J Comput Aided Mol Des (2008) 22:147–159
DOI 10.1007/s10822-007-9150-y

Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery

Ann E. Cleves · Ajay N. Jain

Received: 22 October 2007 / Accepted: 12 November 2007 / Published online: 12 December 2007
© Springer Science+Business Media B.V. 2007

Abstract Inductive bias is the set of assumptions that a person or procedure makes in making a prediction based on data. Different methods for ligand-based predictive modeling have different inductive biases, with a particularly sharp contrast between 2D and 3D similarity methods. A unique aspect of ligand design is that the data that exist to test methodology have been largely man-made, and that this process of design involves prediction. By analyzing the molecular similarities of known drugs, we show that the inductive bias of the historic drug discovery process has a very strong 2D bias. In studying the performance of ligand-based modeling methods, it is critical to account for this issue in dataset preparation, use of computational controls, and in the interpretation of results. We propose specific strategies to explicitly address the problems posed by inductive bias considerations.

Keywords Inductive bias · Ligand-based modeling · Computational evaluation · Molecular similarity · Surflex-Sim

Introduction

Inductive bias is defined informally as the set of assumptions made in order for a person or procedure to make a prediction based on some data. The classic example is that of a naturalist seeing a swan for the first time. Observing that the swan is white, it is not *logically* true that the next observed swan would be white or that all swans are white, but absent other information it is a reasonable guess. For this to be a logical conclusion, one needs the *assumption* that all swans are the same color. The formal definition of inductive bias in machine learning is the assumptions that must be added to the observed data in order to transform the algorithm's outputs into logical deductions [1].

Computational procedures for predicting biological activity based on the structures of some known ligands have different inductive biases. Consider a 2D method based on molecular graph topology. The inductive bias is that a new molecule will be active if, essentially, it can be transformed into a known active molecule by a small number of operations of atom element changes, bond order changes, and the insertion or deletion of atoms and bonds. A 3D pharmacophoric method has a different bias, assuming that if a new molecule has a non-strained pose in which a small number of 3D atom-based points are congruent with those common to a set of active ligands, it will be active. Surface and volume-based approaches have still different biases.

There is a special property within the domain of ligand activity prediction that makes it unique. The active ligands that we have for testing methods, in nearly all cases, represent the specific *design choice* of a human being. We do not have collections of ligands for targets that arose through a natural chemical evolution process that reasonably sampled the space of synthesizable small molecule ligands. We instead have collections of ligands that were the result of focused effort on the part of synthetic chemists to make ligands of targets by design based on knowledge of already existing ligands. The chemists are acting as *predictors* of activity. We assert here, and will show later quite

A. E. Cleves
BioPharmics LLC, 36 Avila Road, San Mateo, CA 94402, USA

A. N. Jain (✉)
University of California, San Francisco, Box 0128,
San Francisco, CA 94143-0128, USA
e-mail: ajain@jainlab.org

directly, that the inductive bias of chemists is closely related to the inductive bias of 2D molecular similarity methods.

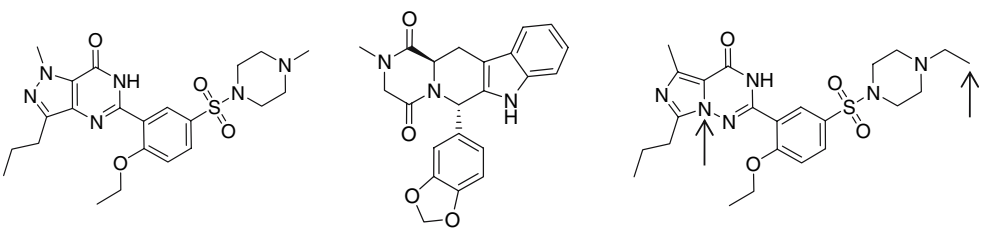
Why is this a problem? It presents a problem for methodological evaluation because the inductive bias of the *generators* of the data matches the bias of a *subset* of the methods. Thus, the methods sharing the bias will have a relative advantage in measured performance comparisons over those that do not, but the advantage may be purely artifactual. Imagine we could identify all small ligands that bound a particular binding site from the space of all synthesizable ligands. Suppose that for some targets these ligands turned out all to be trivial analogs of *one central structure*, but for other targets, the ligands can only be characterized as trivial analogs of *multiple very different-looking structures*. If it were true that the vast majority of targets had the former character, then the assumptions underlying the 2D inductive bias would lead to excellent prediction of activity and excellent coverage of true ligand space. But as we will see, it is likely that many targets fall into the latter category. That is, an assumption of 2D similarity in predicting activity will be, at best, only partially correct and will identify only a subset of active compounds. However, we cannot know the full space of synthesizable true ligands of different binding sites and are limited to those designed by people, so we must consider this effect when we evaluate ligand-based prediction methods.

This bias issue also represents a limitation on drug discovery to the extent that our minds and our tools are making poor assumptions in predicting biological activity. If we can develop effective tools that employ a different and useful inductive bias than the natural bias of humans, the tools will accelerate the discovery of therapeutics that exhibit substantial improvements in efficacy and safety. The subject of this paper is a formal exposition of this issue of inductive bias, with detailed statistical arguments to support the assertions that chemists have historically shown a strong 2D bias and that this bias embeds assumptions that are non-optimal.

First, however, an anecdote is instructive. Figure 1 shows the structures of three phosphodiesterase 5 (PDE5) inhibitors, all patented within 10 years of one-another, and all marketed for the treatment of erectile dysfunction (ED). Sildenafil (Viagra) was the first to market, in 1998, and it established the role of PDE5 inhibitors in treating ED. Vardenafil (Levitra) and tadalafil (Cialis) followed, both in 2003, with the former receiving approval a few months ahead of the latter. The structural differences between sildenafil and vardenafil are subtle, amounting to an additional methyl group and a carbon/nitrogen swap in the heterocycle. In sharp contrast, the structural differences between sildenafil and tadalafil are substantial; however, as seen in Fig. 2, the two drugs bind PDE5 in overlapping volumes, sharing a hydrogen bond with a glutamine residue. While vardenafil is slightly more potent than sildenafil [2], its pharmacokinetics are very similar, notwithstanding a slightly shorter half-life [3]. Tadalafil, on the other hand, has a vastly improved half-life, resulting in an effective duration of action several-fold higher than either sildenafil or its trivial analog vardenafil [4]. Also, the absorption of sildenafil and vardenafil is reduced by high-fat food intake while the absorption of tadalafil is not influenced by food [3]. As of 2006, sildenafil still dominated the market, but tadalafil has outperformed vardenafil in terms of market share gains despite near simultaneous introduction in the marketplace, presumably in part due to its improved therapeutic profile (market data from 2006 annual financial reports of Pfizer, Lilly, GSK, Bayer, and Schering-Plough). This example illustrates two points. First, the human 2D oriented inductive bias is evidenced by the structure of vardenafil, which can be derived by a sequence of very small molecular editing operations starting with sildenafil. Second, therapeutically meaningful improvements can accompany drugs that move beyond "me-too" design.

Methods and data

The results in this paper focus primarily on the utility of methods for virtual screening when making use of a small number of ligands that bind a particular site as a *query* in an attempt to retrieve competitive ligands when they are intermixed among a set of *decoys*. With the exception of a

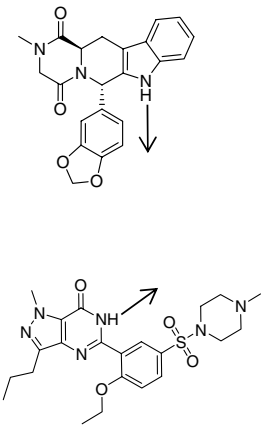


Sildenafil (Viagra)
Pfizer
Patented 1992

Tadalafil (Cialis)
Marketed by Lilly
Patented 1995

Vardenafil (Levitra)
Marketed by Bayer/GSK
Patented 2002

Fig. 1 Structures of three PDE5 inhibitors, all marketed as treatments for erectile dysfunction



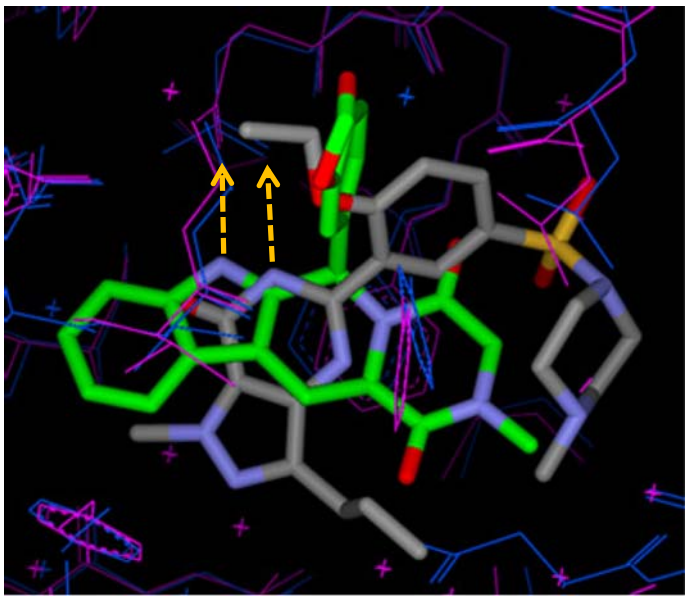


Fig. 2 Structures of Sildenafil and Tadalafil bound to PDE5 (PDB codes 2H42 and 1X0Z)

strawman 2D molecular similarity method, all of the other approaches have been described previously and will be briefly summarized here. The data and scripts for generating the results reported here are available by request from the corresponding author.

Data

Data came from three sources: 40 cases from the DUD dataset [5], four cases from the original report of Surflex-Sim [6], and 22 cases from our recent modeling study of a large portion of the space of known drugs [7]. These will be referred to as the DUD, Similarity4, and Similarity22 sets in what follows.

DUD

The DUD dataset consists of protein structures, known active ligands, and a set of decoys specifically designed with knowledge of known active ligands that are intended to represent a difficult background (i.e., physically similar but topologically distinct). Within the DUD decoy set, a distinction is made between "self" decoys and "all" decoys. The former are target-specific, and the latter are the superset of all self-decoy sets for all 40 targets. We also generated a subsampled set of decoys by randomly choosing one-tenth of the all decoy set. The actives for each target included as few as 10–20 ligands and as many as a few hundred. Each self-decoy set contained roughly 36 times as many. In the version used here, the all decoy set contained 124,413 molecules, and the sampled set contained 12,396 molecules. In the results that follow, the ligands and decoys from DUD were used unmodified.

Similarity4

The Similarity4 dataset consists of known ligands of the Histamine, Muscarinic, Serotonin, and $GABA_A$ receptors [6]. The first three are G-protein coupled receptors for which a significant amount of binding data has been curated within the GPCRDB. The last is a ligand-gated ion channel with multiple binding sites; the ligands used in this case bind what is termed the benzodiazepine binding site, which is named for the primary class of compounds known to bind the site. In our previous work, we have made use of a decoy set due to Rognan's group [8]. However, that set has been characterized as being less drug-like than others [5], and we have begun to make use of a set derived from the ZINC database [9, 10]. This decoy set contains 1,000 molecules, and among the decoy sets used by Huang et al. that *did not* make use of knowledge of active ligands (termed the "Jain set" in that study), it was the most challenging [5].

Similarity22

This set represents 22 different targets of known drugs [7]. The targets were diverse, and the drugs that targeted them were of diverse scaffolds (between targets). The targets included proteins within bacterial, viral, and fungal pathogens, with cognate drugs including azole antifungals, β-lactam antibiotics, sulfa drugs, quinolones, and

nucleoside analogs. Three targets were involved in cardiac indications (e.g., angiotensin converting enzyme), four were nuclear hormone receptors (e.g., estrogen receptor), three were involved in analgesia (e.g., the voltage-gated sodium channel), two in sedation (e.g., the $GABA_A$ receptor barbiturate binding site), and four were G-protein coupled receptors (e.g., the histamine receptor). In the original work, 57 drugs were used to construct ligand-based models of activity (either two or three molecules for each target), with 177 drugs used for testing. A number of targets had very few drugs for testing retrieval, and additional annotation has nearly doubled this number to 330 drugs, which represents nearly one-third of the space of approved small-molecule therapeutics in the United States. As in the Similarity4 case, the original work made use of the Rognan decoy set, but here we used the ZINC decoy set.

Methods

Surflex-Sim

The Surflex-Sim 3D molecular similarity method and its use for virtual screening have been described at length in multiple publications [6, 7, 11]. 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 (and possibly different internal volume overlaps) 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. Surflex-Sim implements a method to produce the optimal superimposition of N ligands (typically two or three) by simultaneously optimizing their joint similarity as well as minimizing the overall volume of the superimposed ligands. This involves deep search of the conformational and relative alignment parameters for the ligands. Virtual screening is done by optimizing the pose of each ligand in a test database to maximize its average similarity to the N ligands in a superimposition. High-ranking molecules are highly similar to the "query" and are likely to bind the same binding site.

For this work, we have used Surflex-Sim version 2.21, which makes relatively minor changes over the previously published version 1.31 [7]. The current version implements for Surflex-Sim a series of search features that had been implemented for Surflex-Dock version 2.11, which are described in detail in a recent paper [10]. These features include dynamic ring search, ligand minimization, and post-alignment all-atom optimization of ligand pose. Generation of molecular superimpositions was done as follows: `surflex-sim-maxdiff 100-ring hypo QueryMolsUnaligned.mol2 loghypo`. This resulted in 100 ranked superimpositions, the top scoring of which was used for virtual screening: `surflex-sim -ring align_list test-db.mol2 loghypo-hypo0.mol2 logtest`. This produced a score for each molecule within the test database, allowing for the computation of ROC curves to measure the separation of active from decoy ligands.

GSIM-2D

In this work, we made use of a 2D similarity method that has been implemented as a strawman, primarily for use as a control in order to identify ligand-retrieval problems that are relatively unchallenging. It is a very simple method. Given two molecules as input:

(1) Identify the subgraphs of molecule A of depth 1, 2, and 3 at each heavy atom.
(2) For each subgraph that contains 3 or more heavy atoms, check for the existence of the subgraph in molecule B.
    (a) Each matched atom pair must be the same element and have the same collection of connected heavy elements attached.
    (b) If the subgraph exists in molecule B, our score is incremented:
        (i) By the number of subgraph atoms (if the root atom is carbon).
        (ii) By 5 times the number of subgraph atoms (if the root is not carbon).
    (c) If the subgraph does not exist, the score is not incremented.
(3) We repeat the procedure for molecule B looking for its subgraphs within molecule A.
(4) The two scores are normalized to the interval [0,1] based on the maximum possible score in each direction.
(5) We compute the minimum of the ratio of number of heavy atoms in molecule A to molecule B and vice versa.
(6) 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 similar subgraphs, especially those rooted at heteroatoms. Minor differences in saturation will not affect the GSIM-2D similarity. The GSIM-2D functionality is a pose-independent method for virtual screening, and it is implemented in Surflex-Sim version 2.21. Screening a database against a query was done as follows: `surflex-sim gsim_list test-db.mol2 query.mol2 logtest`. This produced a score for each molecule within the test database, allowing for the computation of ROC curves to measure the separation of active from decoy ligands. Note that the query may be one or several molecules, and the average GSIM-2D similarity is reported.

Results and discussion

We will first discuss the issue of inductive bias in terms of comparing different types of methods: those that make use of a protein structure and those that make use of ligands instead. We will then show that the effects of a 2D similarity inductive bias in drug design are evident using two independent arguments. Last, we will show how to construct test cases for evaluating ligand-based methods that are both relevant and challenging and demonstrate how to employ 2D similarity methods *as a control* in evaluating the performance of more complex techniques.

Comparing docking and ligand-based virtual screening

Two recent papers have compared ligand-based methods with protein structure-based docking methods for performance in virtual screening. Hawkins et al. compared ROCS (which is ligand-based) with multiple docking approaches, but the study did not directly compare other ligand-based methods or 2D methods [12]. The study did consider the 2D diversity of actives retrieved by ROCS, but there remains a problem of interpretation on a fundamental level. The information used by docking (a protein structure) is distinct from the information used by a similarity method (known ligands). Since the input is different for each method, the inductive biases of the methods are definitionally very different, and the effects of dissimilar biases can be large. McGaughey et al. [13] compared multiple ligand-based methods with multiple docking methods, and this comparison included both sophisticated 3D similarity approaches (e.g., ROCS) and simple 2D methods (e.g., Daylight fingerprints). Their conclusion was that ligand-based methods worked better than docking and that 2D approaches (in particular TOPOSIM) perform better, marginally, than the 3D approaches.

The difficulty with both types of comparison can be illustrated by making use of the DUD dataset. Figure 3 shows the performance of Surflex-Dock and GSIM-2D on the PPAR-gamma case using the DUD self decoys as the decoy set. The two query ligands used by GSIM-2D are shown in the box. Clearly, GSIM-2D outperformed Surflex-Dock with respect to enrichment, both overall, and especially with respect to early enrichment. However, Fig. 3 also shows the structures of three different ligands from the retrieval set. The bottom one was the highest-ranked ligand by GSIM-2D. It differed from one query ligand by only deletion of a fluorine and saturation of a phenyl. The middle ligand was also an obvious analog of the query ligands. The top ligand, which was among the last actives in the GSIM-2D ranked list, was also an obvious analog. Figure 4 shows that whereas the decoy set matters quite a lot for docking, it matters very little for GSIM-2D. This is because the information implicitly captured in the two query ligands contains *nearly all* information required to identify the ligands to be retrieved. Note that the choice to employ two instead of one query ligand here makes little difference, since the actives are generally so self similar. In cases where greater structural diversity exists, using multiple ligands for the query is beneficial, which is why we have taken that approach in previous work and have repeated the approach here. The GSIM-2D method can be used with any number of query ligands.

Some raise the question of why this matters: the 2D method worked well so why should one care why it did? The theoretical reason is that one is *not* comparing the two methods fairly, since one method has a tremendous information advantage over the other. The practical reason is more subtle. It is useful to distinguish between virtual screening and SAR expansion. In the latter case, one may have identified a lead from a high-throughput screen and have good reason to believe that a large number of analogs might be readily available. If the PPAR-gamma ligands in Fig. 3 came from a non-target-directed combinatorial library, and the query ligands were identified from a high-throughput screen, this would be an appropriate case to test SAR expansion methods. In this case, retrieval of obvious analogs is the primary desired behavior of the method. By contrast, when one is performing a virtual screen in practice, it is typically desired to find *novel* lead compounds. It is *not* expected that numerous direct analogs of an active scaffold will be available in some library that is the subject of the virtual screen.

We performed an analogous experiment to that shown in Fig. 3 for GSIM-2D on PPAR-gamma for all 40 DUD targets. For each target, three queries were constructed by randomly selecting two actives each time. For each query, the remaining actives along with the DUD decoy set were

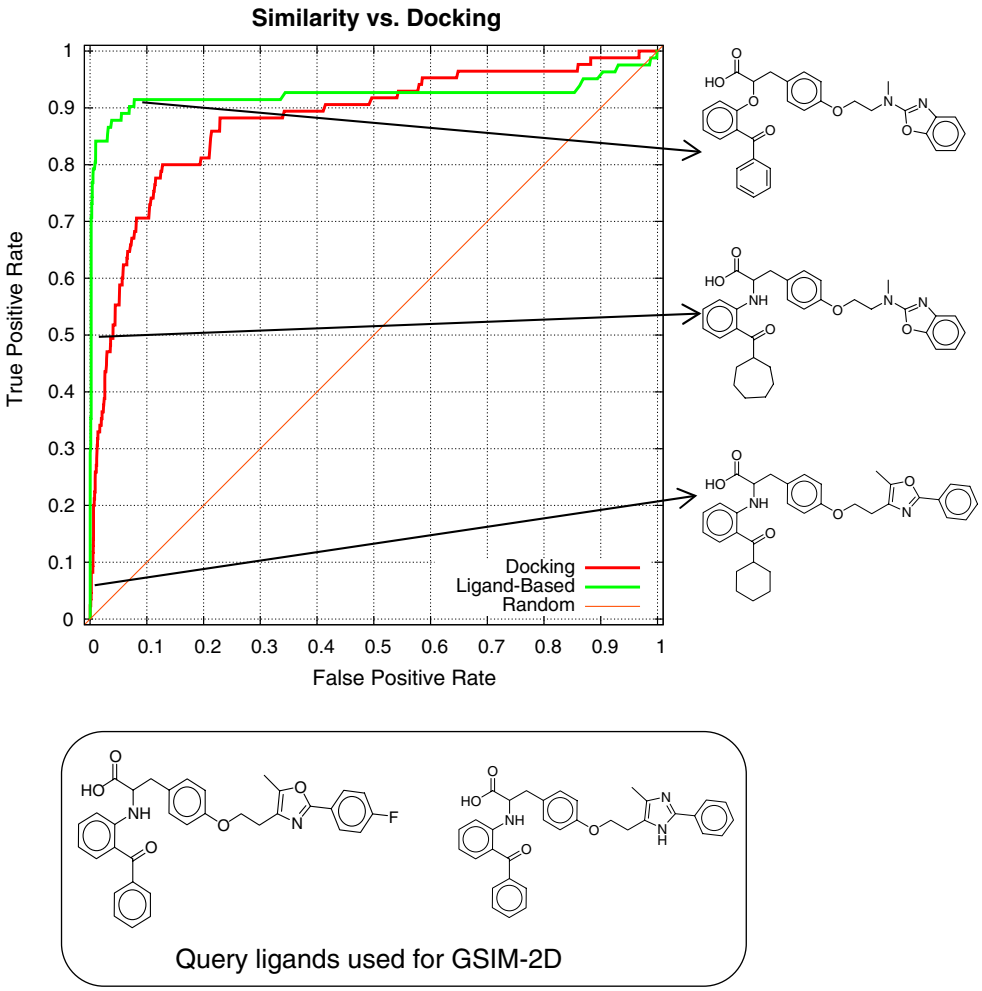


Fig. 3 Relative performance of docking and 2D similarity on the DUD target PPAR-gamma, using the "DUD self" decoys. Docking results were from Surflex-Dock, and ligand-based retrieval results were from GSIM-2D. The ligands in the box were used as the query structures for GSIM-2D. In the case of docking, no *direct* knowledge of any ligands is used to make predictions of activity, but 2D similarity makes use of active ligands. In cases such as this, where the ligands to be retrieved represent minor variations of the ligands used for the query, the meaning of such a comparison is questionable. Any ligand-based method should be able to perform well in a case such as this

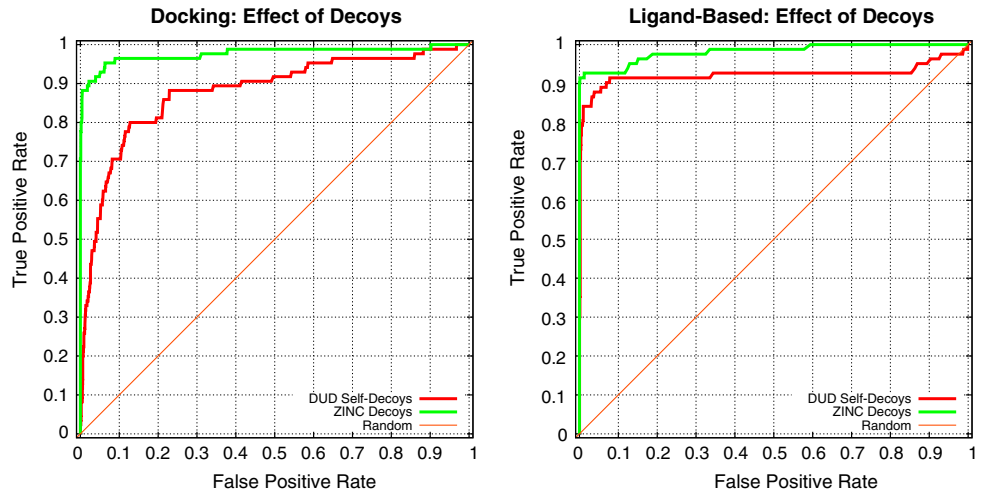


Fig. 4 Different decoy sets have a strong impact on the docking results (left), but the 2D-similarity-based results are affected relatively little (right). This is because, the latter case is dominated by the 2D-self-similarity of the actives, which are retrievable against essentially any background

used to test GSIM-2D's virtual screening performance. ROC areas were computed for each run, and the three values for each target were averaged. We also computed mean enrichment factors for each target in order to make a direct comparison to the values reported by Huang et al. for DOCK [5]. In 37/39 cases, GSIM-2D yielded higher EF 1% values than DOCK 3.5.54. However, this does not tell the full story. Figure 5 plots the performance of GSIM-2D on all 40 DUD targets (Y-axis) against a measurement of scaffold diversity among the DUD actives for each target (Andy Good, personal communication). For the cases in which scaffold diversity was low, GSIM-2D performed

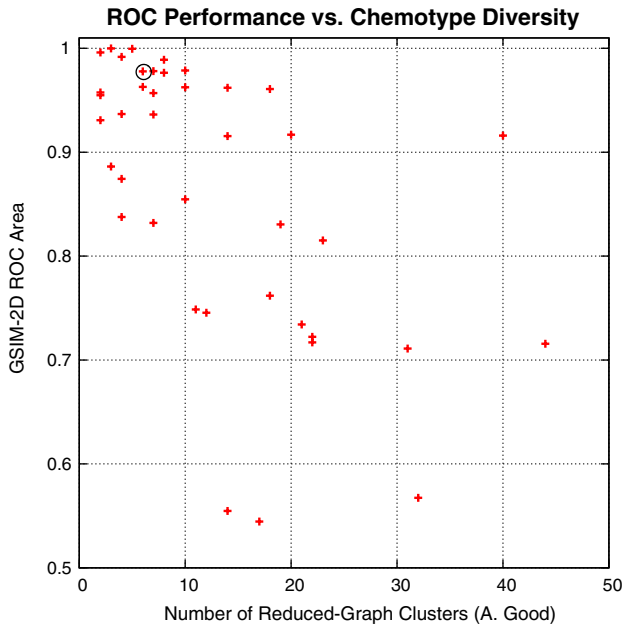


Fig. 5 Performance (ROC AUC) for GSIM-2D on all 40 DUD targets plotted against a measurement of chemotype diversity. In cases of low chemotype diversity (10 or fewer scaffold graphs), GSIM-2D showed an average performance of 0.94. In cases of higher chemotype diversity, the performance averaged 0.77 ($p << 0.001$ by *t*-test). The PPAR-gamma case from Fig. 3 is circled

extremely well (mean ROC area of 0.94). But as scaffold diversity increased, performance decreased (mean ROC area of 0.77).

Given these observations, it would be *true* to say that "GSIM-2D exhibited robust and substantially better performance than DOCK." But it would be *fair* (and true) to say instead that "the DUD screening dataset provides an excellent means for testing the effects of bias in the form of obvious analogs in evaluating molecular similarity methods." Datasets for which simplistic 2D methods perform well in virtual screening *do not* serve as appropriate test sets for assessing virtual screening performance based on ligand structures. These "2D-biased" datasets *may* be appropriate to test the relative merits of approaches for rapid SAR expansion, but many investigators believe that to be a largely solved problem.

Inductive bias in the design of drugs

We will now turn to the issue of whether a 2D inductive bias can be shown in the design of drugs and whether it matters. Recall Fig. 1, where the structures of three different PDE5 inhibitors were shown. Both tadalafil and vardenafil were introduced shortly after sildenafil, in order to compete for patients within an important therapeutic niche. One was a close analog of sildenafil, and one was not. What has been the typical case? Figure 6 shows the cumulative histograms of 2D similarities of drug pairs that hit the same primary (desired) target and those whose primary targets are different. These are highly separable distributions (ROC area 0.86). Therefore, based on the structure of a query drug, 2D similarity methods are capable, in very many cases, of identifying other cognate drugs against a background of non-cognate drugs.

Figure 7 shows the same plot, but it adds two additional cumulative histograms. The blue curve shows the 2D similarities for drug pairs that hit the same primary target *and* which were patented within 10 years of one another. The magenta curve shows the analogous data for drug pairs patented 20 years or more apart. There are two shifts. First, the distribution of 2D similarities for drugs with close patent years shifts to the right, indicating that they have relatively higher 2D similarity than typical drug pairs. Second, for distant patent years, we see a marked shift to the *left*, indicating that it is much more common to find different-looking structures for drugs hitting the same target when a long time has passed between their discovery. The case of tadalafil is clearly an outlier: a very different looking structure discovered in a short time period from the first drug in a class.

It is impossible to *prove* why this is the case, but a plausible explanation is straightforward. If one is approaching drug design for a therapeutic area that is

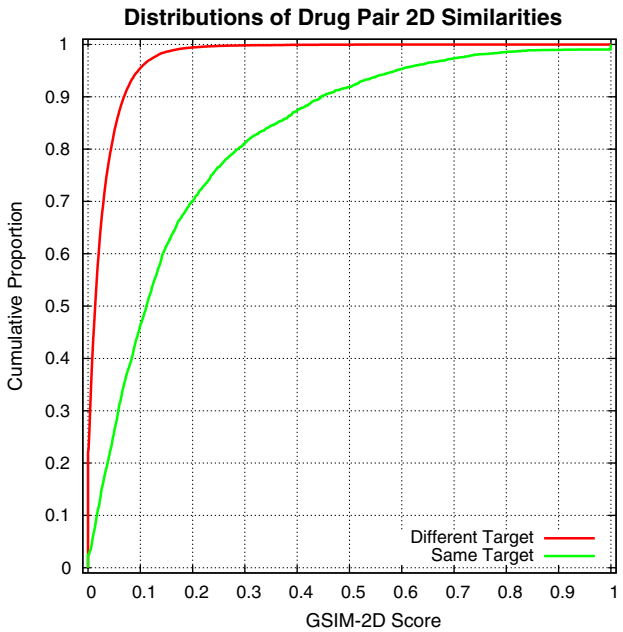


Fig. 6 Cumulative histograms of 2D drug-pair similarity for pairs that hit the same primary target and pairs that hit different primary targets. These are highly separable distributions (ROC AUC 0.86, ROC curve not shown). Each distribution is made up of many thousands of drug-pair comparisons, from a total population of over 850 deeply annotated drugs

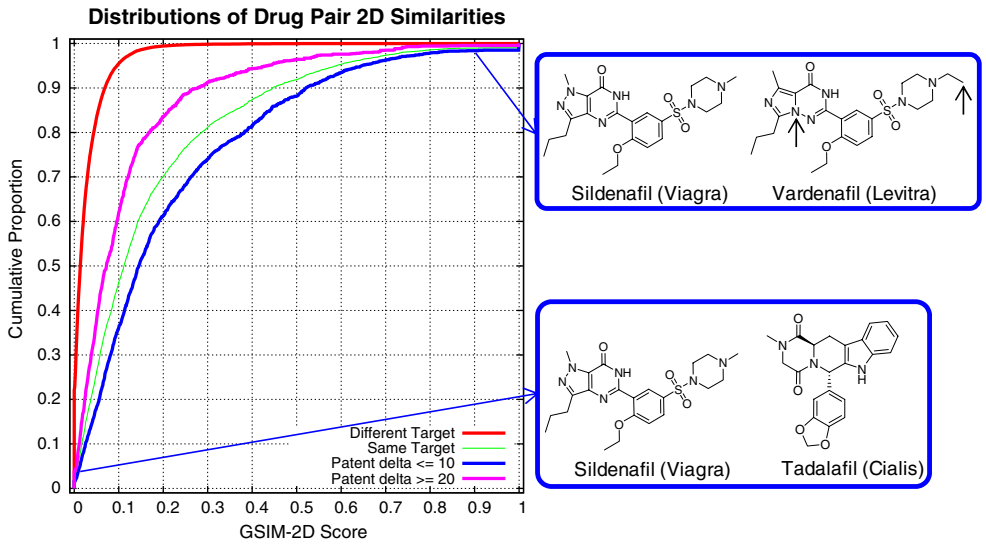


Fig. 7 Cumulative histograms of 2D drug-pair similarity for pairs that hit the same primary target and pairs that hit different primary targets. In addition, the same-target pairs have been split into pairs with relatively close patent years (10 or fewer) and distant patent years (20 or more). We see that the tendency toward a simple "me-too" analog is more pronounced among drug pairs that are patented close in time, reflecting the economic viability of showing little improvement in therapeutic benefit. However, with distant patent years, we see the opposite tendency, with drug pairs showing much less 2D similarity. We believe this reflects the economic need to show substantial therapeutic advantage in competing with off-patent medications, and that this therapeutic improvement is aided by creativity in design that goes beyond obvious analogs. The example of Vardenafil, with high 2D similarity to the recent predecessor Sildenafil is relatively common. The example of Tadalafil, with extremely low 2D similarity is an outlier among drugs with close patent years. It is twice as common to see drug pairs with low 2D similarity (<0.1) when their patent year difference is 20 years or greater than 10 years or less

currently served by on-patent medications and where the economic opportunity is large enough that another drug with nearly identical properties would be profitable, a sensible strategy is to design a molecule that is extremely similar to an existing drug (e.g., vardenafil designed from sildenafil). However, when there are a significant number of off-patent drugs to serve an indication, new therapies would need to represent a significant improvement with respect to either efficacy or safety in order to be economically viable. Our conjecture is that the decrease in 2D similarity between drugs that have distant patent years results from the *necessity* to design non-obvious analogs in order to yield therapeutic improvements.

It is possible to get at the question of whether human beings actually exhibit this 2D inductive bias by considering the types of molecules they *design* to hit a particular target and comparing those to the molecules that they designed to hit a *different* target but which, inadvertently, hit the first target. Consider drug A, whose desired target is X, and drug B whose desired target is Y, but which has a side effect due to binding target X. Figure 8 compares the distribution in 2D similarities of three classes of drug pairs: (1) drug pairs that do not hit the same primary targets (red curve), (2) drug pairs that hit the same primary target (green curve), and (3) pairs for which the drugs hit different primary targets *but* where one of the two hits the other's primary target as an unintended side-effect. The relative 2D similarities of the "side-effect pairs" are quite low. The separation between the cross-target pairs and the side-effect pairs is *smaller* than that between the side-effect pairs and the same-target pairs. If there were no 2D inductive bias, the side-effect pair distribution *would be the same* as the same-target distribution. The green box shows typical examples of *intentional* molecular design, with each pair exhibiting the characteristics of obvious analogs as with sildenafil and vardenafil. However, the blue box shows examples of *unintentional* design, where the intended target was different from what was hit. Here, the relative 2D similarity is much lower, but the therapeutically relevant physical binding of the pairs is real. So, when people are actively acting as predictors of molecular activity, they show a strong bias toward obvious 2D similarity. When we control for the issue of design intention, removing the aspect of activity prediction toward an intended target, we see this phenomenon greatly reduced. The very high level of 2D similarity of drugs that bind the same target does not follow from physics or biology. It is related to the inductive bias of the design process.

Figure 9 shows three specific cases of drug pairs that have relatively low 2D similarity but which have reasonable levels of 3D similarity. Each point within the central plot shows the 2D and 3D similarity of a single pair of drugs (4,918 unique pairs total, reflecting the pairwise comparisons of 865 drugs with annotated primary targets).

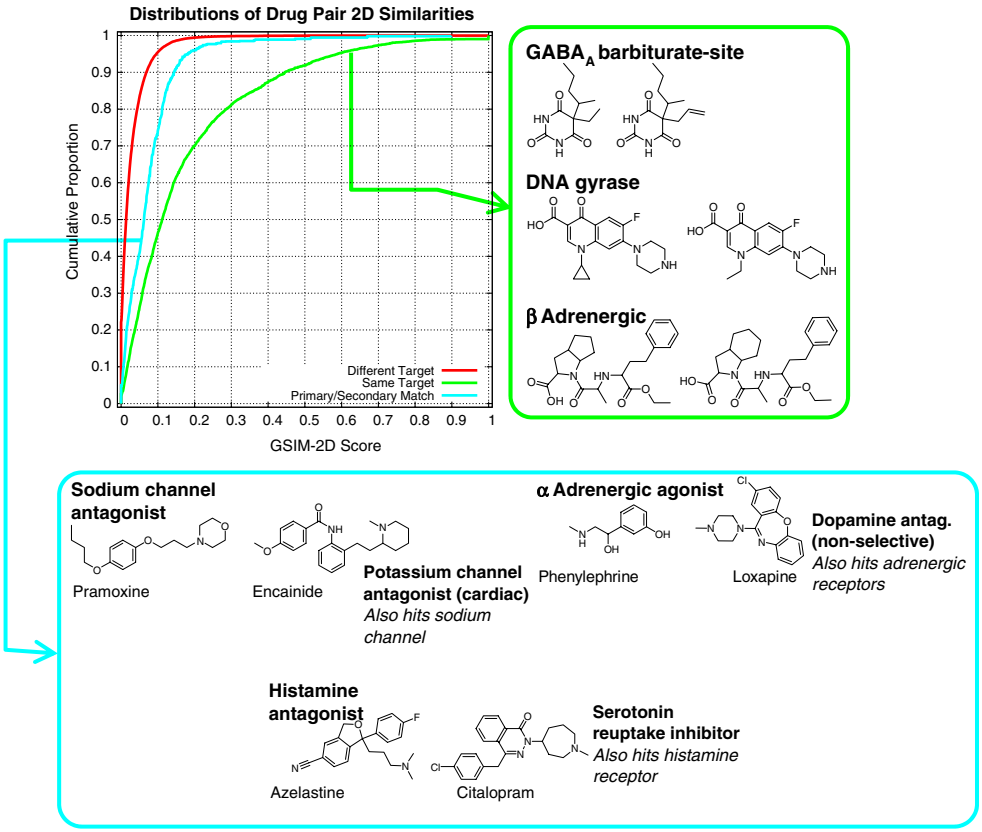


Fig. 8 The cumulative histograms of 2D similarity were computed for drug pairs that hit the same primary (desired) target (green curve), pairs that do not hit the same primary target (red curve), and pairs where one drug hits the primary target of the other drug as an undesirable side-effect (cyan curve). Typical examples of primary target drug pairs are shown in the green box. Typical examples of "side-effect pairs" are shown in the cyan box. The effect of strong 2D similarity was very significantly reduced in the cases where molecular activity was the result of undesired effects, illustrating the inductive bias present in the intentional design of ligands for specific targets

The upper right portion of the plot contains drugs such as the examples in the green box of Fig. 8 very similar by both 2D and 3D similarity. The upper left, where the three examples of Fig. 9 reside, contains drug pairs with low 2D similarity but where 3D similarity, as measured by Surflex-Sim's imprint-based approach [7, 14], is moderate to high. The PDE5 case has been discussed, but the other two cases (left and bottom) present interesting examples as well. At left are two competitive ligands for the benzodiazepine binding site of the GABA$_A$ receptor. Zolpidem (Ambien) has a clearly distinct chemical structure and is purported to be less addictive than diazepam [15]. At the bottom of

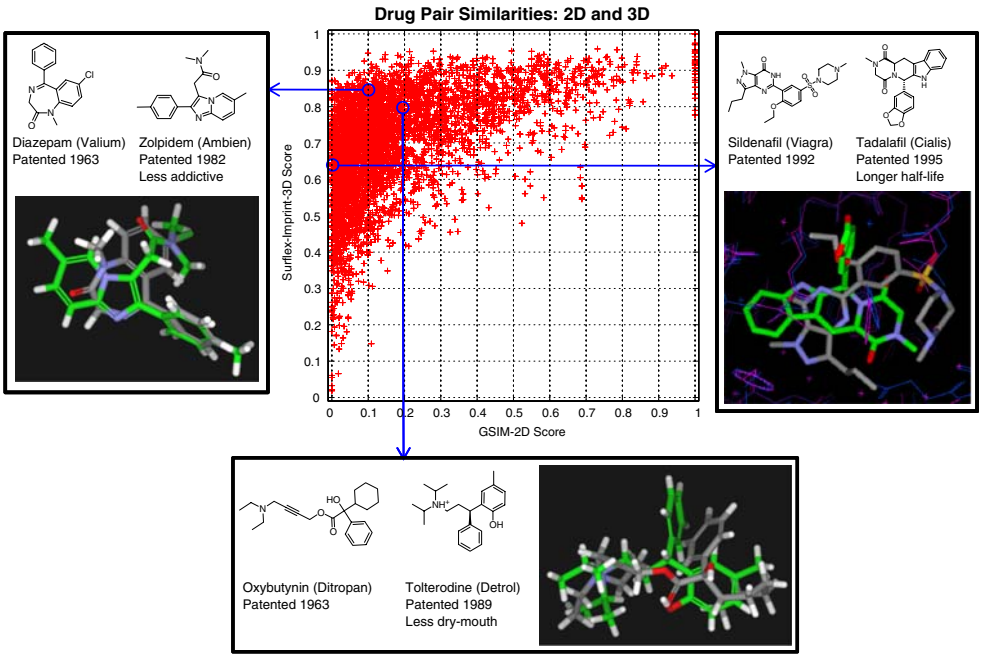


Fig. 9 Molecular similarities were computed for drug pairs, with 2D similarity plotted on the *X*-axis and 3D similarity on the *Y*-axis. Three examples are shown of pairs that show low 2D similarity but which have moderate to high 3D similarity. The 3D overlays were generated by Surflex-Sim for the cases at left and bottom, and were experimentally determined for the case at right

Fig. 9 are two drugs used in the treatment of urinary urge-based incontinence based on muscarinic receptor antagonism, with tolterodine (Detrol) exhibiting reduced dry-mouth side-effects relative to the older oxybutinin [16]. In all three cases, a newer drug with a significantly different underlying scaffold yielded a therapeutic benefit. In each case, whether the 3D alignment was made experimentally (for PDE5) or computationally (for the other two), it is clear that there are some problems for which a 3D approach must be undertaken.

Relevant and hard cases for testing ligand-based modeling

Clearly, in cases where an experimental crystal structure of a drug target is known, docking methods frequently perform well and largely moot the issue of inductive bias. However, for large classes of targets, including G-protein coupled receptors, ligand-gated ion channels, and membrane-bound transporters, atomic-resolution protein structures are extremely challenging. In these cases, which certainly do represent important therapeutic indications, ligand-based approaches remain important. The particularly critical problem is, of course, *not* the one that has been solved: rapidly finding analogs of a structure. If, given a small number of ligands, one can identify novel ligands that are likely to bind the *desired* target site but which also possess *other* beneficial properties lacking in the known ligands, that would be a genuinely useful contribution. So how does one construct test cases that can measure the performance of methods that seek to address this problem?

The key is to control for human inductive bias by reducing the linkage between the molecules used as queries and the active molecules to be retrieved as part of the test of a method. As shown above, the degree to which 2D similarity relates drugs with distant patent years is much lower than in the case of drugs designed close to one another in time. So, making use of *query* ligands that come from a time period that is *distant* from the ligands to be retrieved can help control for this issue. We also saw that in cases where drugs hit *unintended* targets, they are relatively dissimilar in a 2D sense from the drugs that were *designed* to hit those targets. So, one should endeavor to find ligands for retrieval tests that represent examples from both designed ligands and *accidental* ones.

The report introducing Surflex-Sim took exactly this approach [6], although the motivation at that time was not as carefully refined in the context of inductive bias as is laid out here. Figure 10 shows the structure of four sets of query ligands for four pharmaceutical targets, three being GPCRs and one being the $GABA_A$ receptor. In the top two cases, the query ligands represented effort from within single projects during short periods of time. The ligands to be retrieved included a diverse set of molecules. The diversity came from both the time axis (decades worth of ligands) and the intention axis (many ligands had their activity annotations derived from GPCRDB, and those activities were not the intended ones for the ligands). In the third case, the retrieval ligands shared these characteristics, but the query ligands were chosen from among the earliest marketed drugs targeting the histamine receptor. In the last case, the query ligands were early discoveries in $GABA_A$ receptor modulation (all binding the "benzodiazepine" binding site). While effort was made to identify a diverse set of retrieval ligands, very few non-benzodiazepines have been shown to be competitive for their binding site, and this case had the least structurally diverse retrieval set.

Table 1 shows the results of using Surflex-Sim and GSIM-2D on these four targets for virtual screening. In this work, we have made use of a more challenging decoy set (see Methods) as well as quantifying both ROC area and confidence intervals in order to rigorously establish performance. We see that in the three GPCR cases, we observed significantly better performance for Surflex-Sim over GSIM-2D. Also note that the GSIM-2D performance is much lower than the performance observed on the 40 targets from the DUD set (see Fig. 5). By making an explicit effort to control for the effects of inductive bias in the construction of the test, it was possible to establish the value of a 3D approach for ligand-based virtual screening. Note, however, in the case of the $GABA_A$ receptor, we do not see a significant difference between GSIM-2D and Surflex-Sim. This reflects the high degree of 2D structural similarity of the retrieval ligands to the queries.

Our more recent paper that focused exclusively on the space of known drugs for ligand-based modeling showed robust performance across 22 targets for Surflex-Sim [7]. While covering nearly one-quarter of small-molecule therapeutics, the number of ligands for model testing was limited to a very small number in some cases. Here, we present the results of ligand-based virtual screening using 2 or 3 cognate ligands as the query with ligand-retrieval sets that have nearly doubled in size. We have also employed the more challenging ZINC background and made explicit quantitative comparisons to the GSIM-2D method. Table 2 shows the result for the 22 targets. They parallel the results from the $GABA_A$ receptor case above, with no statistically significant differences between the two methods in *any* of the 22 cases. This, of course, makes perfect sense, given the very strong inductive bias shown in Fig. 9. There is generally so little genuine structural variation among the scaffolds of known drugs that were designed to hit the same targets that 2D methods will very frequently perform very well in such tests. Note, however, that this inductive bias degeneracy does not extend itself to unintended

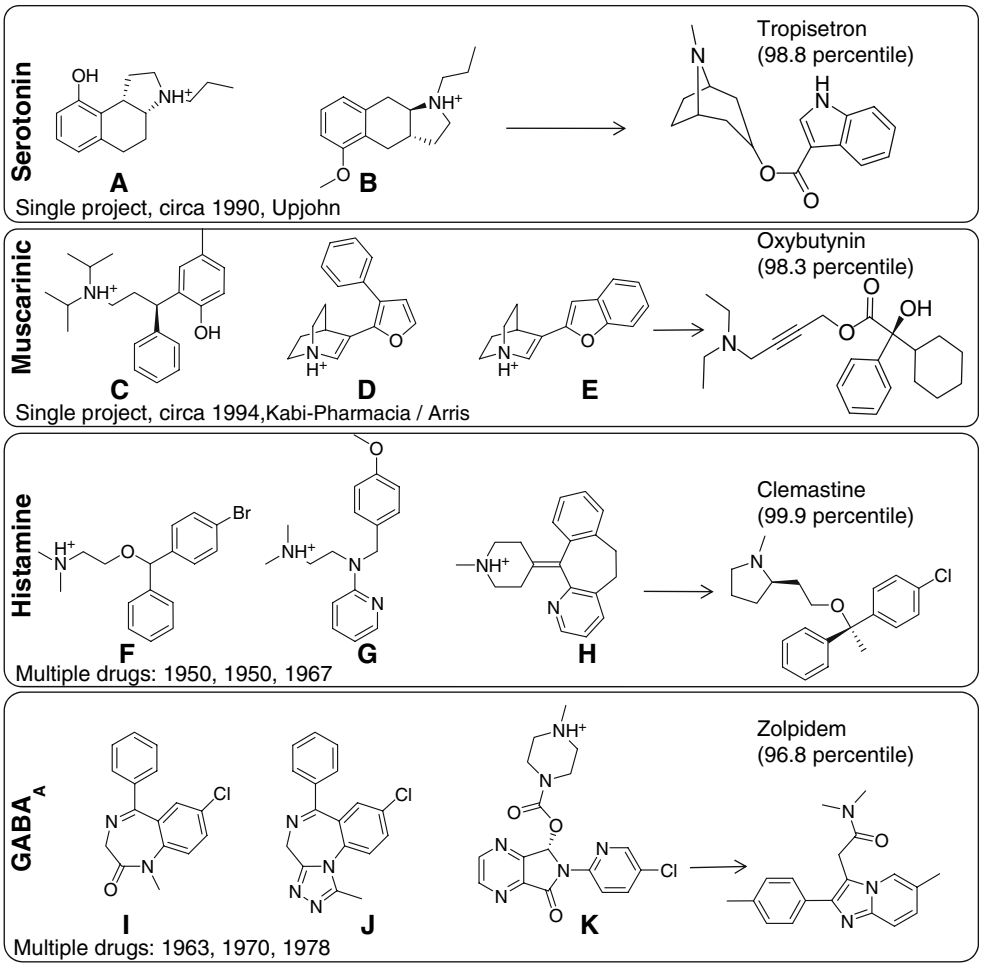


Fig. 10 Query structures (left) and examples of retrievable drugs (right) for four cases of pharmaceutical relevance. Despite substantial scaffold variation, the ligands were retrieved at high ranks from a virtual screening database. These represent non-trivial examples where an inductive bias that presumed 2D similarity would yield poor predictions of activity

Table 1 Performance of Surflex-Sim, with an explicitly quantified 2D control

Target	Ntest	Surflex-Sim (ZINC)	95% CI	GSIM-2D (ZINC)	95% CI
Histamine	48	0.946	0.91–0.98	0.736	0.66–0.82
Muscarinic	44	0.926	0.89–0.96	0.752	0.66–0.83
Serotonin	30	0.883	0.83–0.93	0.768	0.70–0.83
$GABA_A$ BZR	15	0.867	0.73–0.97	0.959	0.90–1.00

targets, as we showed in our previous study in considering the off-target effects of drugs [7].

Conclusions

Recall the anecdote earlier about PDE5 inhibitors. The leap from sildenafil to vardenafil was not large, and could have been managed without the use of screening or computational methods. However, the leap from sildenafil to tadalafil was large. In the former case, the very high similarity in structure was accompanied by little change in overall pharmacology. In the latter case, the jump to tadalafil led to a vast increase in the temporal therapeutic window (days not hours). The example of PDE5 inhibitors for treatment of ED is not a singular anecdote. Figure 11 shows the structures of lovastatin and atorvastatin, which are both HMG-CoA reductase inhibitors used to treat cardiovascular disease. Lovastatin established the market and was followed by a number of me-too analogs. The novel scaffold of atorvastatin brought improvements in therapeutic profile, and atorvastatin (marketed as Lipitor) is currently the single best selling drug in the marketplace, dominating the statin segment. The most significant side effects of statins are myopathies including fatal rhabdomyolysis [17], and atorvastatin exhibits significantly reduced risk of myopathies compared with other statins [18]. The binding modes of many statins have been determined by crystallography [19], and the large, relatively rigid, and non-obvious component of atorvastatin reaches a different part of the binding pocket within HMG-CoA reductase. It appears likely that this component is responsible for the lack of interaction with some secondary

Table 2 Performance of Surflex-Sim on drug targets, exclusively predicting drug molecules against the ZINC background, with an explicitly quantified 2D control

Target	Target name	Ntest	Surflex-Sim (ZINC)	95% CI	GSIM-2D (ZINC)	95% CI
target_a	Lanosterol demethy-lase	9	0.983	0.96–1.00	0.899	0.75–1.00
target_b	Carboxypeptidase	36	0.972	0.94–0.99	0.987	0.97–1.00
target_c	Dihydropteroate synthase	13	0.996	0.99–1.00	0.962	0.90–1.00
target_d	DNA gyrase	10	0.808	0.63–0.94	0.973	0.92–1.00
target_e	HIV Rev. transcriptase	6	0.988	0.97–1.00	0.856	0.65–0.98
target_f	L-Type calcium channel	5	0.923	0.81–1.00	0.999	1.00–1.00
target_g	Acetylcholinesterase	9	0.636	0.43–0.82	0.800	0.57–0.95
target_h	Angiotensin I Conv. Enz.	8	0.938	0.81–1.00	0.999	1.00–1.00
target_i	b-Adrenergic receptor	10	1.000	1.00–1.00	0.992	0.98–1.00
target_j	Opioid receptor Mu	22	0.958	0.91–0.99	0.918	0.87–0.95
target_k	Sodium channel	28	0.719	0.61–0.81	0.774	0.67–0.86
target_l	Estrogen receptor	12	0.981	0.96–1.00	0.990	0.98–1.00
target_m	Progesterone receptor	7	0.993	0.98–1.00	1.000	1.00–1.00
target_n	Androgen receptor	13	0.886	0.75–0.99	0.976	0.95–1.00
target_o	Gluco/Corticosteroid Rec.	26	0.998	1.00–1.00	1.000	1.00–1.00
target_p	COX-I COX-II	27	0.715	0.63–0.79	0.704	0.60–0.80
target_q	$GABA_A$ barbiturate site	12	0.993	0.98–1.00	0.940	0.86–1.00
target_r	$GABA_A$ benzodiazepine	14	0.950	0.89–1.00	0.946	0.87–1.00
target_s	Muscarinic receptor	21	0.961	0.89–1.00	0.952	0.92–0.98
target_t	Histamine receptor	30	0.917	0.85–0.98	0.900	0.84–0.94
target_u	Nacl cotransporter	6	0.997	0.99–1.00	0.995	0.99–1.00
target_v	Sulfonyl urea receptor	6	0.745	0.60–0.89	0.918	0.78–1.00

target underlying the myopathy side effects. Examples of the non-obvious jumps shown here represent challenges for 3D methods (both ligand-based and docking), and they are clearly beyond the capability of 2D methods.

Ligand-based modeling is an important and valuable approach for drug design. For targets lacking an accurate structure, it is a principal method, and even in cases where a protein structure exists, ligand-focused approaches can be complementary. There are relevant and unsolved problems within this subfield of modeling. However, the issues of inductive bias can serve to obscure the differences among methods if the issues are not addressed explicitly in dataset construction, use of proper controls in evaluations, and in the interpretation of results. We must realize that most of the ligands on which we test our methods have been made as part of a predictive modeling process in which human beings have a demonstrable 2D bias. This bias leads to an under-representation of novel jumps like those from sildenafil to tadalafil and lovastatin to atorvastatin. Uncritical acceptance of nominally high performance of methods on tests in which the bias issue has been ignored will lead to the erroneous conclusion that the methods under study are adequate to the problems we face. Drug discovery can be improved by employing sophisticated 3D methods for ligand-based modeling, where the implicit bias in prediction is a close match to the *physical interactions* of ligands with proteins as opposed to a close match to the *thought process* evidenced in our largely incremental pharmacopoeia.

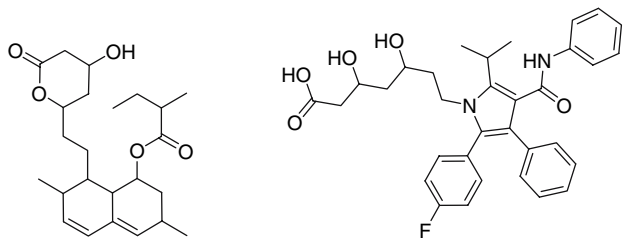


Fig. 11 Statin drugs further illustrate the importance of creativity in drug design. While Lovastatin established the statin class, Atorvastatin is currently the world's best selling pharmaceutical, based in part on a significant improvement with respect to non-target-related side effects. Note that the lactam of lovastatin opens up in vivo to the identical $\beta\beta$-hydroxy acid seen with atorvastatin

Acknowledgements The authors gratefully acknowledge NIH for partial funding of the work (grant GM070481). Drs. Jain and Cleves have a financial interest in BioPharmics LLC, a biotechnology company whose main focus is in the development of methods for computational modeling in drug discovery. Tripos Inc., has exclusive

commercial distribution rights for Surflex-Sim, licensed from Biopharmics LLC.

References

1. Mitchell TM (1997) Machine learning. McGraw-Hill, New York
2. Crowe SM, Streetman DS (2004) Vardenafil treatment for erectile dysfunction. Ann Pharmacother 38(1):77–85
3. Wright PJ (2006). Comparison of phosphodiesterase type 5 (PDE5) inhibitors. Int J Clin Pract 60(8):967–975
4. Supuran CT, Mastrolorenzo A, Barbaro G, Scozzafava A (2006) Phosphodiesterase 5 inhibitors–drug design and differentiation based on selectivity, pharmacokinetic and efficacy profiles. Curr Pharm Des 12(27):3459–3465
5. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801
6. Jain AN (2004) Ligand-based structural hypotheses for virtual screening. J Med Chem 47(4):947–961
7. Cleves AE, Jain AN (2006) Robust ligand-based modeling of the biological targets of known drugs. J Med Chem 49(10):2921–2938
8. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43(25):4759–4767
9. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45(1):177–182
10. Jain AN (2007) Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. J Comput Aided Mol Des 21(5):281–306
11. Jain AN (2000) Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. J Comput Aided Mol Des 14(2):199–213
12. Hawkins PC, Skillman AG, Nicholls A (2007) Comparison of shape-matching and docking as virtual screening tools. J Med Chem 50(1):74–82
13. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon J-F, Cornell WD (2007) Comparison of topological, shape, and docking methods in virtual screening. J Chem Inf Model 47(4):1504–1519
14. Ghuloum AM, Sage CR, Jain AN (1999) Molecular hashkeys: a novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. J Med Chem 42(10):1739–1748
15. Jaffe JH, Bloor R, Crome I, Carr M, Alam F, Simmons A, Meyer RE (2004) A postmarketing study of relative abuse liability of hypnotic sedative drugs. Addiction 99(2):165–173
16. Clemett D, Jarvis B (2001) Tolterodine: a review of its use in the treatment of overactive bladder. Drugs Aging, 18(4):277–304
17. Schreiber DH, Anderson TR (2006) Statin-induced rhabdomyolysis. J Emerg Med 31(2):177–180
18. Waters DD (2005) Safety of high-dose atorvastatin therapy. Am J Cardiol 96(5A):69F-75F
19. Istvan ES, Deisenhofer J (2001) Structural mechanism for statin inhibition of HMG-CoA reductase. Science 292(5519):1160–1164