Molecular comparisons of similar molecules that are close in size will tend to
yield results that vary little, whether one is using a volume-based method or a
surface-based method, and will also be relatively insensitive to fast and aggressive
search protocols. However, as the degree of similarity decreases, sharp differences
between metrics and optimization methods for identifying global maxima will
begin to appear.
3.2.8 Approximations for 3D Similarity: Extreme Speed

In making large-scale computations of molecular similarity, even with fast methods, the problem is computationally challenging. For example, pairwise comparison of all small-molecule drugs (e.g. for a clustering analysis) requires on the order of 1,000,000 similarity comparisons, each of which may require multiple individual pose optimization instantiations. A number of methods have been developed to approximate the direct computation of 3D molecular similarity in order to yield speeds that compare favorably with 2D approaches. These methods generally take the form of representing a molecule as a real-valued vector of fixed dimension, where the vector values are chosen based on the similarity of a molecule to a basis set of reference molecules. Each molecule is flexibly aligned to a fixed conformation of each of multiple basis molecules (our work has typically made use of 20 such basis molecules). For each input molecule, the result is a vector, where each value within each vector represents a single similarity computation (yielding a value between 0 and 1). Distances between these vectors can be used as a computationally cheap surrogate for the direct molecular similarity computation. Additional details on this method, including a more sophisticated algebraic treatment of vector derivation, and its application to molecular diversity computation, bioavailability prediction, and molecular clustering can be found in previous reports [63, 24, 13]. The basic concept has been exploited by other groups as well, initially in the middle-1990’s [45, 7]. Later, very closely related ideas were proposed by multiple groups [31, 71]. Here, as discussed in the next section, we make use of 20-dimensional molecular imprints in order to make inferences about the shape of the distribution of similarity of a given molecule to a background set of randomly selected molecules.
3.2.9 Summary

Methods for computation of 3D molecular similarity involve complex functions and require careful search and optimization protocols to produce results. However, their relationship to the physical processes that underlie protein-ligand recognition events is an *ipso facto* advantage over approaches that are fundamentally non-physical in nature (e.g. many 2D topological similarity methods). Because small molecules are made generally by humans in an active design process, the relationship of a similarity method to either the causative physical basis for biological activity or to the means by which molecular design took place have special implications. These have been discussed to some extent in the previous chapter. Three-dimensional molecular similarity computation owes its utility to making molecular comparisons in a manner that is congruent to the direct causal basis for two molecules to behave similarly in biological interactions.

3.3 Imprint-based p-value normalization

In the initial work developing the approach (see previous chapter), we observed that the background distributions of nearly all drug-line small molecules are well-approximated by normal distributions. Figure 3.10 shows a histogram of the background scores of carbamazepine along with the normal curve where the mean and standard deviation was computed directly from the background set. Instead of counting the number of occurrences of similarity score of 8.8 or greater and dividing it by the cardinality of the background set to compute $p$-values (as was the case for empirical $p$-value computation), the $p$-value can be directly computed by the area under the curve that is greater than 8.8 from the cumulative distribution function, the $p$-value is therefore easily computed by Equation 3.15. Given
Figure 3.10: Molecular imprint-based $p$-value normalization. Empirically derived distributions used to normalize raw 3D similarity scores to $p$-values follow the normal distribution (green line, for carbamazepine). Therefore, normalization can be achieved by estimating the mean and standard deviation of the normal distributions from which a $p$-value can be computed given a similarity score $s$. Two linear models for the mean and standard deviation based on molecular imprints were fitted using a random set of drugs and drug-like molecules along with their computed mean and standard deviations. Given two molecules (in this example, carbamazepine and levetiracetam) with their corresponding imprints and similarity score between them $s$, the mean and standard deviation are estimated by the linear models for each molecule. For example, carbamazepine’s estimated mean and standard deviation was 6.48 and 0.57 which is nearly identical to the empirical mean and standard deviation of 6.69 and 0.59. Then, $p$-values are computed based on $s$ for each fitted normal distribution and the largest $p$-value is retained.
the normal distribution of levetiracetam, the same computation is performed for
levetiracetam and the largest of the two \( p \)-values is retained.

\[
p-value(x, \mu, \sigma) = 1 - \frac{1}{2}[1 + erf\left(\frac{x - \mu}{\sqrt{2} \sigma^2}\right)] \tag{3.15}
\]

In the previous example, the mean and standard deviation were directly com-
puted from the background distribution, but for high-throughput application of
the approach, a faster method was sought that did not require full computation of
the background distribution. Linear models relating the pre-computed molecular
imprints to the mean and standard deviation were constructed and proved to be
excellent predictor of the background distributional characteristics of drug-like
molecules. Molecular imprints are an abstract way to represent molecules as a
vector of similarity scores to a fixed set of structurally diverse basis molecules.
Previously, we have used imprints as a surrogate for 3D similarity by computing
the Euclidean distance between vectors \([63, 24, 13]\).

In Figure 3.10, the linear equations at the top for \( \mu \) and \( \sigma \) offer a means to
estimate these molecule-specific parameters, where \( b_i \) is the similarity score of a
molecule against a basis set molecule and \( w_i \) are the weights of the model. The
linear models were constructed using data for 200 randomly selected drugs and
drug-like molecules from the ZINC2 data set \([70]\). Background distributions were
generated for each molecule, from which the true mean and standard deviations
were computed. Using the imprint-based models, estimates of the mean and
standard deviation of a new molecule all that is required is the new molecule’s
molecular imprint. This needs to be computed just once, and its computation
is part of the conformational pre-searching protocol. Figure 3.10 shows the es-
timated means for carbamazepine and levetiracetam (6.66 and 6.48) and the
estimated standard deviations (0.57 and 0.59). These values were comparable
to the values computed directly from the backgrounds (shown in parenthesis). Finally, the estimated and empirical $p$-values agreed, both being significantly less than 0.01.

Table 3.1 and 3.2 show the fitted parameters for $w_i$ and $v_i$. To estimate the mean and standard deviation of a new molecule all that is required is the molecular imprint which only needs to be computed once where previously generating the background distribution was expensive because it required 1000 similarity computations.

$$
\mu = w_0 + w_1 b_1 + ... + w_1 b_1 \quad (3.16)
$$

$$
\sigma = v_0 + v_1 b_1 + ... + v_1 b_1 \quad (3.17)
$$

Using the imprint based normalization, we recomputed the $p$-values and the log-odds from our previous work. Figure 3.11A shows the 2D histogram of estimated versus the empirical $p$-values of the all-by-all similarity of 358 drugs showing that they are highly correlated (R-squared of 0.947 and Kendall’s tau of 0.814). Figure 3.11B shows the 2D histogram of the estimated versus the empirical log-odds of 358 drugs against 44 targets showing that they are also highly correlated (R-squared of 0.955 and Kendall’s tau of 0.761).

### 3.4 Conclusion

In this chapter, we have shown how log-odds calculations for predictive pharmacology can be made fast enough for broad application to problems of significant scope. Speed improvements within the underlying similarity computation engine (from several seconds per similarity computation to roughly fifty per second) had
Figure 3.11: Relationship between empirical and estimated p-values and Log Odds scores. The new normalization scheme is close related to the empirical normalization. (A) The all-by-all pairwise similarities of over 300 drugs is shown, the Pearson correlation between empirical and estimated p-values is 0.947 and the Kendall’s tau rank correlation is 0.814. (B) Log Odds computation of over 300 drugs against more than 40 pharmacologically relevant protein targets is shown, the Pearson correlation between empirical and estimated Log Odds is 0.955 and the Kendall’s tau rank correlation is 0.761.
a substantial impact. Further, approximation of molecule-specific background distributions, required for normalization of similarity scores, have removed the requirement for empirical distribution generation. The combined effect of these enhancements allows a computation of a log-odds score in a few minutes rather than several hours.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w_0)</td>
<td>2.208905</td>
<td>(w_1)</td>
<td>1.074314</td>
</tr>
<tr>
<td>(w_2)</td>
<td>0.939088</td>
<td>(w_3)</td>
<td>0.499798</td>
</tr>
<tr>
<td>(w_4)</td>
<td>0.204079</td>
<td>(w_5)</td>
<td>0.576666</td>
</tr>
<tr>
<td>(w_6)</td>
<td>1.288041</td>
<td>(w_7)</td>
<td>0.675059</td>
</tr>
<tr>
<td>(w_8)</td>
<td>0.968887</td>
<td>(w_9)</td>
<td>-1.226087</td>
</tr>
<tr>
<td>(w_{10})</td>
<td>-1.032411</td>
<td>(w_{11})</td>
<td>1.016418</td>
</tr>
<tr>
<td>(w_{12})</td>
<td>1.040275</td>
<td>(w_{13})</td>
<td>0.715834</td>
</tr>
<tr>
<td>(w_{14})</td>
<td>0.655881</td>
<td>(w_{15})</td>
<td>1.744321</td>
</tr>
<tr>
<td>(w_{16})</td>
<td>1.013583</td>
<td>(w_{17})</td>
<td>-0.363244</td>
</tr>
<tr>
<td>(w_{18})</td>
<td>-0.543741</td>
<td>(w_{19})</td>
<td>-0.301841</td>
</tr>
<tr>
<td>(w_{20})</td>
<td>0.009863</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Parameters for the mean linear model.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_0$</td>
<td>1.131467</td>
<td>$v_1$</td>
<td>0.297707</td>
</tr>
<tr>
<td>$v_2$</td>
<td>-0.122146</td>
<td>$v_3$</td>
<td>0.283737</td>
</tr>
<tr>
<td>$v_4$</td>
<td>-0.129003</td>
<td>$v_5$</td>
<td>0.036350</td>
</tr>
<tr>
<td>$v_6$</td>
<td>0.118103</td>
<td>$v_7$</td>
<td>-0.051102</td>
</tr>
<tr>
<td>$v_8$</td>
<td>0.258132</td>
<td>$v_9$</td>
<td>-0.006762</td>
</tr>
<tr>
<td>$v_{10}$</td>
<td>-0.358383</td>
<td>$v_{11}$</td>
<td>-0.123263</td>
</tr>
<tr>
<td>$v_{12}$</td>
<td>0.275085</td>
<td>$v_{13}$</td>
<td>0.027236</td>
</tr>
<tr>
<td>$v_{14}$</td>
<td>-0.095684</td>
<td>$v_{15}$</td>
<td>-0.087388</td>
</tr>
<tr>
<td>$v_{16}$</td>
<td>-0.399395</td>
<td>$v_{17}$</td>
<td>-0.229595</td>
</tr>
<tr>
<td>$v_{18}$</td>
<td>-0.618184</td>
<td>$v_{19}$</td>
<td>-0.154896</td>
</tr>
<tr>
<td>$v_{20}$</td>
<td>-0.672807</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Parameters for the standard deviation linear model.
CHAPTER 4

Prediction of Off-Target Biological Effects through Data Fusion

4.1 Abstract

Information from the linkage between chemical structures and biological targets has been shown to support prediction of clinically relevant off-target biological effects through computation of chemical structural similarity. Here, information about the clinical effects of chemicals, derived from patient package inserts, is shown to support computation of similarity based on lexical analysis of pharmacological descriptions. Similarity based on structure and on lexical analysis is shown to be synergistic, allowing drug-target linkage data to be used to predict new drug-target interactions. Using a simple probabilistic framework, chemical similarity values derived from 2D structure comparison, 3D structure comparison, and lexical comparison of drug package inserts are combined in order to make predictions of surprising off-target effects. The method supports predictions that cannot be confidently made using simple topological structure comparison and is easily generalized to include information from other sources such as docking calculations.
4.2 Introduction

In prior work, we introduced a methodological approach for data fusion which was used to predict the protein targets of small molecules based on molecular similarity [91]. Given a test molecule and a set of small molecules with a known shared biological effect, the method produces a single scalar value that corresponds to the likelihood that the test molecule will share the same activity. The framework allowed us to investigate the types of pharmacological inferences that 2D and 3D structural similarity methods can generate, and this methodology was applied to roughly one third of FDA approved drugs. We showed that for predicting primary targets (i.e. targets modulating intended therapeutic effects) the performance advantage of a 3D similarity method over a 2D method was relatively small. This not especially surprising because it has been shown that there is a substantial human 2D bias in drug design results in the production of numerous “me-too” drugs [14, 91]. However, for predicting secondary targets (i.e. sources of side-effects) 3D similarity was much more effective than 2D topological comparisons. Additionally, we showed that drug pairs with novel chemical scaffolds (quantified by high 3D but low 2D similarity) are much more likely to have pharmacologically relevant differences in protein target modulation compared to drug pairs that are structurally similar (high 2D and high 3D similarity).

In the previous work, we also showed that clinical effects of drugs could be used as a surrogate for biochemical characterization [91]. This was achieved by a detailed manual curation of a large set of drugs for which no specific positive or negative annotation as muscarinic receptor ligands existed. For each drug, the package insert was surveyed for muscarinic related side-effects such as dry mouth, urinary retention, and blurred vision, and a binary label was applied indicating the best interpretation about the existence of muscarinic effects for each drug.
Using chemical similarity to a set of established muscarinic ligands, predictions were made on the clinically annotated set, resulting in strong separation of likely muscarinic modulators from those with no evidence of such effects. While the manual curation of nominal target interactions based on either literature or drug package insert information is possible, performing large scale analyses using manual curation strategies is not feasible. In the current work, we report a means to automatically extract relevant textual terms from drug package inserts and to compare sets of such terms quantitatively.

To illustrate that clinical effects reported in patient package inserts are relatable to the underlying protein targets, consider the molecules shown in Figure 4.1. The top row shows two first generation sulfonylureas, tolbutamide and tolazamide, that are prescribed to treat diabetes. Tolbutamide was the first in class drug discovered for treatment of type 2 diabetes. It was developed by the Upjohn Company and approved by the FDA in 1961. At the time, the only treatment available for diabetes was insulin, but the discovery of tolbutamide stimulated the development of other small molecule therapeutics. Development of structurally similar sulfonylureas continued, and tolazamide was shown to have similar pharmacological effects as tolbutamide [89], as expected given their high structural similarity. Tolazamide was subsequently approved in 1966, and it was later discovered that the therapeutic benefit of these drugs was due to interactions with ATP-sensitive K+ channels. Specifically the $K_{ir}6.2$-SUR1 complex is thought to be responsible. It is expressed in pancreatic beta cells and, when blocked, promotes the release of insulin [65]. A few examples of the common clinical terms present in the patient package inserts are: “diabetes mellitus” (the indication), “blood glucose” (the physiological target), and “hypoglycemia” (an undesired side effect). Commonality of such terms coincides with high structural 2D and 3D similarity. In addition to interactions with potassium channels,
Figure 4.1: Relationship between small molecules based on molecular similarity, protein target modulation, and clinical effects. The top row shows two antidiabetics, tolbutamide (first in class) and tolazamide (follow on) which are highly structurally similar, interact with similar proteins, and have similar clinical effects. The bottom row shows two anti-epileptic drugs, carbamazepine and levetiracetam, thought not to share protein targets but having significant 3D molecular similarity and similar clinical effects. Surflex-Sim’s 3D overlay of carbamazepine (green) and levetiracetam (atom color) is shown. Green sticks correspond to regions of significant surface shape similarity and blue/red sticks correspond to regions of significant polar similarity.
the drugs share interactions with the peroxisome proliferator-activated receptor gamma-retinoid X receptor heterodimers (PPARγ-RXR).

Next, consider the two anticonvulsants on the bottom of Figure 4.1, carbamazepine and levetiracetam. Carbamazepine was one of the first anticonvulsants approved by the FDA in 1968 and its therapeutic benefit is attributed to stabilizing the inactivated state of voltage-gated sodium channels (Nav1.1) [72]. Levetiracetam is a newer anticonvulsant that has been suggested to have a novel mechanism of action through interaction with synaptic vesicle glycoprotein 2A (SV2A) [58]. As expected, the two package inserts have clinical effect terms in common that relate to the therapeutic target of these drugs which include “status epilepticus” and “grand mal.” High 3D similarity between carbamazepine and levetiracetam is shown in the overlay at the bottom of Figure 4.1, despite little topological commonality. Here, as will be discussed in more detail later, given the agreement between clinical pharmacological terms and high structural similarity, it would be surprising if these drugs shared no molecular targets, especially because the effects of anticonvulsants can be mediated through a wide range of protein targets [82].

The present study establishes a computational method to draw relationships between drugs based on the clinical effects present in Patient Package Inserts (PPI). As described above in the muscarinic example, small molecules that share protein targets have similar clinical effects beneficial or not. Therefore, it is likely that drugs with shared targets will also share pharmacologically relevant terms present in their package inserts. Use of such information from package inserts, specifically side-effect information, to predict drug target interactions has previously been demonstrated [9]. The present study makes three primary contributions. First, we show that the means by which drug similarity is auto-
matically computed from package inserts is directly correlated with drug struc-
tural comparisons. Drug pairs that were structurally similar (both by 2D and
3D comparison) more frequently shared clinical effects than drug pairs that were
structurally novel (low 2D but high 3D similarity), and that drug pairs sharing
no structural similarity exhibited very low clinical effects (or PPI) similarity. Sec-
ond, we established that the combination of 2D, 3D, and PPI similarity yielded
better predictive performance over any single similarity computation, especially
when predicting off-target interactions. This was shown both by leave-one-out
cross-validation within our internal Structural Pharmacology Database (SPDB)
and on an entirely new set of drug-target annotations derived from ChEMBL.
Roughly speaking, recovery of roughly 40–50% of secondary target annotations
was possible with false positive rates of 1–3%. Third, we applied the method
systematically across all drugs and targets within the SPDB and identified three
cases from among the highest scoring predictions where drug-target effects out-
side of either the SPDB or ChEMBL have either been shown definitively or have
very strong literature support.

4.3 Methods and Data

The following describes the molecular data sets, computational methods, and
specific computational procedures. Data and protocols are available for download
(see http://www.jainlab.org for details).

4.3.1 Molecular Data Sets

In the present study two molecular data sets are used. The Structural Pharma-
cology Database (SPDB) is a deeply curated drug target database that is used as
the basis to make predictions. A set of drug target annotations from ChEMBL that were not annotated in our database were used as a blind test set.

4.3.1.1 Structural Pharmacology Database

The details of the SPDB and its relationship to other databases has been extensively described elsewhere [13, 14, 91]. Two features that are important for the present study will be discussed. First, we consider targets as specific binding sites on proteins or protein complexes. For example, the GABA	extsubscript{A} receptor has distinct binding sites for benzodiazepines and barbiturates, each of which is annotated as a separate target. It is critical to make such distinctions in order to make inferences about small molecule activity based on structural similarity. Second, we distinguish between primary targets, those that are believed to be therapeutically beneficial, and secondary targets which mediate pharmacologically relevant off-target effects. In many cases, predicting primary targets is neither challenging nor especially relevant due to the presence of me-too analogs and a strong 2D bias in drug design. For inference of secondary targets, representing unexpected interactions, such a structural bias does not exist. By making this distinction, it is possible to explicitly quantify performance of methods for prediction of relatively surprising effects.

For each drug in the SPDB (~1000), we sought a corresponding patient package insert (PPI) having the drug as the sole active ingredient (drug combinations were excluded to avoid mistakes in ascribing clinical effects to the incorrect drug). PPIs were downloaded from DailyMed [1], provided by the National Library of Medicine, which has current prescribing information from the FDA. A total of 760 drugs had corresponding PPIs. For drugs with multiple PPIs (due in part to different manufacturers and forms of administration), a representative PPI was
randomly selected. From all 760 PPIs, pharmacologically relevant text was extracted and filtered to include only terms that were part of medically controlled vocabularies (more details follow below). The set of PPIs were filtered to only include those that had at least 100 medical terms, reducing the set to 602 PPIs. The corresponding 602 drugs had specific annotations against 257 targets in the SPDB, of which 91 targets had at least 5 drugs annotated as ligands. The 91 targets were comprised of 83 human proteins, including 28 aminergic GPCRs, 19 ligand and voltage gated ion channels, 13 human enzymes, 7 nucleotide and short peptide GPCRs, 5 tyrosine kinases, 5 steroid receptors, 3 reuptake transporters, 2 ion transporters, and 1 transcription factor. The remaining 8 targets were bacterial, fungal, and viral proteins.

4.3.1.2 ChEMBL Blind Test Set

To test our methodology, we sought a set of drug target interactions that were not present in the SPDB that could serve as a blind test set. To find new interactions, ChEMBL version 14 was used as a data source. ChEMBL has vast coverage of small molecule to target bioactivities including over a million distinct compounds with over 10 million bioactivities against 9,000 targets [23]. For each of the 602 drugs, corresponding ChEMBL compounds were identified based on direct structural equivalence. Equating SPDB targets to ChEMBL targets required manual curation in order to identify which ChEMBL activities corresponded to the specific binding sites curated within the SPDB. Out of the 91 targets used in this study, there were 65 corresponding ChEMBL targets. A threshold to define significant activity was set as those that had $K_d$, $K_i$, or $IC_{50}$ activities that were less than or equal to 1$\mu$M. The common intersections between the annotations of SPDB and ChEMBL included roughly 50% of the total number
from both. Annotations within the SPDB are based on careful inspection of primary literature and pharmacology texts such as Goodman & Gilman’s the Pharmacological Basis of Therapeutics [32], the latter of which ChEMBL does not index. There were 380 drug-target interactions present in ChEMBL that were missing from the SPDB matrix of 602 drugs and 91 targets. This set served as a blind test set and will be referred to as the ChEMBL set in what follows.

4.3.2 Patient Package Insert Similarity

The well established vector space model [74] has been extensively used in the field of information retrieval for a wide range of applications such as document relevance ranking and classification [30]. This method was employed to model patient package inserts (PPIs). In the vector space model, text documents are modeled as vectors in high dimensional space where each dimension corresponds to a term with an associated weight. Conceptually, if two documents are related, they will share a significant portion of terms, quantified by the cosine similarity metric. As for drugs, if two drugs have similar clinical effects, they will share a significant portion of pharmacologically relevant terms present in their respective PPIs.

The process to transform PPIs into weighted term vectors is shown in Figure 4.2A. Given a PPI as input, the process involves four steps. The first step is to extract pharmacologically relevant sections from PPIs which will serve as the lexical basis to draw relationships between drugs: Indication, Contraindications, Precautions, Adverse Reactions, Drug Interactions, and Clinical Pharmacology. The second step is to generate terms of varying sizes using a cutoff of five words or less per term. This is achieved by removing punctuation and employing a window of varying word sizes. This normalization step may generate incorrect terms
Figure 4.2: Modeling patient package inserts in the vector space model. (A) Flow chart of term vector generation. Pharmacologically relevant sections from a patient package insert are extracted and combined including the following sections: Indication, Contraindications, Precautions, Adverse Reactions, Drug Interactions, and Clinical Pharmacology. From the combined sections, terms are generated that have five words or less and are filtered using the MedDRA and MeSH controlled vocabularies to include only clinically relevant terms. Terms that appear less frequently (e.g. “syncope and collapse”) have a higher weight than more frequent terms (e.g. “grand mal”). To agnostically weight terms, the Google 1T data was used. The output is a vector where each dimension corresponds to a clinically relevant term with an associated weight. (B) Excerpt of the procedure outlined in (A) applied to the PPI of the anticonvulsant carbamazepine.
by removing punctuation, which can join adjacent words that should not be part of the same term. Also, terms that are not pharmacologically important will also be generated. The third step addresses these issues by only retaining those terms that are part of the Medical Subject Headings (MeSH [2]) or low-level Medical Dictionary for Regulatory Activities (MedDRA [8]) controlled vocabularies. The final step involves assigning term weights, because some terms are more information rich than others (e.g. “generalized seizures” has more content than just “seizures”). We used the term frequencies in the Google Web 1T 5-gram Corpus Version 1 [6] to compute term weights. The corpus is the result of a comprehensive effort by Google to count the number of occurrences of terms up to 5 words in length in approximately 1 trillion web pages [6]. The term weights are computed by dividing the number of occurrences of a term by the total number of sentences parsed by (95,119,665,584), taking the log of the quotient, and inverting the sign. For example, “seizures” occurs 1,746,834 times in the corpus and therefore its weight is \(-1 \times \log(1,746,834/95,119,665,584)\) which is 4.74. However, the more specific term “generalized seizures” occurs 12,255 times and therefore its weight is \(-1 \times \log(12,255/95,119,665,584)\) which is 6.89. The overall effect is that terms which occur less frequently in the corpus have higher weight than those that occur more frequently. The final output of the process is a drug PPI vector composed of weighted terms present in pharmacologically relevant sections of the PPI. This work employed 602 drugs that had corresponding PPI vectors of 100 terms or greater.

Figure 4.2B shows an example of how a short phrase is processed (from the Indication Section of the PPI of the anticonvulsant carbamazepine). The input text states: “patients with the following seizure types: partial seizures with complex symptomatology (psychomotor, temporal lobe).” The process begins by removing the punctuation and generating terms of varying sizes. The unigrams
that are extracted are the following: patients, with, the, following, seizure, types, partial, seizures, with, complex, symptomatology, psychomotor, temporal, and lobe. Sliding a window of 2 words generates all possible adjacent bigrams, which are: patients with, with the, the following, following seizure, seizure types, types partial, partial seizures, seizures with, with complex, complex symptomatology, symptomatology psychomotor, psychomotor temporal, and temporal lobe. The same procedure is repeated using a sliding window of size 3, 4, and 5 words. A partial listing of the generated trigrams are shown in Figure 4.2B. The terms are filtered to discard terms that are not medically relevant (e.g. with the following) or incorrect (e.g. seizure types partial seizures) by only retaining MeSH and MedDRA terms. The terms that were retained are: patients, seizure, seizures, partial seizures, and temporal lobe. Finally, the weights for each of the terms are computed by the scheme described above. A partial listing of the 371 terms of the carbamazepine’s PPI vector are shown at the bottom of Figure 4.2B. The two most heavily weighted terms along with terms as well as terms relevant to the therapeutic benefit are shown.

To quantify the similarity between a pair of drug PPI vectors, the cosine similarity metric is used (shown in Equation 4.1). A raw similarity score of 1 corresponds to having identical terms, a score of 0 corresponds to no shared terms, and a score in between corresponds to having some overlap. In Equation 4.1, \( n \) is 6,591 which is the total number of terms present in all 602 PPIs and \( A_i \) and \( B_i \) are vectors of the size \( n \) with either the weight for the term if it is present in the package insert or 0 otherwise. The numerator is the sum of the squares of the weights of the terms that are in common and the denominator is a normalization term that considers the total number of terms that \( A_i \) and \( B_i \) have. To be able to assess the significance of raw similarity scores and to integrate PPI similarity into our data fusion framework, the raw similarity scores were normalized to
This was achieved by generating a background set of PPI similarity scores from 1000 random drug pairs that had low 2D and low 3D similarity (\(p\)-values greater than or equal to 0.50). This \(p\)-value cut-off was selected because we have previously shown that structurally unrelated drug pairs very infrequently share targets [91]. Given a PPI similarity score \(S\) between a drug pair, the \(p\)-value is simply the number of occurrences of \(S\) or greater in the background set divided by the number of PPIs in the background set (1000). Figure 4.3 shows the raw PPI similarity between carbamazepine and levetiracetam along with the corresponding \(p\)-value plus the top 10 most heavily weighted terms between them. The raw similarity score between the pair was 0.286 and \(p\)-value was 0.044 which indicates significant shared clinical effects between the small molecules.

\[
PPI_{Similarity}(A, B) = \frac{\sum_{i=1}^{n} A_i \ast B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \ast \sqrt{\sum_{i=1}^{n} B_i^2}}
\]  

(4.1)

4.3.3 Target Prediction using Patient Package Insert Similarity

We have previously developed a general framework for data fusion which allows for the integration of similarity scores into a single value [91]. Briefly, given a molecule \(A\) and a set of molecules with a shared biological effect, \(B_n\), the similarity between molecule \(A\) and each molecule in \(B_i\) is computed. The similarity scores are normalized to \(p\)-values using a suitable normalization scheme such as assessing score significance against a random background set of scores. The multinomial distribution is then used to compute the likelihood, \(M\), of observing the set of \(p\)-values and of the converse probabilities, \(M^*\). The log-odds score \(L\) is then computed by taking the log of the ratio of \(M\) and \(M^*\) and inverting the sign. A detailed discussion of the computation and corresponding 2D and 3D similarity example can be found in the original publication [91].
Carbamazepine (371 terms)  
Levetiracetam (292 terms)  
PPI Sim: 0.286  
P-value: 0.044

<table>
<thead>
<tr>
<th>term (118 in common)</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>pancytopenia</td>
<td>6.64</td>
</tr>
<tr>
<td>cytochrome p450</td>
<td>6.57</td>
</tr>
<tr>
<td>primidone</td>
<td>6.54</td>
</tr>
<tr>
<td>albuminuria</td>
<td>6.52</td>
</tr>
<tr>
<td>grand mal</td>
<td>6.52</td>
</tr>
<tr>
<td>antiepileptic drugs</td>
<td>6.46</td>
</tr>
<tr>
<td>arthralgia</td>
<td>6.43</td>
</tr>
<tr>
<td>diplopia</td>
<td>6.42</td>
</tr>
<tr>
<td>hepatic failure</td>
<td>6.40</td>
</tr>
<tr>
<td>partial seizures</td>
<td>6.37</td>
</tr>
</tbody>
</table>

Figure 4.3: Example of Patient Package Insert (PPI) similarity between carbamazepine and levetiracetam. To quantify the clinical effects in common between a pair of drug term vectors, the cosine similarity (Equation 4.1) is used where 1.0 corresponds to having identical clinical terms and 0.0 to having no terms in common. Shown are the top 10 weighted terms between carbamazepine and levetiracetam. Highlighted in blue are terms parsed in Figure 4.2. The raw PPI similarity score between the pair is 0.286. To normalize similarity scores into \( p \)-values, a background similarity score distribution is derived from which significance can be computed. If the raw score is higher than 950/1000 of the background distribution, a significant \( p \)-value (0.05 or less) is obtained. In this case, the \( p \)-value was 0.044.
An attractive feature of our methodology is that it is able to integrate the results of different similarity computations into a single value. Figure 4.4 shows an example of a log-odds calculation for tolazamide interacting with the peroxisome proliferator-activated receptor gamma-retinoid X receptor heterodimer (PPARγ-RXR) using PPI similarity. The log-odds for this interaction resulting from the individual similarity metrics were 11.35 for PPI, 7.57 for 3D, and 5.49 for 2D. Combining the similarity methods gives a stronger prediction compared to using any single method alone with 3D+PPI log-odds = 18.37 and 3D+2D+PPI log-odds = 23.43.

4.3.4 3D Similarity and \( p \)-value Computation with Surflex-Sim

The Surflex-Sim 3D molecular similarity method and its use for virtual screening and off-target prediction has been extensively described in multiple publications [13, 14, 41, 42, 66]. Briefly, the method uses a molecular similarity function that computes, given two molecules in specific poses, a value from 0 to 1 that reflects the degree to which their molecular surfaces are congruent with respect to both shape and polarity. The function is computed based on the differences in distances from observer points surrounding the molecules to the closest points on their surfaces, including both the closest hydrophobic surface points and the closest polar surface points. So, two molecules that may have very different underlying chemical scaffolds may exhibit nearly identical surfaces to the observer points, which are intended to be analogous to a protein binding pocket, which also “observes” ligands from the outside. Additional details regarding the theory and underlying algorithmic details can be found in the previous chapter.

Previous versions of Surflex-Sim performed dynamic conformation search during the search process, which performs thorough exploration of the conformational
Figure 4.4: PPI log-odds calculation example. To test if tolazamide is predicted to interact with the PPARγ-RXR complex, the PPI $p$-values for tolazamide to nine known agonists were computed. Shown are the $p$-values of tolazamide with two agonists (nateglinide and repaglinide) along with the terms in common that are related to their therapeutic effects. The PPI log-odds score was 11.4 indicating that there is a high likelihood that tolazamide interacts with the target due to significant clinical effect similarity to known modulators. Also, shown are the 2D and 3D structural log-odds along with their combinations. In this case, combining structural and PPI similarity gives a log-odds score of 23.4 which is greater than that of any of the methods independently illustrating a synergistic effect when combining methods.
space of small molecules but comes with some computational cost. A new mode has been recently added to Surflex-Sim that allows for using pre-searched library molecules, and this has made similarity computations more easily scalable to applications such as the log-odds computations described here. The new command “search library” using default parameters generates up to 200 conformations for each molecule in a library, and it generates standard molecular imprints [13] that are used for similarity score normalization (described below). Briefly, given a molecule pair A and B that has been pre-searched, the conformations of molecule A are aligned to the conformations of molecule B, the same procedure is repeated in the opposite direction while maintaining a list of the similarity scores that are computed. Both the average bidirectional and maximum similarity score is reported. In the present work, the maximum similarity score is used. The scores produced range from 0 to 10. The bottom of Figure 4.1 shows the highest scoring alignment of the pre-searched conformations of the anticonvulsants carbamazepine and levetiracetam. The overlay shows excellent concordance with regards to overall surface shape and correspondence of charge, illustrated by the overlay of the carboxamide group. The maximum similarity score between these molecules was 8.8.

In prior work, normalized similarity scores were transformed to \( p \)-values empirically by assessing their significance against a background set of similarity scores. These were derived from a random set of 1000 screening molecules from the ZINC database [91]. The background similarity distributions computed in this fashion were observed to be nearly always well-approximated by a normal distribution. The previous chapter describes in detail a method for rapidly estimating the mean and standard deviation for a given molecule in such a way as to accurately ascertain \( p \)-values.
Figure 4.5: Relationship between empirical and estimated $p$-values and log-odds scores. The new normalization scheme is closely related to the empirical normalization. (A) The all-by-all pairwise similarities of over 300 drugs is shown (Pearson correlation between empirical and estimated $p$-values was 0.947, and Kendall’s Tau rank correlation was 0.814). (B) Log-odds computation of over 300 drugs against more than 40 pharmacologically relevant protein targets is shown (Pearson correlation of 0.955, and Kendall’s Tau rank correlation of 0.761).
Using the imprint based normalization, we recomputed the $p$-values and the log-odds from our previous work. Figure 4.5A shows the 2D histogram of estimated versus the empirical $p$-values of the all-by-all similarity of 358 drugs showing that they are highly correlated (R-squared of 0.947 and Kendall’s tau of 0.814). Figure 4.5B shows the 2D histogram of the estimated versus the empirical log-odds of 358 drugs against 44 targets showing that they are also highly correlated (R-squared of 0.955 and Kendall’s tau of 0.761). Taken together, the rapid similarity computation and the $p$-value estimation approach allow for typical 3D log-odds computations to be made in a few minutes for a given molecule against a target characterized by twenty known ligands.

4.3.5 GSIM-2D

The GSIM method was implemented as a strawman approach to identify relatively unchallenging ligand retrieval problems [14] and drug target interactions [91]. In many cases, the method is successful because there are many structural analogues present within many activity classes. Briefly, given two molecules, $A$ and $B$, identify all subgraphs at each heavy atom of molecule $A$ at depths of 1, 2, and 3. For every identified subgraph with 3 or more heavy atoms, check for its existence in molecule $B$. If the subgraph exists, increment a counter by the number of atoms in the subgraph or by 5 times the number of atoms in the subgraph if it is rooted at a non-carbon atom. The procedure is then repeated in the opposite direction by identifying subgraphs in molecule $B$ and checking for their existence in molecule $A$. The two scores are normalized to the interval [0,1] based on the maximum possible score in each direction. The ratio of the number of heavy atoms in molecule $A$ to molecule $B$ is computed along with the inverse ratio (molecule $B$ over molecule $A$). The overall similarity is then the minimum
of the two subgraph scores multiplied by the minimum of the heavy atom ratios. To yield high similarity, molecules A and B must be roughly the same size and contain similar subgraph composition, especially those rooted at heteroatoms.

4.4 Computational Procedures

Detailed scripts for generating the results presented here are available in the data archive associated with this paper. All computations were performed using default parameters. Briefly, a search library was generated for all of the drug structures: sf-sim search_library drug-list processed. The 3D molecular similarity computations using imprint based score normalization were performed as follows: sf-sim -lscreenopt simcover drug-list drug-list log-file-3d. Similarly, all 2D molecular similarity computations between drugs were made as follows: sf-sim gsimcover drug-list drug-list log-file-2d. To compute the 2D background distributions to normalize the similarity values, the analogous computations were done using the 1000 molecule ZINC decoy set: sf-sim gsimcover druglist zinc-list log-file-2d-norm.

4.5 Results and Discussion

4.5.1 Relationship between Structural Novelty and Clinical Effects

Previously, we quantified the effect of me-too drugs by showing that drug pairs with high 2D and high 3D similarity had four times more likelihood of having identical primary and secondary targets than drugs pairs that were structurally novel [91]. In the present study, we further quantified the effects of structurally similar drugs by comparing clinical effects. Clearly, because highly similar mol-
ecules often share target profiles, they also *should* share pharmacological effects, evidenced in their package insert information. Significant pairwise clinical effect *p*-values correspond to significant overlap in the respective patient package inserts. Conversely, insignificant *p*-values correspond to drug pairs sharing little in the way of pharmacologically relevant terms. Both to establish the relevance of the PPI similarity metric and to quantify the degree to which structural novelty is related to changes in clinical effects, we computed the pairwise 2D, 3D, and PPI similarity of all 602 drugs. The drug pairs were separated them into four categories based on chemical structural similarity: high 2D and 3D similarity, low 2D but high 3D, high 2D but low 3D, and low 2D and 3D. High similarity *p*-values were those less than or equal to 0.01 and low similarity *p*-values were those greater than or equal to 0.50.

Figure 4.6A shows the histogram of the PPI *p*-value distributions for each of the four structural categories. It is clear that the “me-too” drug distribution (red line) is different than the others, especially at significant PPI *p*-values. At a PPI *p*-value less than 0.01, a large fraction of the high 2D and high 3D similarity drug pairs had highly similar clinical effects. Drug pairs with high 3D similarity but with low 2D similarity (green line) exhibited a significantly smaller fraction with PPI *p*-values that were less than 0.01, indicating less similarity in clinical effects. The high 2D and low 3D pairs had fewer still with significant PPI *p*-values, and only a very small portion of structurally dissimilar drug pairs (low 2D and low 3D) shared clinically similar effects. Clearly, drug pairs with very high structural similarity (both by 2D and 3D methods) were much more likely to have closely shared clinical effects than molecule pairs of any other category, even those sharing high 3D similarity but low 2D similarity.

We also considered if drug pairs with high clinical effects similarity also had
Figure 4.6: Relationship between me-too drugs and clinical effects based on PPI similarity. (A) Drug pairs were segregated based on 2D and 3D similarity p-values into the four bins shown above (number of pairs per bin are shown in parentheses). The PPI p-value distributions illustrate that drug pairs that are structurally very similar (red line) have a higher likelihood of sharing clinical effects than drug pairs with diverse chemical structures (green line). Molecules that have high 2D/low 3D similarity (blue line) and low 2D/low 3D similarity (pink line) very infrequently have similar clinical effects. (B) Drug pairs were segregated based on PPI p-values into two bins, those that had high PPI similarity (p-values < 0.01) and those that had low PPI similarity (p-values ≥ 0.50). A high portion of drug pairs with high PPI similarity also have high 3D similarity (green line) but less 2D similarity (red line). As expected, drugs that have low PPI similarity, very infrequently exhibit high 3D (brown line) or 2D (blue line) similarity.
## Figure 4.7

(Upper right) Mepivacaine and ropivacaine are local anaesthetics first synthesized in 1957. The only structural difference is in the length of the carbon chain extending from the amine in the piperidine. The molecules have identical targets and, as expected, very similar clinical effects (PPI $p$-value $< 0.001$).

(Upper left) Felbamate and rufinamide are antiepileptics that are structurally dissimilar in 2D but have significant 3D similarity. The molecules share primary targets but differ in off-target effects; felbamate interacts with cardiac sodium channels and rufinamide does not, reflected in the PPI $p$-value of 0.130.

(Lower right) Neostigmine and pyridostigmine are cholinesterase inhibitors that have high 2D similarity, have identical targets, and similar clinical effects (PPI $p$-value = 0.007).

(Lower left) Mepivacaine and decitabine have different chemical structures, share no targets, and as expected, have no overlap in clinical effects (PPI $p$-value = 0.934).
high 3D or 2D similarity. The drug pairs were separated into two bins based on PPI p-values, those with high PPI similarity (p-value ≤ 0.01) and those with low PPI similarity (p-values ≥ 0.5). Figure 4.6B shows a histogram with the 3D and 2D p-value distributions of the two bins (four distributions in all). The 2D and 3D similarity p-value distributions for drug pairs with high PPI similarity (low PPI p-values) are shown as red and green lines. A large portion of these drug pairs had high 3D structural similarity. There was some enrichment for 2D similarity, but this was markedly less than that observed for the 3D approach. As expected, drug pairs that had low PPI similarity (blue and brown lines) also had low 3D and 2D structural similarity.

Figure 4.7 shows examples for each of the 2D and 3D structural similarity categories. Mepivacaine and ropivacaine (top right) are local anaesthetics that were part of a series of compounds first reported in 1957 [20]. The molecules are structurally similar, share identical annotated primary and secondary targets and have very similar pharmacokinetics [18]. Their PPI p-value was significant, indicating that these two molecules have very similar clinical effects, as expected for nearly identical chemical structures. In contrast, the antiepileptics felbamate and rufinamide are structurally different (top left). The drugs share the same annotated primary targets (neuronal voltage-gated sodium channels) but have different off-target profiles. Felbamate interacts with cardiac sodium channels, driving adverse effects. Cardiac-related terms such as “palpitation” and “tachycardia” are present in the felbamate package insert but are not seen for rufinamide. The PPI p-value of this drug pair was 0.13, reflecting some similarity in the desired anticonvulsant effects but also the dissimilarity in secondary target effects. Chemical scaffold diversity resulting in differences in clinical effects was a typical feature in these analyses. An example of a high 2D/low 3D structural similarity pair are the cholinesterase inhibitors neostigmine and pyridostigmine
These molecules are parasympathomimetics that have topologically similar structures, but where the location change of the charged nitrogen has a significant impact on their relative 3D similarity. These drugs share identical target annotations, have no substantive differences in clinical effects, and thus have a significant PPI p-value. An example of a low 2D/low 3D similarity drug pair is mepivacaine and decitabine. These molecules are structurally very different, have no target overlap, and an insignificant PPI p-value.

4.5.2 Internal SPDB Validation: On and Off Target Effects

An attractive aspect of the log-odds framework is that it allows us to combine different types of similarity computations into a single value. For each of the 602 drugs in our dataset, we computed the 2D, 3D, PPI, and combination log-odds scores of interacting with each of the 91 targets that had at least 5 drugs as ligands in the SPDB (see Table 4.1 and 4.2). In each case, any self/self comparisons were omitted from the calculations, making this exercise a leave-one-out cross-validation of the log-odds predictive methodology. The three methods were used independently and in combination to predict the log-odds of known primary and secondary target interactions. For known primary and secondary interactions, the three methods produced scores exceeding a threshold of 0.089% of the time or greater. Significant differences in performance between the methods were seen at higher log-odds thresholds. For primary target prediction, 2D similarity performed almost as well as 3D, which is expected since many drugs are designed as minor derivatives of pre-existing drugs in the same class. The PPI metric performed as well as the structural similarity methods for primary target prediction. This supports our observations that drugs are often designed as minor variants of others in the same class, and that drugs with highly similar structures will
Figure 4.8: Proportion of drug target annotations correctly predicted for primary and secondary targets using 3D (green line), 2D (red line), and PPI (blue line) similarity. (A) For primary target prediction there is only a slight benefit in using 3D over 2D, using PPI alone is more predictive than 2D. (B) For secondary target prediction there is a significant benefit in using 3D over 2D. However, using clinical effects (PPI) alone yields a greater portion correctly predicted off-targets than 2D.

have highly similar pharmacological profiles. Drug packet inserts are generally very explicit with respect to primary indication descriptions but tend to include varying amounts of information with respect to side effects especially when the effects were determined post-market. Thus, the standardization of medical terms reported for the primary drug indication within a class of drugs is reflected in the high accuracy of our PPI metric on predicting primary targets.

In contrast, for secondary target prediction, 3D similarity performed significantly better than 2D, again as expected because off-target effects are often due to drugs with chemical scaffolds that are highly-divergent from drugs designed for
a target. These performance trends for 2D and 3D similarity on the 602 drug/91 target set were analogous to those observed in our previous work with a 358 drug and 44 target set [91]. The PPI metric performed less well than 3D similarity but better than 2D similarity within this set of secondary target predictions. Weaker performance relative to the 3D approach reflects the non-standardization of off-target effects reported in package inserts, but performance in excess of the 2D chemical similarity computation was surprising.

Finally, as a surrogate for a measurement of false positive rates for our similarity methods, we determined the number of drug/target predictions for interactions that were unannotated. In our SPDB, a missing annotation between a drug and a target does not mean that the interaction does not occur. Authentic interactions within our 602 drug/91 target set may have been published after our curation or have yet to be biochemically characterized. At log-odds thresholds of 5, 10, and 20, predictions for non-existent SPDB annotations for both 3D similarity alone and 3D+2D+PPI were 3%, 1%, and 0.2%. These are upper limits of false positive predictions. As will be described below, the false positive rate was actually lower since many of the new predictions were validated as true by incorporating annotations from the ChEMBL database.

Figure 4.8 shows the log-odds distributions for primary and secondary targets using the individual methods and Figure 4.9 shows the results of combining the methods for secondary target prediction. For completeness, Table 4.1 and Table 4.2 shows the percentages correctly predicted at various log-odds for all of the methods for both the primary and secondary targets.

For primary target prediction, there was only a slight benefit in using 3D over 2D, which we have previously shown on a smaller set of pharmacologically relevant targets; however, PPI was also slightly better than 2D. At a stringent
Figure 4.9: Proportion of SPDB secondary drug target annotations correctly predicted using 3D (green line), 3D+PPI (aqua line), and 3D+2D+PPI (orange line). Combining orthogonal similarity methods leads to a benefit in predicting off-targets. At a log-odds of 10 and 20, the proportion predicted increased by more than 25%, at even more stringent log-odds thresholds the trend continues.
Table 4.1: Proportions correctly predicted for primary and secondary targets at various log-odds thresholds using different similarity methods.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>log-odds</th>
<th>3D</th>
<th>2D</th>
<th>PPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPDB PRIMARY</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>97</td>
<td>89</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>14</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>SPDB SECONDARY</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>97</td>
<td>90</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>43</td>
<td>7</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

log-odds threshold of 10, 3D+PPI was 10 percentage points more predictive than 2D. The few cases where the log-odds were 35 or greater for 2D and PPI similarity were due to antibiotics that inhibit D-alanyl-D-alanine carboxypeptidase. These molecules have very high 2D similarity (i.e. oxacillin and carbenicillin) and have similar pharmacological effects. Combining the methods has a clear advantage, increasing predictability to greater than 25% at a threshold of 10 and 20 (table 4.2).

As for secondary targets, consistent with previous observations, 3D similarity had a clear advantage over 2D for predicting off-targets [91]. When a small molecule is not designed toward a specific target but unintentionally interacts with it, 2D similarity to other ligands tends to be significantly diminished. But considering molecular surface congruence, which is scaffold independent, supports detection of these unexpected and important relationships. At threshold of 10, 3D similarity correctly predicted 47%, but 2D only predicted 7% of these relationships. The PPI metric recovered twice as many as 2D (14%). Combining
Table 4.2: Proportions correctly predicted for primary and secondary targets at various log-odds thresholds using different combinations of similarity methods. Combining all of the methods (3D+2D+PPI) is more predictive than any method alone.

The methods can potentially lead to more confident predictions, though this does not always occur. Log-odds scores are not additive; including more observations (p-values) within the computation greatly affects the combinatorics. Consider a collection of p-values that gives a minimally positive log-odds score. If those p-values are combined with a set already producing a high score, the magnitude of the score may be reduced. One can identify cases of synergy where the combination of multiple types of comparisons improves upon use if single-mode calculations. Figure 4.9 shows the results when 3D is combined with PPI (aqua line) and with 2D and PPI (orange line). It is evident that combining 2D, 3D, and PPI similarity calculations yields a greater percentage of off-targets. At a log-odds threshold of 10, the proportion correctly predicted was 64% (21 points better than 3D alone). This effect is mostly due to the added benefit of the PPI information. At a more stringent log-odds of 20, combining all of the methods
correctly predicted 38% of the secondary targets. Neither PPI or 2D can independently make such inferences that confidently, but the methods are contributing to the 3D similarity computations in a positive manner when combined.

Examples of specific secondary target predictions are shown in Figures 4.10 and 4.11. Figure 4.10 shows an example where 3D similarity alone identifies important off-target effect. Mesoridazine is an antipsychotic whose therapeutic benefit is due to interaction with dopamine receptors. However, mesoridazine has serious cardiac effects, attributed to interacting with hERG, and it was ultimately removed from the market [81]. Computing the similarity to 25 other hERG modulators, the log-odds scores were 1.1, 12.3, and 2.3 using 2D, 3D, and PPI, respectively. Combining the methods gave slightly lower log-odds compared to the 3D. Shown are two molecules that interact with hERG, desipramine and clomiphene, along with their 3D overlays and five clinical terms that they have in common with mesoridazine. The 3D similarity of both molecules to mesoridazine was high, indicated by significant p-values (0.017 and 0.004). However, 2D similarity was lower, having p-values of 0.085 and 0.895. PPI similarity yielded p-values of 0.089 and 0.619. It can be difficult to interpret the top two weighted terms in common because clinical effects related to the protein target in question need not be the most heavily weighted. Here, we are considering similarity based on a complete set of clinical effects. In this case, clinical terms that are hERG-related do appear. For example “ventricular tachycardia” and “hypotension” partly drive the similarity calculation.

Figure 4.11 is an example where synergy was obtained among the methods, allowing for a more confident off-target prediction. Ropivacaine is a local anesthetic which achieves its therapeutic benefit by interacting with peripheral NaV1.8 (SCN10A) and NaV1.9 (SCN11A) sodium channels but has serious cardiotoxicity.
Figure 4.10: SPDB Example 1, 3D similarity alone can predict important off-target effects. Mesoridazine is an antipsychotic (dopamine receptor antagonist) that causes serious cardiac side effects which are attributed to interacting with HERG. Shown are the 2D structures, 3D overlays, and clinical terms in common of mesoridazine with two HERG modulators, desipramine and clomiphene. When combined with 23 other HERG modulators, mesoridazine is predicted to be a HERG modulator with log-odds of 1.1, 12.3, and 2.3 using 2D, 3D, and PPI. Combining the methods does not strengthen the prediction since the 3D+PPI log-odds is 11.1 and 2D+3D+PPI is 10.0.
Figure 4.11: SPDB Example 2 - Combining similarity methods can make off-target predictions stronger. Ropivacaine (Naropin) is a local anesthetic (Nav1.8 and Nav1.9 sodium channel antagonist) with unwanted cardiac effects which are attributed to interacting with cardiac Na\textsubscript{v}1.5 sodium channels. Shown are the 2D structures, 3D overlays, and clinical terms in common of ropivacaine with two cardiac sodium channel modulators, flecainide and fosphenytoin. When combined with 23 other Na\textsubscript{v}1.5 modulators, ropivacaine is predicted to be a modulator with log-odds of 2.3, 5.9, and 3.6 for 2D, 3D, and PPI. Combining the methods strengthens the prediction since the 3D+PPI log-odds is 7.9 and 2D+3D+PPI is 9.0.
due to interacting with Na\(_v\)1.5 sodium channels (SCN5A) expressed in the heart. By computing the similarity to 25 other Na\(_v\)1.5 sodium channel modulators, the log-odds were 2.3, 5.9, and 3.6 using 2D, 3D, and PPI, respectively. Combining the methods gave higher log-odds: 9.0 when combining all of the methods. Shown are two molecules that interact with the sodium channel, flecainide and fosphenytoin. The 3D and PPI similarity between ropivacaine and both of the molecules was more significant than the 2D similarity. An example of clinical effects that were in common and clearly related to SCN5A interaction are “heart block” and “cardiac arrest.” The synergistic effect of the methods gave increased confidence in the prediction in part because of the greater magnitude of the score. In addition, because the sources of information are orthogonal (e.g. structural comparisons and quantification of clinical effects), further confidence is provided.

### 4.5.3 Prediction of New Drug-Target Pairs within ChEMBL

As discussed above, a missing annotation within the SPDB between a drug and a target does not necessarily mean that the interaction does not occur. For example, both orphenadrine and mesoridazine had high 3D log-odds against the muscarinic receptor but the interactions were unannotated in the SPDB. Careful

<table>
<thead>
<tr>
<th>Data Set</th>
<th>log-odds</th>
<th>3D</th>
<th>2D</th>
<th>PPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChEMBL</td>
<td>0</td>
<td>87</td>
<td>61</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>43</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>16</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.3: Proportions correctly predicted for new drug target interactions from ChEMBL at various log-odds thresholds using different similarity methods.
Table 4.4: Proportions correctly predicted for new drug target interactions from ChEMBL at various log-odds thresholds using different combinations of similarity methods. Combining all of the methods (3D+2D+PPI) is more predictive than any method alone.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>log-odds</th>
<th>3D+2D</th>
<th>3D+PPI</th>
<th>2D+PPI</th>
<th>3D+2D+PPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChEMBL</td>
<td>0</td>
<td>80</td>
<td>85</td>
<td>68</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>42</td>
<td>41</td>
<td>19</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>20</td>
<td>18</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 4.12: Proportion of ChEMBL drug target annotations correctly predicted using 3D (green line), 2D (red line), and 3D+2D+PPI (orange line) similarity. Challenging off-target interactions are difficult for 2D (red line) to predict but 3D predicts significant portion and combining the methods gives a slight benefit.
inspection of the literature revealed that the drugs were known to antagonize muscarinic receptors [91]. Therefore, drug target annotations that are known but missing from our SPDB can serve as a blind set to test our methodology. To supplement annotations within the SPDB with a blind set for methodological testing, we searched ChEMBL and found 380 biochemically characterized drug/target interactions not present in the SPDB. We then investigated how well the methodology could identify the new ChEMBL annotations based only upon information within in the SPDB as the basis to compute the log-odds.

Figure 4.12 shows the log-odds distributions of the 3D, 2D, and 2D+3D+PPI for the new ChEMBL annotations and Tables 4.3 and 4.4 show the proportions correctly predicted at various log-odds using different methods and combinations. Among individual methods, 3D similarity strongly outperformed 2D- or PPI-based similarity, with the latter two having similar performance. However, the combination of the three methods, overall, yielded better performance than 3D alone. At log-odds thresholds of 10 and 20, using the full combination of methods, the percentage of recovered annotations within the SPDB test set was 22% and 4%, respectively. This compared with 16% and 2% using 3D similarity alone, and 3% and 0% using either 2D or PPI similarity alone. The enrichment ratios for the combination approach, using the upper-bound false positive rates discussed above, corresponded to 22-fold and 40-fold, respectively, at log-odds thresholds of 10 and 20.

Figure 4.13 shows a typical example of a drug/target interaction not annotated in the SPDB where the combination similarity approach confidently identified a pharmacologically relevant target. Sibutramine is an anorexic annotated in the SPDB as a ligand of the serotonin and norepinephrine reuptake transporters. However, it has been shown that sibutramine also interacts with the dopamine...
Figure 4.13: ChEMBL example showing that combination similarity effectively predicts a drug target interaction not covered within the SPDB. Sibutramine is an anorexic (5-HT and norepinephrine reuptake inhibitor) with a ChEMBL annotation as a dopamine reuptake transporter ligand. Shown are the 2D structures, 3D overlays, and common clinical terms between sibutramine and two dopamine reuptake transporter inhibitors, bupropion and nefazodone. When tested with 9 other dopamine reuptake inhibitors, the log-odds were 2.3, 4.2, and 6.9 using 2D, 3D, and PPI. Combining the methods strengthened the prediction (log-odds for 2D+3D+PPI = 9.4).
reuptake transporter and that this interaction contributes to the therapeutic benefit [3]. Computing the similarity between sibutramine and 11 other dopamine reuptake transporter inhibitors, the log-odds were 2.3, 4.2, and 6.9 using 2D, 3D, and PPI, respectively. These predictions were strengthened by combining all three methods, with corresponding log-odds of 9.4. The pairwise PPI similarities between sibutramine and the two dopamine reuptake inhibitors, bupropion and nefazodone, are extremely significant as are the individual 3D similarities. The example illustrates that clinical effects are sometimes sufficient to infer off-targets and that combining similarity methods can make predictions stronger.

4.5.3.1 Structural Novelty of Test ChEMBL Drug-Target Pairs

The ChEMBL drug/target annotations not present in the SPDB were often for drugs that were structurally novel compared with those sharing the target within the SPDB. To quantify structural novelty, we performed a nearest neighbor analysis where, for each drug, the most similar 2D representative from the SPDB was identified (based on \( p \)-value) from within the collection of drugs having the correct target annotation. An analogous computation was performed for each drug target annotation in the SPDB, using a leave-one-out methodology. Figure 4.14 shows a histogram of the distributions of \( p \)-values for the ChEMBL set (red line) and the SPDB (green line). Within the SPDB leave-one-out set, there were substantially more cases with extremely low \( p \)-values than for ChEMBL, illustrating that the nearest structural neighbor for the ChEMBL test molecules was generally more distant. Two examples are highlighted where the nearest neighbor had relatively poor 2D \( p \)-values but much more significant 3D \( p \)-values that provided support for high log-odds scores.

Fluoxetine (blue box) is a selective serotonin reuptake inhibitor which achieves
Figure 4.14: Novelty of ChEMBL drug target annotations. The histogram shows the results of a nearest neighbor analysis for each drug target pair in ChEMBL and SPDB set. At significant 2D $p$-values (less than 0.05) there was a smaller proportion of significant $p$-values in the ChEMBL set than in the SPDB. Fluoxetine (blue box) is a selective serotonin reuptake inhibitor, annotated within ChEMBL against the muscarinic M receptor, and its 2D nearest-neighbor was methadone. Apomorphine (red box) is an antagonist of the dopamine D$_2$ receptors, annotated in ChEMBL with activity against the dopamine D$_2$ receptor. Its 2D nearest-neighbor was structurally distinct: ropinirole.
its therapeutic benefit by inhibiting the 5-HT reuptake transporter. However, the ChEMBL data indicated that fluoxetine also interacts with the muscarinic M₃ receptor. The nearest-neighbor molecule from the SPDB sharing this annotation was methadone (2D $p$-value = 0.041). Considering all of the muscarinic M₃ receptor ligands (38 total), the 2D, 3D, and PPI log-odds were 1.2, 7.7, and 3.7 respectively. Combining all of the methods gave a score of 8.2. Because of the topological structural novelty compared with the set of known muscarinic ligands, the contribution of the 2D comparisons slightly reduced the magnitude of the combination score (8.2) compared to that using only 3D and PPI comparisons (9.8).

Apomorphine (red box) is indicated to treat Parkinson’s disease and its therapeutic benefit is thought to be primarily due to activating dopamine D₂ receptors. However, apomorphine was indicated within ChEMBL to also interact with the dopamine D₃ receptor (which is also known to play a role in the beneficial effects for other anti-Parkinsonian drugs). The SPDB nearest-neighbor drug was ropinirole (2D $p$-value = 0.210), which can be seen to be structurally distinct in a topological sense in Figure 4.14. As in the previous case, when considering all 11 dopamine D₃ ligands, the 3D comparisons provide primary support for a positive log-odds score. The 2D, 3D, and PPI log-dds were -1.1, 7.7, and 1.2 respectively. The combination of all three comparison types yielded a score of 3.3, but considering only 3D and PPI comparisons, the score was 5.5. In this case, the 3D molecular similarity information was the only reliable predictor.

### 4.5.4 Predicting Novel Drug/Target Interactions

The foregoing leave-one-out experiments within the SPDB and the blind test on new drug-target annotations from ChEMBL has established predictive value
in the log-odds approach, especially when combining information derived from 3D, 2D, and PPI molecular comparisons. We have treated the “empty cells,” cases where no drug-target linkage has been established, as a surrogate for false positives. Of course, many of the empty cells within the SPDB existed within ChEMBL as \textit{bona fide} drug-target interactions. Therefore, we expect that a significant number of the cases where no interaction is known, but where a high predicted combination log-odds score exists, may actually represent genuine drug-target activities. To prioritize consideration of predicted new interactions, the 3D+2D+PPI log-odds score was used to rank predictions, and drug-target pairs for which the combination score was higher than any individual method were retained, indicating synergy among the information sources.

The single highest-scoring prediction was cyclobenzaprine, with a combination log-odds score of 36.6 for the 5HT reuptake transporter. This prediction, however, reflects design ancestry more than representing a surprising off-target effect. The log-odds of cyclobenzaprine against the 5-HT reuptake transporter were 10.4, 18.3, and 13.0 using 2D, 3D, and PPI, respectively, reflecting very high similarity across all modes of comparison. Cyclobenzaprine is primarily prescribed as a muscle relaxant, its therapeutic benefit is due to antagonizing the 5-HT$_{2A}$ and 5-HT$_{2C}$ receptors. The molecule is structurally similar to tricyclic antidepressants (e.g. imipramine and amitriptyline) and there have been reported cases where combining cyclobenzaprine with antidepressants leads to serotonin syndrome [46, 17]. The same assertion was made by Mestres et al. [61] in 2011 and was experimentally characterized, showing that cyclobenzaprine is an inhibitor of the 5-HT reuptake transporter (IC$_{50}$ of 108 nM). When considering the history of cyclobenzaprine, this effect is perhaps not very surprising. Cyclobenzaprine was originally synthesized in 1961 (originally known as proheptatriene), based upon the structure of promazine, replacing the sulfur and nitrogen atoms in the
phenothiazine with carbons [86]. Similar chemical modifications resulted in the synthesis of imipramine (a groundbreaking antidepressant) which was later discovered to achieve its effect by inhibiting the 5-HT reuptake transporter [91]. In the 1960's, cyclobenzaprine was clinically tested in comparison with imipramine for treatment of depression, and it was shown to have similar pharmacological effects [28]. It was not until 1997 that cyclobenzaprine approved as a muscle relaxant.

In order to avoid cases like this, where design ancestry drove a prediction, we eliminated cases from further consideration where the known primary or secondary targets of a drug were aminergic GPCRs, owing to the established promiscuity of such drugs among numerous targets [47, 48]. We also eliminated cases where the nominal predicted target was so closely related to one of the established targets as to be unsurprising (e.g. nucleoside analogs targeting a particular reverse transcriptase predicted to target another reverse transcriptase). After elimination of such predictions, the following examples were the top two highest ranked.

The first example illustrates an inference regarding a therapeutic benefit. Levetiracetam is an anticonvulsant that is believed to have a unique mechanism of action when compared with most existing anticonvulsants. The CNS targets of the major classes of anticonvulsant drugs have been known for many years. For example, barbiturates such as pentobarbital are known to act through the GABA$_A$ receptor, and other drugs such as carbamazepine and phenytoin act through neuronal voltage-gated sodium channels. More recent experiments have shown that these drugs also modulate voltage-gated potassium channels and suggest that this contributes to their anti-epileptic effects [94, 5, 68, 4]. Levetiracetam is a newer anticonvulsant with a chemical scaffold significantly different from the aforemen-
tioned drugs in this class. Levetiracetam has been proposed to work through a novel mechanism of action due to high binding affinity to the synaptic vesicle protein SV2A, which is not a known therapeutic target of any drug including antiepileptics [59, 58, 82]. Our methods strongly predict that levetiracetam is a voltage-gated sodium channel modulator with 3D log-odds alone of 14.5. Figure 4.15 shows the pairwise 3D similarity scores and overlays for levetiracetam with carbamazepine and pentobarbital. A recent study, in 2009, showed that levetiracetam inhibited voltage-gated potassium currents, leading to the suggestion that this drug, like other anti-epileptics, acts at least in part through potassium channels. Considering that many antiepileptics modulate both sodium and potassium channels, our prediction supports the notion that levetiracetam shares a similar mechanism of action, perhaps in addition to the interaction with SV2A. Note that since the levetiracetam structure is so different from the other anticonvulsants, the pairwise 2D similarity was almost non-existent and thus 2D similarity was low (log-odds of 0.5). The PPI log-odds of 9.1 and the combined 3D+PPI log-odds of 21.8 were responsible for the strength of this prediction. Although our approach predicted sodium channels as the primary target of levetiracetam, by association, the results also suggest this drug acts through potassium channels as do the other drugs in our set to which it has high 3D similarity.

The second example is shown in Figure 4.16, illustrating the case of selegiline, a monoamine oxidase B (MAO-B) inhibitor originally approved in oral form as adjunctive treatment for Parkinson’s disease. At higher doses, selegiline also inhibits another MAO isozyme (MAO-A) which is understood to be a therapeutic target for treating depression [55]. However, MAO-A is also involved in first pass metabolism of biogenic amines such as dietary tyramine, which sometimes lead to unwanted hypertensive effects (sometimes called the “cheese reaction” owing to high levels of tyramine in many cheeses). Due to drug-food interactions, the ef-
Anticonvulsant Drugs

carbamazepine  

<table>
<thead>
<tr>
<th>term (118 in common)</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>grand mal</td>
<td>6.52</td>
</tr>
<tr>
<td>antiepileptic drugs</td>
<td>6.46</td>
</tr>
<tr>
<td>partial seizures</td>
<td>6.37</td>
</tr>
<tr>
<td>epilepsy</td>
<td>4.87</td>
</tr>
<tr>
<td>seizures</td>
<td>4.74</td>
</tr>
</tbody>
</table>

pentobarbital  

<table>
<thead>
<tr>
<th>term (87 in common)</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>antiepileptic drugs</td>
<td>6.46</td>
</tr>
<tr>
<td>status epilepticus</td>
<td>6.33</td>
</tr>
<tr>
<td>somnolence</td>
<td>6.21</td>
</tr>
<tr>
<td>ataxia</td>
<td>5.60</td>
</tr>
<tr>
<td>seizures</td>
<td>4.74</td>
</tr>
</tbody>
</table>


term (118 in common) weight 
grand mal 6.52 
antiepileptic drugs 6.46 
partial seizures 6.37 
epilepsy 4.87 
seizures 4.74

Figure 4.15: Predicted interaction between levetiracetam and voltage-gated potassium channels. Levetiracetam is an anticonvulsant targeting synaptic vesicle glycoprotein 2A (SV2A). Shown are the 2D structures, overlays, and clinical effects in common with two anticonvulsants, carbamazepine and pentobarbital. Combining molecular and clinical effects similarity of known sodium channel modulators gave a log-odds score of 21.4.
fects of orally administered selegiline (and other MAO inhibitors) can be difficult to manage, limiting the utility of such drugs. In 2006, a transdermal formulation of selegiline was approved to treat major depressive disorder (the transdermal administration substantially reduces dietary restrictions by bypassing first-pass liver involvement). While the relationship between MAO inhibition and therapeutic treatment of depression is established, other clear targets include amine reuptake transporters. Examples include selective serotonin reuptake inhibitors (SSRIs, e.g. escitalopram) and serotonin norepinephrine reuptake inhibitors (SNRIs, e.g. duloxetine).

The combination log-odds scores of selegiline against the 5-HT reuptake transporter (SERT) and the norepinephrine reuptake transporter (NET) were 30 and 38, respectively (see Figure 4.16 for the breakdown of individual scores). We postulate that part of the beneficial anti-depressive effects of selegiline is mediated through modulation of these transporters. Careful review of the literature revealed that selegiline had been previously tested in rat brain synaptosomes for reuptake inhibition of biogenic amines. The corresponding IC₅₀ for reuptake inhibition of serotonin and norepinephrine were 460µM and 26µM, respectively [53]. Selegiline is not a potent inhibitor of MAO-A, with in vitro IC₅₀ of 67µM [75]. The amine reuptake effects occurred at lower concentrations for the norepinephrine transporter. It seems likely that the therapeutic benefit of selegiline is not only due to inhibiting the deamination of serotonin, dopamine, and norepinephrine by MAO-A but also by blocking norepinephrine and/or 5-HT reuptake directly, as with many other major depression drugs. Further biochemical experiments would more fully characterize the pharmacological effects of these interactions and confirm direct effects of unmodified selegiline on the two predicted transporters.
Figure 4.16: Predicted interaction between selegiline and the 5-HT and norepinephrine reuptake transporters. Shown are the 2D structures, 3D overlays, and clinical effects in common with two antidepressants, clomipramine and sertraline.
4.6 Conclusion

In the present study, we report a means to combine chemical similarity between molecules with information derived from computing similarity based upon lexical analysis of patient package inserts (PPI). As expected based on our prior work, drugs that were highly structurally similar (both by 2D and 3D comparison) were much more likely to have significant overlap of their clinical effects compared to drugs that were structurally different (low 2D similarity but high 3D similarity). Our prior work illustrated a similar effect with respect to specifically annotated molecular targets: me-too drugs tend to have nearly identical target profiles [91]. The correlation between lexical and chemical similarity also served to validate the lexical comparison methodology.

We extended a probabilistic data fusion method to include observations from both molecular and clinical effects similarity and reported performance on predicting protein targets of small molecules. This was done both by leave-one-out cross-validation on our internal database of drug-target interactions (the SPDB) as well as a blind test on new interactions present in ChEMBL. For off-target prediction within the SPDB, 3D similarity was the most effective single information source. However, combining the methods predicted a larger proportion of secondary targets than any of the individual methods, while maintaining a similar nominal false positive rate. On the test against previously unseen ChEMBL drug-target linkages, again 3D similarity was the single most effective predictor, but gains were derived from combining the different data sources.

We also illustrated three predictions that were not annotated in the SPDB or ChEMBL. The single most confident such prediction was recently confirmed in the literature, but careful study revealed that the predicted off-target effect had been contemplated many years ago. However, in two other cases, more striking
predictions of activity against specific targets were made, in both cases helping to explain therapeutically beneficial effects. These effects have strong support in the literature but further direct characterization is warranted.
CHAPTER 5

Conclusion

We have reported a new method for combining information from molecular computations in order to effectively make inferences based on the known activity of sets of molecules. The approach is general and can be applied to any computational method, offering a unified means to fuse the output from multiple sources to produce a single log-odds score. Two aspects are particularly important: (1) mapping of raw computed values to $p$-values in a context-specific way because raw scores have different significance depending on the complexity of the molecules being considered, and (2) offering a means of data fusion that balances evidence in favor of an assertion with evidence against the assertion while avoiding degeneracies that arise from literal interpretation of empirically estimated $p$-values.

By comprehensively applying the method to a large set of drugs with overlapping pharmacological effects, we were able to identify differences in the predictive ability of 2D similarity methods compared with 3D ones. In assessing the predictive value of the approach, the most comprehensive analyses were done considering recovery of known primary and secondary targets. Particularly for the prediction of unanticipated off-target effects, 3D performed much better than 2D, although the combination of both was beneficial. Comprehensive analysis of false positive predictions was impossible due to the lack of systematic profiling of drugs against large panels of targets. However, using the well-known mus-
carinic side effects as a surrogate for direct muscarinic modulation, we assessed the behavior of drugs lacking explicit annotation. Within this set, we established excellent separation of drugs with muscarinic side-effects from those with none apparent. Alternative approaches such as belief theory and SEA performed less well.

We also considered the relationship between chemical structural novelty and pharmacological novelty. The key finding was that for a me-too drug pair (exhibiting both high 2D and 3D similarity), the chances were essentially even of observing modulation of identical sets of biological targets compared with non-identical sets. However, in the case where the drugs show high 3D similarity but show differences at the topological level, these odds shifted to roughly three-to-one in favor of observing novel effects at the gross level of target modulation. Clearly, even very subtle changes in chemical structure can yield sufficient differences in potency, selectivity, or ADME/toxicity characteristics to make for novelty in pharmacological action that can provide benefits to patient populations over existing therapies. This is especially true for therapeutics that target rapidly evolving pathogens, where minor structural modifications can overcome resistance. But it is clear that introduction of a novel scaffold brings significantly higher risk and potentially higher reward in terms of novelty that might be beneficial for patients.

We developed a means to compute similarity between drugs based upon English descriptions of their clinical pharmacology, as represented in the patient package inserts of approved drugs. This built upon standard methods for lexical comparative analysis, but it also made use of modern medical terminology dictionaries as well as comprehensive data on word and phrase usage from the World Wide Web. We have described this a patient package insert (PPI) sim-
ilarity method and combined it with the data fusion framework. Application was to the majority of approved of drugs, with demonstrations of predictive performance being done both using leave-one-out experiments and a blind test on data derived from the ChEMBL data curation effort. In assessing the predictive value of the approach, comprehensive analyses were done considering recovery of known primary and secondary targets. For the prediction of unanticipated off-target effects, inferences driven by 3D similarity were much more valuable than 2D- or PPI-based approaches individually. However, the combination of all three was beneficial. Based upon clear enrichment for *bona fide* true positives over nominal false positives, we applied the combined approach to identify as yet uncharacterized drug-target interactions. While additional characterization of the predictions will be an ongoing effort, particularly using direct biochemical tests, the highest scoring examples identified had strong literature support.

There are three primary contributions of this research effort. First is the development of a simple and scalable means to combine multiple data sources in order to make surprising predictions about potential biological activities of drugs. The second has been a careful analysis on large corpora of data, elucidating the reasons behind predictive success and failure of different methods and combinations thereof. In particular, we have shown that reasoning solely based upon 2D topological computations of molecular similarity often has the roots of its success in the design ancestry of molecules, suggesting that validation strategies for such predictive methodologies should be structured *temporally*, making predictions on molecules that were made after those on which inferences are being made. Third, the successful demonstration of the synergistic combination of different methods, including those that detect phenotypic effect similarities, offers some hope that the disparate data sources that impinge upon drug discovery in the pre-clinical research phases can be exploited to make better decisions.
REFERENCES


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

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

[Signature]
Author Signature

[Date]
6-7-13
Date