Do molecular fingerprints identify diverse active drugs in large-scale virtual screening? (no)

Vishwesh Venkatraman,[*,†] Jeremiah Gaiser,[‡,¶] Daphne Demekas,[§] Amitava Roy,[¶,‖] Rui Xiong,[⊥] and Travis J. Wheeler[*,§,‡]

†*Norwegian University of Science and Technology, Department of Chemistry, 7491 Trondheim, Norway*

‡*Department of Computer Science, University of Montana, Missoula, MT, USA*

¶*Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rocky Mountain Laboratories, Hamilton, MT, USA.*

§*Department of Pharmacy Practice and Science, University of Arizona, Tucson, AZ, USA*

‖*Department Of Biomedical And Pharmaceutical Sciences, University of Montana, Missoula, MT, USA.*

⊥*Department of Pharmacology & Toxicology, University of Arizona, Tucson, AZ, USA*

E-mail: vishwesh.venkatraman@ntnu.no; twheeler@arizona.edu

Abstract

Computational approaches for small-molecule drug discovery now regularly scale to consideration of libraries containing billions of candidate small molecules. One promising approach to increased speed in evaluating billion-molecule libraries is to develop succinct representations of each molecule that enable rapid identification of molecules with similar properties. Molecular fingerprints are thought to provide a mechanism for

producing such representations. Here, we explore the utility of commonly-used fingerprints in the context of predicting similar molecular activity. We show that fingerprint similarity provides little discriminative power between active and inactive molecules for a target protein based on a known active – while they may sometimes provide some enrichment for active molecules in a drug screen, a screened dataset will still be dominated by inactive molecules. We also demonstrate that high-similarity actives appear to share a scaffold with the query active, meaning that they could more easily be identified by structural enumeration. Furthermore, even when limited to only active molecules, fingerprint similarity values do not correlate with compound potency. In sum, these results highlight the need for a new wave of molecular representations that will improve the capacity to detect biologically active molecules based on similarity to other such molecules.

Introduction

Methods for computational identification of small molecules likely to bind to a drug target (virtual screening) are increasingly intended to explore a space of billions of candidate molecules.[1–4] One strategy for exploring this massive molecular search space is to begin with a collection of known or presumed active molecules (seeds), and use those seeds as the basis of a rapid search among billions of candidates[5] for other molecules expected to demonstrate similar activity.[6,7]

The notion that small molecules with similar structure are likely to share biological properties, coined the *similar property principle* (SPP),[8,9] is central to this sort of search strategy. The SPP is simple and intuitive, and has served as the basis for predictions of biological activity,[10,11] toxicity,[12–14] aqueous solubility[14,15] ($\log S$), and partition coefficient[16] ($\log P$); it is, however, difficult to assess objectively, and may not necessarily reflect the similarity in biological activity from a medicinal chemist's perspective.[17] Furthermore, proper definition of structural similarity depends on the context. For example, the quantitative

structure-activity relationship focuses on similarity between local structural features of two molecules, while similarity in biological activity typically depends on more global features of the molecules[18–20] (though even these notions of *local* and *global* similarity are also not well defined).

The most common way to quantify structural similarity of two small molecules begins with calculation of a so-called *molecular fingerprint*, a binary or count vector that encodes structural and often chemical features.[21–23] Such a fingerprint is computed for each molecule, then the fingerprints of molecules are compared for overlap to approximately assess molecular similarity, motivated by the notion that this measure of fingerprint similarity will correlate with similarity at the level of biological activity. Fingerprint similarity has been used to effectively estimate $\log S$ and $\log P$ values,[24] likely because these values can largely be approximated from the small molecule itself without explicitly considering interacting partners.

Other molecular properties involve a dependency on context, placing greater strain on the utility of the SPP. For example, biological activity of a small molecule depends on the interaction between that molecule and the target protein binding region. Such binding regions (or pockets) vary between different proteins, and therefore impose strong context dependence in biological interactions. Consequently, small molecule ligand-based fingerprint similarity may not be sufficient to capture the wide spectrum of similarities in biological activities. Similarly, the toxicity of a small molecule depends on the molecule's interaction with multiple proteins, suggesting a limit to the inference power provided by similarity at the level of molecular fingerprints.

Despite previous demonstrations of the limitations in using SPP to predict similarity in biological activity,[25,26] the technique is heavily used in drug development. This is especially true in fingerprint-based virtual screening (VS), in part due to the computational simplicity and speed of searching the vast chemical space of small molecules.[2,27–29] A variety of molecular fingerprints have been devised for use in ligand-based virtual screening (LBVS), to aid in identifying biologically active small molecules[29,30] (hereafter referred to as actives), from

within a library of small molecules. LBVS begins with a small number of known/predicted actives (queries) for a target protein pocket, and explores a library of other small molecules, seeking a set that is expected to also be active. This expectation is based on the SPP, so that LBVS seeks molecules with high fingerprint similarity to one of the queries, under the assumption that fingerprint similarity to known actives will generally assign a higher ranking to actives than to non-actives (decoys). Martin et al.[25] have highlighted a disconnect between empirical and computational perception of similarity, and suggested resulting limitations in the utility of LBVS for drug binding predictions. Here, we explore the shortcomings of simplistic molecular fingerprints in the context of LBVS, and demonstrate that all commonly used fingerprint methods fail to sufficiently enrich for binding activity in a library of mostly decoy molecules. We are not oblivious to a prevailing perception that fingerprint similarity is suggestive of similarity in drug interaction potential, and hope that this manuscript contributes to a better understanding of the limitations of fingerprint similarity in the context of large-scale drug screening.

Methods and Materials

Fingerprint Representations

The palette of fingerprints evaluated in this study (Table 1), can be broadly classified into those based on (i) path, (ii) circular features, (iii) pharmacophore, and (iv) pre-defined generic substructures/keys.[31]Circular and path-based fingerprints are generated using an exhaustive enumeration of (linear/circular) fragments up to a given size/radius, which are then hashed into a fixed-length bit vector. The SIGNATURE descriptor[32] generates explicitly defined substructures, which are mapped to numerical identifiers (no hashing involved). The LINGO fingerprint[16] works directly with the SMILES strings (rather than the usual dependence on a molecular graph representation), by fragmenting the strings into overlapping substrings.

Table 1: Molecular fingerprints evaluated in this study. Abbreviations: Topological torsion (TT), Extended Connectivity Fingerprint (ECFP), Functional Class Fingerprint (FCFP), Atom pair (AP2D), Atom Triplet (AT2D), All Star Paths (ASP), Depth First Search (DFS).

TYPE	**FAMILY**	**DESCRIPTION**	**SIZE (bits)**
AP2D[33]	SUBSTRUCTURE	Topological Atom Pairs	4096
ASP[34]	PATH	All-Shortest Path encoding	4096
AT2D[34]	SUBSTRUCTURE	Topological Atom Triplets	4096
AVALON	SUBSTRUCTURE	Enumerates paths and feature classes	1024
DFS[35]	PATH	All-path encodings	4096
ECFP_0[36,37]	CIRCULAR	Extended-connectivity fingerprint of diameter 0	1024
ECFP_2[36,37]	CIRCULAR	Extended-connectivity fingerprint of diameter 2	1024
ECFP_4[36,37]	CIRCULAR	Extended-connectivity fingerprint of diameter 4	1024
ECFP_6[36,37]	CIRCULAR	Extended-connectivity fingerprint of diameter 6	1024
ESTATE[37,38]	SUBSTRUCTURE	Fingerprint based on E-State fragments	79
FCFP_0[36,37]	CIRCULAR	Feature-class fingerprint of diameter 0	1024
FCFP_2[36,37]	CIRCULAR	Feature-class fingerprint of diameter 2	1024
FCFP_4[36,37]	CIRCULAR	Feature-class fingerprint of diameter 4	1024
FCFP_6[36,37]	CIRCULAR	Feature-class fingerprint of diameter 6	1024
FP2[39]	PATH	Indexes linear fragments up to 7 atoms in length	–
FP3[39]	SUBSTRUCTURE	Based on 55 SMARTS patterns defining functional groups	–
FP4[39]	SUBSTRUCTURE	Based on SMARTS patterns defining functional groups	–
KR[37,40]	SUBSTRUCTURE	Klekota-Roth SMARTS based fingerprint	4860
LINGO[16,37]	TEXT	fragmentation of SMILES strings	–
LSTAR[34]	PATH	Local Path Environments	4096
MACCS[41]	SUBSTRUCTURE	Molecular ACCess System structural keys	166
MAP4[42]	CIRCULAR	Combines substructure and atom-pair concepts	2048
MHFP[43]	CIRCULAR	Encodes circular substructures	2048
P2PPHAR2D[44]	PHARMACOPHORE	Pharmacophore pair encoding	4096
P3PPHAR2D[44]	PHARMACOPHORE	Pharmacophore triplet encoding	4096
PUBCHEM[37,45]	SUBSTRUCTURE	Substructure fingerprint	881
RAD2D[19]	CIRCULAR	Topological Molprint-like fingerprints	4096
RDK5[46]	SUBSTRUCTURE	Encodes substructures at most 5 bonds long	1024
RDK6[46]	SUBSTRUCTURE	Encodes substructures at most 6 bonds long	1024
RDK7[46]	SUBSTRUCTURE	Encodes substructures at most 7 bonds long	1024
SIGNATURE[32,37]	SUBSTRUCTURE	Based on an array of atom signatures	–
TT[33]	PATH	Based on bond paths of four non-hydrogen atoms	–

All fingerprints were generated using open-source software. Routines in the RDKit[46] library were used to compute the AVALON, ERG, RDK5, RDK6, RDK7, MHFP, and TT fingerprints. FP2, FP3 and FP4 fingerprint similarities were calculated directly using the OpenBabel toolbox.[39] The other fingerprints were calculated using custom software that makes use of the *jCompoundMapper*[34] and Chemistry Development Kit[37] libraries.

Although a number of similarity metrics have been used,[47] the most common measure of fingerprint similarity is the Tanimoto coefficient:[48] $T_{ab} = |F_a \cap F_b|/|F_a \cup F_b|$, where F_a and F_b are the fingerprints of molecules a and b, respectively – this is equivalent to the Jaccard index over fingerprint bit vectors. Tanimoto coefficient values range between 1 (identical fingerprints, though not necessarily identical compounds) and 0 (disjoint fingerprints).

Benchmarking Data Sets

Numerous benchmarking data sets have been developed over the years to evaluate VS methods.[49,50] Each data set contains a set of active compounds (with known/documented activity for the target of interest) and a corresponding set of inactives/decoys. While the definition of actives is consistent, there is some variance in the question of what should be considered a 'decoy'. Some benchmarks include only confirmed inactive molecules, while others add compounds presumed to be non-binding.[51–53] Data set composition can impact VS evaluation, such that both artificial under- and over-estimation of enrichment have been well documented.[52,54,55]

In this study, we employ four different VS data sets to explore the utility of molecular fingerprinting strategies for prediction of similar activity. These data sets are briefly summarized in Table 2 and described here:

DUD-E Directory of Useful Decoys, Enhanced:[57] DUD-E is a widely-used data set for VS benchmarking, containing data for 102 protein targets. Each target is represented by an average of ∼224 active ligands, ∼90 experimental decoys, and ∼14,000 computational decoys. Compounds are considered active based on a 1 μM experimental affinity

Table 2: Comparison between different VS data sets. In all cases, the actives may not bind to the same pocket of the target. [a] [56]

Data set	Active source	Comments	Decoy generation	Comments
DUD-E	ChEMBL09	No rigorous method to remove false positives	0.65% from experiments. 99.35% generated choosing different topologies with similar chemical properties using 2D similarity methods.	Decoys biased towards 2D similarity methods.
MUV	PubChem BioAssay	Removal of false positives and assay artifacts	Choosing unbiased distribution of decoys from experimentally available data.	Low active to decoy ratio.
DEKOIS	DUD[a] (from literature)	Decoys generated choosing different topologies with similar chemical properties using 2D similarity methods.	Decoys biased towards 2D similarity methods.	Low active to decoy ratio.
LIT-PCBA	PubChem BioAssay	Removal of false positives and assay artifacts.	Decoys were chosen from experimentally available data and pruned to have similar chemical properties.	High active to decoy ratio. Actives may not bind to the same pocket of a target. Variable performance in 2D and 3D similarity search and docking across different target sets.

cutoff, and experimental decoys are ligands with no measurable affinity up to 30 μM. Computational decoy ligands are selected from ZINC[51] to have 50 physical properties (rotatable bonds, hydrogen bond acceptors/donors, molecular weight, logP, net charge) similar to the actives, but with low fingerprint (Daylight[58]) Tanimoto coefficient $T_{ab} < 0.5$. (Note: this means that computational decoys are, by construction, expected to have low fingerprint similarity).

MUV Maximum Unbiased Validation[59] MUV data sets are based on bioactivity data available in PubChem.[60] This benchmark consists of sets of 30 actives (taken from confirmation assays) and 15,000 decoys (drawn from corresponding primary screens) for each of the 17 targets. The goal of the experimental design is to obtain an optimal spread of actives in the chemical space of the decoys. Since the data are taken from high-throughput screening assays that can be affected by experimental noise and artifacts (caused by unspecific activity of chemical compounds), an assay filter is applied to remove compounds interfering with optical detection methods (autofluorescence and luciferase inhibition) and potential aggregators.

DEKOIS The Demanding Evaluation Kits for Objective In silico Screening (DEKOIS)[61] benchmark is based on BindingDB[62] bioactivity data (K_i, K_d, or IC_{50} values). The DEKOIS data set is derived from a set of 15 million molecules randomly selected from ZINC, which are divided into 10,752 bins based on their molecular weight (12 bins), octanol–water partition coefficient (8 bins), number of hydrogen bond acceptors (4 bins), number of hydrogen bond donors (4 bins), and number of rotatable bonds (7 bins). Active ligands are also placed into these pre-defined bins. For each active ligand, 1500 decoys are sampled from the active's bin (or neighboring bins, if necessary). These are further refined to a final set of 30 structurally diverse decoys per active. The DEKOIS data set includes 81 protein targets found in the DUD-E data set.

LIT-PCBA The LIT-PCBA benchmark[63] is a curated subset of the PubChem BioAssay

database, containing data from experiments where more than 10,000 chemicals were screened against a single protein target, and dose-response curves identified at least 50 actives. Active ligands identified in a bioassay experiment are not guaranteed to bind to the same pocket of the target protein; to overcome this concern, LIT-PCBA includes only targets with representative ligand-bound structures present in the PDB, such that the PDB ligands share the same phenotype or function as the true active ligands from the bioassay experiments. The LIT-PCBA data set was further refined to contain only targets for which at least one of the VS methods (2D fingerprint similarity, 3D shape similarity, and molecular docking) achieved an enrichment in true positives (i.e. the most challenging protein targets have been removed, so that enrichment results are, by design, expected to show good results). Targets in the LIT-PCBA have a variable active to decoy ratio that ranges from as low as 1:20 to 1:19000.

Virtual Screening Evaluation

A common measure for the efficacy of a method's discriminatory power depends on the receiver operating characteristic (ROC) curve, which plots the sensitivity of a method as a function of false labels.[64] If a classification method assigns better scores to all true matches (actives) than to any false matches (decoys), then the area under that curve (AUC) will be 1. A random classifier will have an AUC of 0.5.

AUC provides a measure of the sensitivity/specificity trade-off across the full sensitivity range, but medicinal chemists are typically more interested in early recognition of active molecules,[29,65] since there is little actionable value to gains in recall of true actives buried among thousands (or more) of decoys. As an example, consider an imaginary method that assigns the highest scores to 10% of all active molecules, then afterwards loses discriminative power and assigns essentially random scores to all remaining molecules (actives and decoys). The AUC for such a method would be not particularly good (roughly 0.6), even though the early enrichment (ranking 10% of actives with superior score to all decoys) provides some

experimental utility.

To address this shortcoming of ROC AUC, a number of other metrics have been devised to assess early enrichment.[29,66,67] Unfortunately, it can be difficult to extract an intuitive meaning from these measures,[68] and they are often not comparable across test sets because their scale and value depends on set size and number of decoys in the test set. Here, we introduce a simple new early enrichment measure, the *decoy retention factor* (DRF); DRF is easy to interpret, and generalizes across input size. We note that DRF is only applicable in situations in which the number of active and decoy ligands is known beforehand. For analysis of fingerprint benchmarks, we present both DRF and AUC values. Additional metrics such as BEDROC[65] and sum of log rank[29] are summarised visually in the Supplementary Information.

The purpose of DRF is to identify, for a parameterized fraction p of the active molecules, how effectively decoys are filtered from the score range containing those actives. Consider an input containing n active compounds and d decoys, and an enrichment threshold of $p = 0.1$. Since we are interested in the score of the top p fraction of actives, let $x = \lceil pn \rceil$, and let s_p be the score of the xth element (so that s_p is the score threshold that would recover 10% of actives). Define d_p to be the number of decoys that exceed s_p – this is a fraction of the d total decoys. DRF_p measures the extent to which decoys have been filtered out of the range containing the top p actives:

$$DRF_p = \frac{d_p}{\lfloor pd \rfloor} \tag{1}$$

A system that assigns scores randomly will recover a fraction p of the decoys at roughly the same score as it sees p of the actives, so that $\mathrm{DRF}_p = 1$. In an ideal case, no decoys have score greater than the xth active element, meaning that $d_p = 0$, and thus $\mathrm{DRF}_p = 0$. A $\mathrm{DRF}_p = 0.2$ indicates that the number of decoys remaining is 20% of the number expected if p of the decoys were kept (there is a 5-fold reduction in decoys). Meanwhile a $\mathrm{DRF}_p > 1$ indicates that the method *enriches* for decoys.

We find DRF to be a useful measure because it enables prediction of the number of decoys expected to remain in a score-filtered result set, based on the size of the underlying library. For example, consider a library of 1 million molecules – this will consist almost entirely of inactives (decoys), so that $d \approx 1,000,000$. If we hope to discover 10% of actives, and we have previously established that $\mathrm{DRF}_{0.1} = 0.05$ (a 20-fold reduction in decoys relative to random chance), then we expect to observe $d_p \cdot DRF_p \approx 1,000,000 \cdot 0.1 \cdot 0.05 = 5,000$ decoys mixed with the surviving actives. This simple calculation is important, because it can highlight that apparently-good enrichment (10- or 100-fold) may not be enough to effectively filter out inactives when the target set includes billions of candidates.

Results

To gain insight into the utility of various fingerprinting strategies for billion-scale virtual drug screening, we explored the capacity of fingerprint similarity to extract a small set of candidates that is highly enriched for molecules with activity similar to the seed query molecules. First, we computed measures of enrichment for 32 fingerprints on four benchmark data sets, presenting both classical ROC AUC calculations and our new decoy retention factor (DRF) scores. We then explored the distributions of fingerprint similarity scores across a variety of target molecules, and show that the score distributions for actives and decoys are not sufficiently separable for billion-scale search. We also found that, where fingerprints did occasionally find good matches among actives, the "new" actives almost always shared a scaffold with the query active (meaning that they would be found by more traditional shape exploration methods) We further considered whether there is a correlation between compound potency and active-active similarity scores, and found that there is not. Finally, we used a data set containing more than 300,000 experimentally-confirmed inactive compounds, and found that fingerprint similarity to an active molecule does not enable discrimination between actives and inactive. In total, these results indicate that fingerprint similarity is

not a reliable proxy for likely similar binding activity or particularly for discovering diverse active structures.

Enrichment for active molecules

To assess the utility of fingerprinting strategies for selecting compounds with similar expected activity, we computed similarities of all compounds to a query active molecule, and tested whether active molecules tend to be more similar to other actives than to decoys. Specifically, for each target protein, we computed the fingerprints of each molecule associated with that target protein. Then, for each active compound, we computed the similarity of its fingerprint to each of the other compounds (actives and decoys) affiliated with that target. The union of these distance calculations was merged and sorted by similarity. $DRF_{0.1}$ and ROC AUC were computed from these ordered lists.

Table 3 presents the resulting enrichment values on each benchmark data set. The performance of all fingerprints is poor for both the MUV and LIT-PCBA data sets, with AUC values generally <0.6, and $DRF_{0.1}$ values close to 1.0. Performance is somewhat better on DEKOIS and DUD-E, but not particularly strong, and is offset by concerns of benchmark bias highlighted elsewhere, such as artificial enrichment[50,69–73] (enrichment due to bias in the actives or decoys), analogue bias (limited diversity of the active molecules), and false negative bias (risk of active compounds being present in the decoy set), all of which can cause misleading VS results.[52,72,74] Table 3 also provides a summary of the VS performances obtained for the fingerprint types (substructure, circular, path, text, pharmacophore). No particular fingerprint strategy appears to be better suited to the problem of virtual screening. Among all tests, the best observed $DRF_{0.1}$ value was 0.09, equating to a roughly 11-fold enrichment in actives versus inactives; in a screen of a billion molecules, this means that roughly 9 million inactive molecules are expected to show similarity at the score threshold where 10% of actives are retained)

Most circular and path-based fingerprints employ a standard length of 1024 bits. O'Boyle

Table 3: Summary of the VS performances in terms of the AUC and DRF ($p = 0.1$) for the 32 fingerprints tested on the DEKOIS, DUDE, MUV and LIT-PCBA data sets.See also heatmaps of the different metrics for the data sets in Figures F1–F5 in the Supplementary Information.

FP	**AUC**				**DRF**			
	DEKOIS	**DUDE**	**MUV**	**LIT-PCBA**	**DEKOIS**	**DUDE**	**MUV**	**LIT-PCBA**
AP2D	0.64	0.66	0.49	0.51	0.55	0.39	1.24	1.00
AT2D	0.78	0.79	0.58	0.55	0.20	0.12	0.86	0.83
AVALON	0.72	0.73	0.60	0.55	0.30	0.18	0.85	0.97
ESTATE	0.71	0.75	0.53	0.50	0.31	0.17	0.98	0.94
FP3	0.68	0.77	0.51	0.52	0.45	0.20	0.89	0.83
FP4	0.74	0.80	0.58	0.54	0.30	0.12	0.89	0.91
MACCS	0.71	0.75	0.55	0.54	0.33	0.18	0.99	0.93
PUBCHEM	0.76	0.76	0.55	0.54	0.28	0.20	1.06	0.97
RDK5	0.76	0.75	0.58	0.56	0.24	0.16	0.90	0.89
RDK6	0.70	0.70	0.59	0.58	0.38	0.29	0.86	0.82
RDK7	0.62	0.63	0.58	0.59	0.74	0.70	0.98	0.85
KR	0.72	0.74	0.54	0.51	0.31	0.19	1.05	0.96
SIGNATURE	0.72	0.72	0.55	0.54	0.33	0.23	0.97	0.99
SUBSTRUCTURE	0.71	0.73	0.56	0.54	0.36	0.24	0.96	0.91
ASP	0.80	0.79	0.58	0.53	0.17	0.12	0.89	0.99
DFS	0.79	0.78	0.55	0.52	0.18	0.14	0.98	1.01
FP2	0.79	0.78	0.55	0.54	0.20	0.14	1.01	0.90
LSTAR	0.78	0.78	0.54	0.51	0.19	0.13	0.97	1.03
TT	0.80	0.80	0.61	0.56	0.15	0.10	0.72	0.85
PATH	0.75	0.75	0.57	0.54	0.18	0.13	0.91	0.96
ECFP0	0.70	0.77	0.53	0.50	0.33	0.13	0.97	0.96
ECFP2	0.77	0.81	0.54	0.51	0.19	0.09	0.99	1.01
ECFP4	0.76	0.80	0.54	0.51	0.19	0.09	0.99	1.00
ECFP6	0.75	0.78	0.54	0.52	0.20	0.10	0.99	0.98
FCFP0	0.66	0.69	0.54	0.52	0.35	0.23	0.41	0.42
FCFP2	0.76	0.75	0.55	0.52	0.24	0.19	0.93	1.01
FCFP4	0.78	0.76	0.54	0.52	0.20	0.15	0.93	1.01
FCFP6	0.78	0.75	0.54	0.52	0.20	0.15	0.96	0.98
MAP4	0.81	0.83	0.56	0.54	0.14	0.07	0.91	0.85
MHFP	0.81	0.81	0.54	0.53	0.17	0.10	0.97	0.94
RAD2D	0.76	0.77	0.53	0.53	0.23	0.14	0.99	0.93
CIRCULAR	0.76	0.77	0.54	0.52	0.26	0.11	0.98	0.99
P2PPHAR2D	0.66	0.74	0.51	0.54	0.50	0.25	1.20	0.94
P3PPHAR2D	0.71	0.76	0.52	0.55	0.33	0.16	1.17	0.93
PHARMACOPHORE	0.75	0.76	0.54	0.53	0.42	0.21	1.19	0.94
LINGO	0.77	0.79	0.54	0.54	0.21	0.10	1.02	0.91

and Sayle[75] suggested that increasing the bit-vector length from 1024 to 16384 can improve VS performance, though at a cost of space and run time for comparison. We evaluated the utility of longer fingerprints for the MUV and LIT-PCBA data sets, and found that longer fingerprints yield little to no gain in efficacy (see Table S1 in Supplementary Information).

Tanimoto Similarity Distributions Are Generally Indistinguishable

To explore the distribution of similarities between actives and decoys, we computed Tanimoto coefficients for active-active and active-decoy molecule pairs in the DEKOIS data set. For each target protein in DEKOIS,[61] we randomly selected an active molecule, and computed the molecular fingerprint Tanimoto similarity to all other actives and decoys for that target. Figure 1 shows the resulting score distributions for 32 fingerprints. In DEKOIS, the distributions of active-active (blue) and active-decoy (red) Tanimoto values substantially overlap, so that the vast majority of actives fall into a score range shared by most decoys.

Most of the fingerprints in Figure 1 present a thin high-Tanimoto tail for actives (blue) that is not seen for decoys (red), suggesting that perhaps a small fraction of actives could be discriminated from decoys by establishing a high score threshold – in other words, it may be possible to select a threshold that delineates regions of early enrichment. However, consider the ECFP2 fingerprint, which shows an apparently compelling right tail in the active-active plot (blue), such that it appears to be reasonable to establish Tanimoto cutoff of 0.5. In DEKOIS, there are 423 active matches to active queries above this threshold. Though the right tail of the active-decoy distribution (red) is imperceptible in this plot, it still contains $\sim$0.0064% of the decoys. Extrapolating to a library of 3.7 billion candidates, as we used in Venkatraman et al.[2], we expect to see $\sim$23.7M decoys with Tanimoto $\geq$ 0.5, so that the active-to-decoy ratio is $\sim$1:56,000. Setting the Tanimoto threshold to 0.75 leads to an expected ratio of $\sim$1:68,000 (57 actives to $\sim$3.9M expected decoys). This is likely an overly pessimistic view of the decoy ratio risk, since the compounds in the decoy set are intended to be more similar to the actives than would be a random molecule from a billion-scale library. Even so, it highlights the fact that even small false discovery rates can produce an overwhelmingly large number of false matches when the target base is sufficiently large. Moreover, see the next section for exploration of the expected novelty of these high-scoring matches.

Figure 1: Ridgeline plots showing the distribution of the Tanimoto fingerprint similarities calculated between a randomly-selected active molecule for each target protein and all other actives (shown in blue) and decoys (in red) for that target. Data taken from the DEKOIS data set. The distribution of similarity scores between the active query and other active molecules is largely indistinguishable from the distribution of similarity scores to random molecules. Where the active (blue) distribution does show a fatter high-scoring tail than the inactive distribution (suggesting potential for early enrichment by using a high score threshold), search against a large target database will still produce filtered sets that are massively dominated by inactives (see text).

Tanimoto score distributions, and their utility in scaffold hopping

The LIT-PCBA data set[63] contains binding activity data for 15 proteins, each with 13 to 1,563 confirmed active molecules and 3,345 to 358,757 confirmed inactives/decoys (see Table 4. We computed the ECFP4 fingerprint for each of the 2.6M active and inactive small molecules. For each protein, we first selected the molecule with best affinity to serve as the 'active query', then computed the Tanomito similarity measure (T_c) between seed and all other molecules (actives and decoys) for that protein. Figure 2 presents T_c distributions. Each plot shows, for a single protein, the fraction of actives (blue) and decoys (red) that are recovered (Y-axis) as a function of T_c (X-axis). One important observation is that *for most proteins*, the molecule with highest similarity (active or decoy) to the active query has a $T_c < 0.5$, which is below commonly-used thresholds for expected shared bioactivity.[76–78] In addition, for most proteins, T_c distributions are indistinguishable, though three proteins (GBA, OPRK1, and PPARG) demonstrate early enrichment for actives. We manually inspected all high-scoring actives ($T_c > 0.5$) for these three enriched proteins (see Figure 3 and supplementary Figures F6 and F7), and observed that high-scoring actives were almost entirely bioisosteres, easily predictable results of exchanging an atom or atom-group with a similar atom or group. The lack of scaffold diversity among matches with scores discernible from high-volume decoy noise casts doubt on the utility of fingerprint similarity for exploring novel candidate drug space.

Evaluation on a target with many validated inactive molecules

The previous experiments depend on benchmarks containing computationally-identified decoys that almost entirely have not been experimentally validated as inactive. The MMV St. Jude malaria data set[79] provides an alternative perspective on the utility of fingerprint similarity for activity prediction in the context of verified decoys. It contains a set of 305,810 compounds that were assayed for malaria blood stage inhibitory activity. Among these molecules, 2507 were classified as active, while the remaining 303,303 compounds were

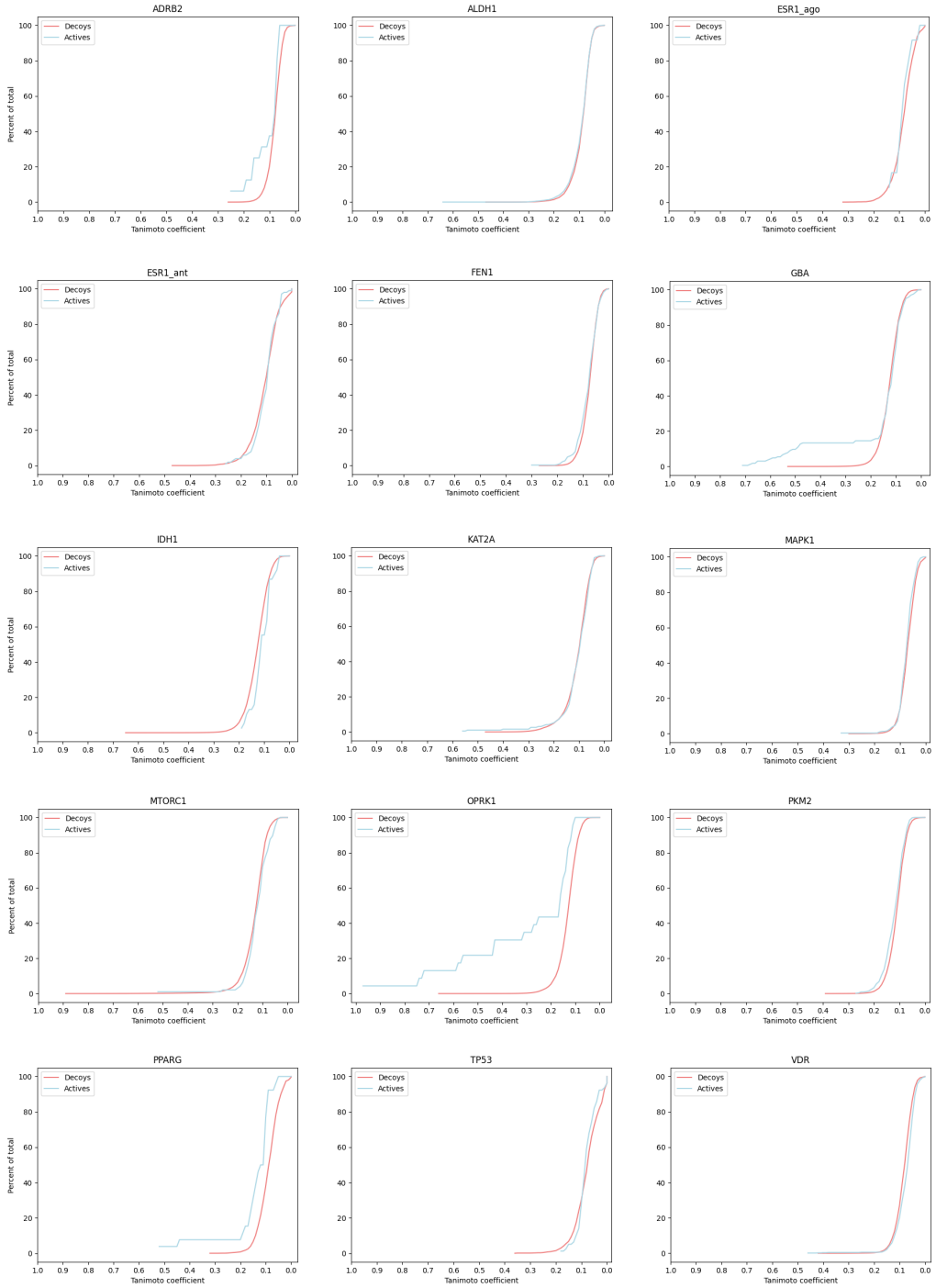

Figure 2: Cumulative percent of total actives (blue) and decoys (red) encountered (Y-axis) as a function of decreasing Tanimoto coefficient, using ECFP4 fingerprints. For each protein, the molecule with best binding affinity was used as the query molecule (the molecule for which the Tanimoto score was computed for each other molecule, active or inactive).

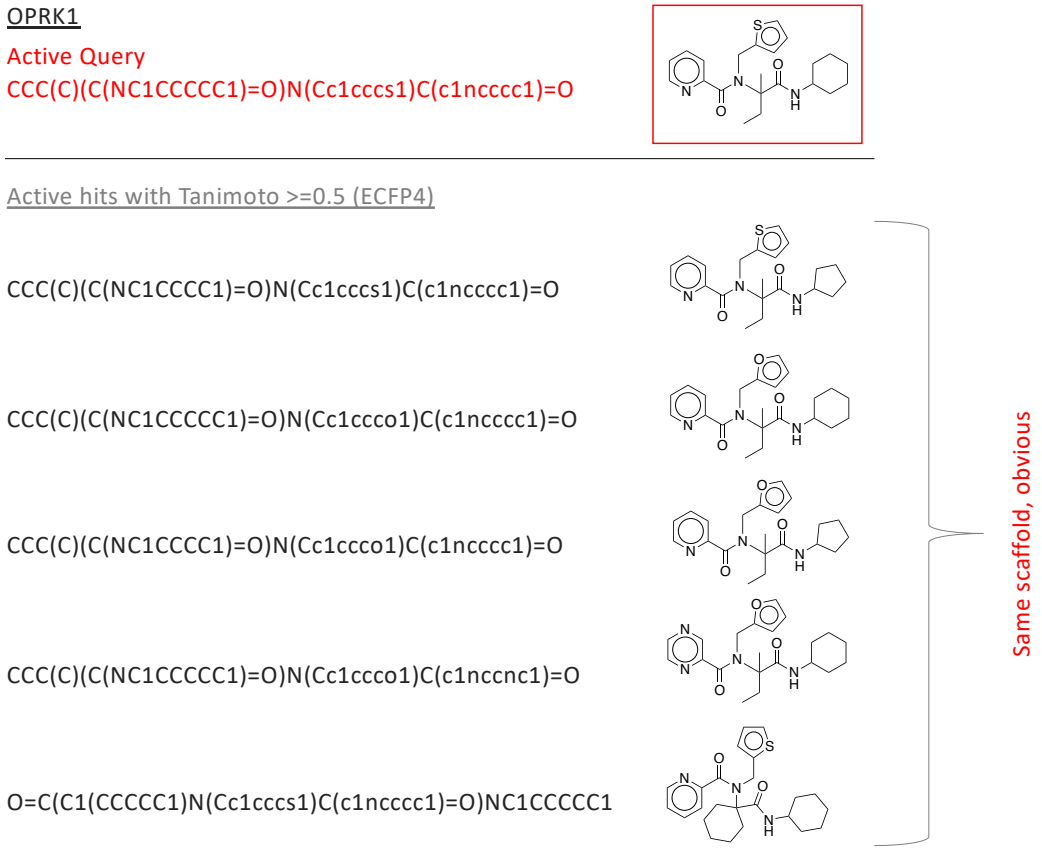

Figure 3: In the LIT-PCBA analysis, early enrichment was observed for OPRK1 – for all Tanimoto coefficients $t >\sim 0.2$, the fraction of actives with Tanimoto score $> t$ is much larger than the fraction of decoys with that score (see Figure 2). This suggests that fingerprints produce useful early enrichment of active molecules, at least for this one protein target. We sought to understand if the high-scoring actives represented novel drug candidates that could not be easily identified by simple modifications to the active drug used as a fingerprint query There are only 24 actives in the OPRK1 data set, and only 5 of these showed Tanimoto score > 0.5 to the initial query; we manually inspected the structures of these compounds. All 5 are built on the same scaffold as the query (in red), and are obvious variations that should be identified through standard enumeration (i.e. no new scaffold are explored). Similar plots are provided for GBA and PPARG in Supplementary Figures F6 and F7.

Table 4: LIT-PCBA provides sets of confirmed actives and inactives for 15 protein binding partners.

Protein	Actives	Inactives
ADRB2	17	311,748
ALDH1	5,363	101,874
ESR_ago	13	4,378
ESR_antago	88	3,820
FEN1	360	350,718
GBA	163	291,241
IDH1	39	358,757
KAT2A	194	342,729
MAPK1	308	61,567
MTORC1	97	32,972
OPRK1	24	269,475
PKM2	546	244,679
PPARG	24	4,071
TP53	64	3,345
VDR	655	262,648

classified as inactive.

For each active molecule, we computed Tanimoto similarity to each other active and to each inactive. Figure 4 shows bar plots for each fingerprint, with each plot showing the fraction of inactives (red) and actives (blue) with Tanimoto similarity values $T_{ab} \geq c$ for values of c = (0.1, 0.2, ... , 0.9) and 0.99. In general, the remaining fraction of actives only slightly exceeds the remaining fraction of inactives, indicating minimal enrichment of actives at increased Tanimoto similarity values. MAP4 shows an apparent relative abundance of actives, but note enrichment is still only ∼10-fold; also note that <1.5% of actives show Tanimoto similarity >0.1 to another active, raising concerns about how to establish meaningful MAP4 thresholds.

One important caveat for this dataset is that the target binding pocket of each active is unknown, so that it is possible that some actives target one pocket while other actives target another pocket. Even so, the lack of visible signal of predictive enrichment from fingerprint similarity is notable.

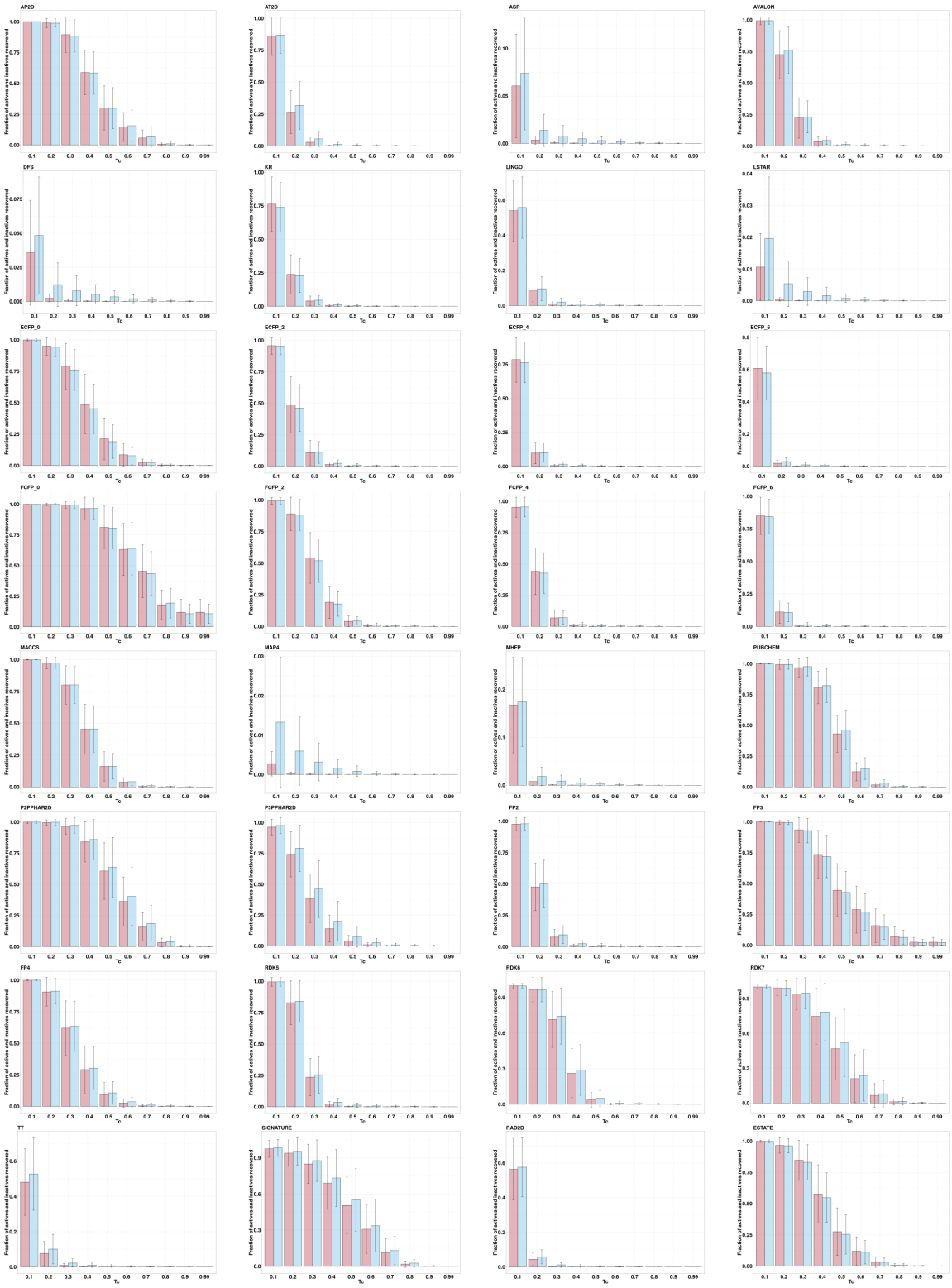

Figure 4: For compounds in the St. Jude malaria data set, bar plots show the fraction of the inactives (in red) and actives (in blue) exceeding Tanimoto similarity cutoffs by the different fingerprints. Tanimoto similarities were calculated using each active as the query; mean and standard deviation (based on the 2507 actives) are shown as error bars.

Fingerprint similarity values do not correlate with compound potency

The previous sections demonstrate that fingerprint similarity has limited utility in discriminating active molecules from decoys. An alternative use of fingerprints could be to take a set of candidates that have already (somehow) been highly enriched for active compounds, and rank them according to expected potency. The LIT-PCBA data set provides a measure of molecule potency for each active (specifically, the concentration at which the compound exhibits half-maximal efficacy, AC_{50} μM), and therefore provides a mechanism for evaluating this value proposition.

For each target protein in the LIT-PCBA data set, we selected the most potent active molecule, and computed fingerprint similarities for all other actives for the corresponding target. We evaluated the correlation of fingerprint similarity value to observed AC_{50} by computing the Kendall rank correlation.[80] Figure 5 presents a heatmap of these correlation values (τ) for each fingerprint across 15 protein targets, and demonstrates that all fingerprints exhibit poor correlation, with values ranging between -0.53 to 0.54, and generally only slightly higher than zero. This indicates that the fingerprints evaluated are unlikely to yield a ranked set of enriched highly potent compounds, in agreement with the observations of Vogt and Bajorath[81].

To give insight into these summary statistics represented by the heatmap squares in Figure 5, we prepared scatter plots corresponding to three of the target/fingerprint pairings (Figure 6). The middle plot shows the relationship between fingerprint similarity and AC_{50} values for the target/fingerprint pair with median Kendall correlation (target=VDR, fingerprint=SIGNATURE), and is representative of most of the target/fingerprint pairs; it shows essentially no correlation between fingerprint similarity and AC_{50} values ($\tau = 0.01$). The first and last scatter plots show fingerprint similarity and AC_{50} values for the target/fingerprint with the highest (ADRB2, FFCP0) and lowest (PPARG, ASP) correlation values. Note that Kendall rank correlation values for FCFP0 with targets other than ADRB2 ($\tau = 0.54$) vary

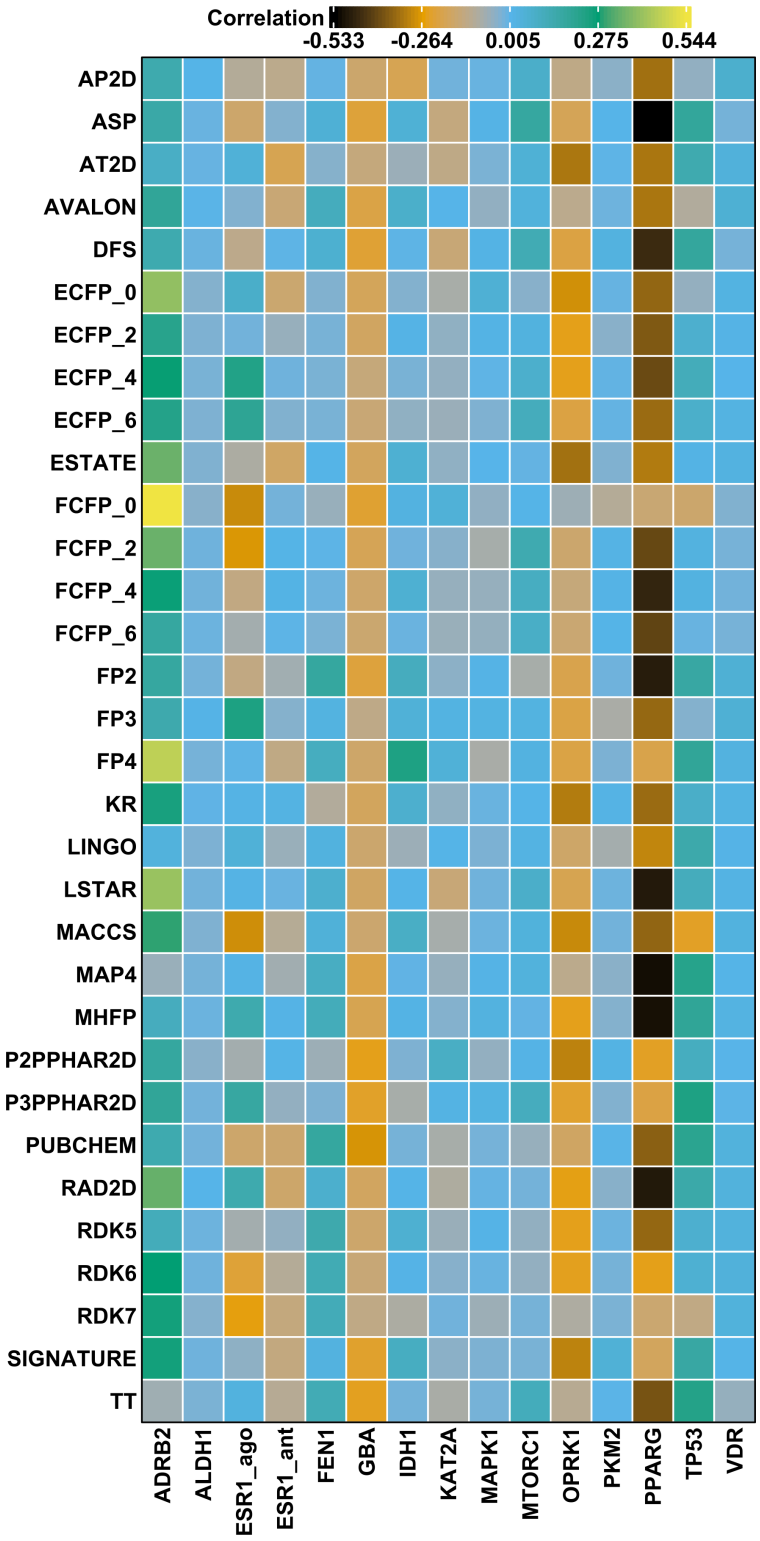

Figure 5: Heatmap of the Kendall rank correlation (τ) between fingerprint Tanimoto (T_c) similarities calculated between the most active compound for a given target and the potency values (AC_{50}) of the actives for that target.

from -0.30 to 0.05 and for ASP with targets other than PPARG ($\tau = -0.53$) from -0.23 to 0.18. Though some protein/fingerprint combinations show apparent correlation between fingerprint similarity and AC_{50}, most show a near total lack of correlation; deviations (positive or negative correlation) are observed in proteins with a small number of actives, which are thus more prone to stochastic or biased observations of correlation.

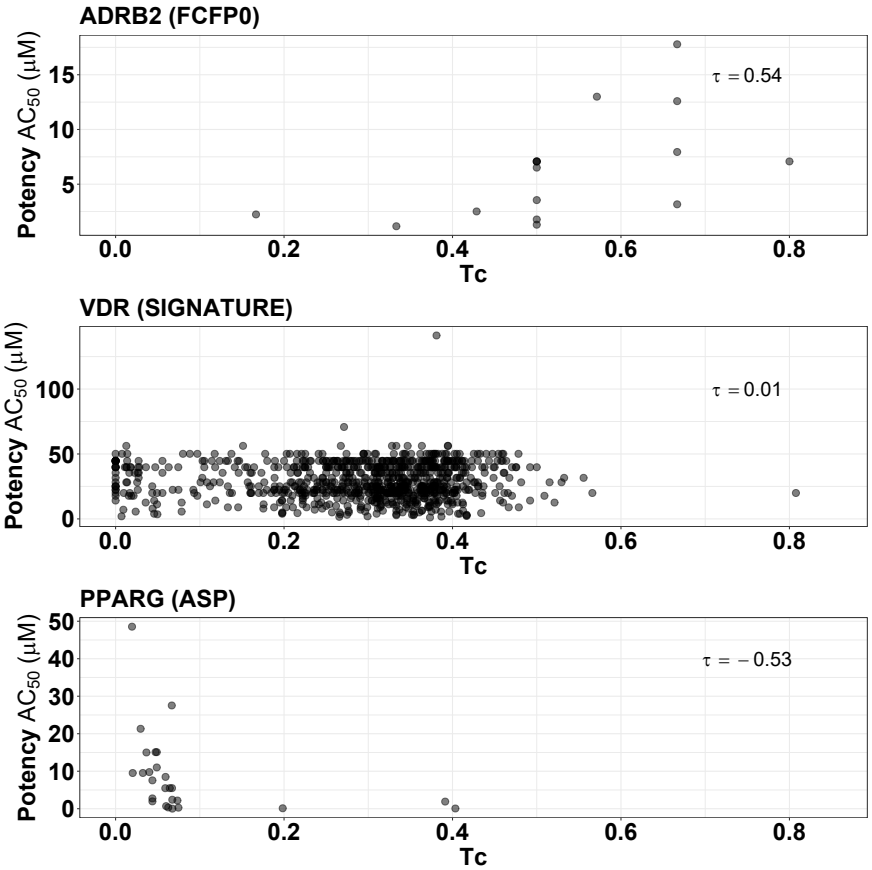

Figure 6: Scatter plot of the fingerprint similarities (T_c) and the potencies (AC_{50}) of active compounds for ADRB2 (using FCFP0 fingerprint), VDR (SIGNATURE fingerprint) and PPARG (ASP fingerprint).

Discussion

Computational methods hold the promise of expanding the diversity of small molecule drug candidate targeting for specified protein binding pockets. To fulfill this promise, these methods must be supported by succinct molecular representation schemes that enable rapid iden-

tification of molecules that are functionally similar to known or suspected active molecules while presenting non-trivial molecular similarity. The results of this study demonstrate that molecular fingerprints, and specifically measurement of molecular similarity based on those fingerprints, are not generally effective at discriminating molecules with similar binding activity from those with dissimilar activity; furthermore, we have shown that where some early enrichment is observed, it tends to highlight candidate molecules that would arise from a simple structural perturbation. This suggests that the task of rapid evaluation of molecular similarity must move beyond fingerprint representation of molecules. In recent years, deep learning strategies have entered the drug discovery toolkit (e.g.[82–84]), but these have not yet solved the problem of rapid virtual screening. Though the path forward is not clear, we suggest that it is vital that molecules be represented in such a way that the potential *context* of the molecule (i.e. information about the potential binding target) can be considered when evaluating the similarity of molecules. We suspect that future successful strategies will emphasize the surface properties of the small molecule,[85,86] and will represent both the compound and the target protein not as a monolith, but as a collection of surface patches.[87–89] These, we believe, will enable a more context-dependent emphasis on features of importance to particular interactions, without interference from unimportant features.

Data and Software Availability

Data and software used for the calculation of fingerprints and scripts to reproduce the results are available from `https://osf.io/d3cbr/`.

Acknowledgement

We gratefully acknowledge the computational resources and expert administration provided by the University of Montana's Griz Shared Computing Cluster (GSCC) and the high performance computing (HPC) resources supported by the University of Arizona TRIF, UITS,

and Research, Innovation, and Impact (RII) and maintained by the UArizona Research Technologies department.

VV was supported by the Research Council of Norway through grant no. 275752. AR was supported by the NIH National Institute of Allergy and Infectious Diseases (NIAID), Department of Health and Human Services under BCBB Support Services Contract HHSN316201300006W/HHSN27200002 to MSC, Inc. JLG was supported by the Federal Pathway Program at NIH, NIAID. RX was supported by NIH National Institute of Allergy and Infectious Diseases (NIAID) grant AI168165. TJW and DD were supported by NIH National Institute of General Medical Sciences (NIGMS) grant GM132600, and by the DOE Biological and Environmental Research (BER) Program, grant DE-SC0021216.

Supporting Information Available

The supporting information includes additional plots.

References

(1) Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X.-P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F., et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **2022**, *601*, 452–459.

(2) Venkatraman, V.; Colligan, T. H.; Lesica, G. T.; Olson, D. R.; Gaiser, J.; Copeland, C. J.; Wheeler, T. J.; Roy, A. Drugsniffer: An Open Source Workflow for Virtually Screening Billions of Molecules for Binding Affinity to Protein Targets. *Front. Pharmacol.* **2022**, *13*.

(3) Luttens, A.; Gullberg, H.; Abdurakhmanov, E.; Vo, D. D.; Akaberi, D.; Talibov, V. O.; Nekhotiaeva, N.; Vangeel, L.; De Jonghe, S.; Jochmans, D., et al. Ultralarge virtual

screening identifies SARS-CoV-2 main protease inhibitors with broad-spectrum activity against coronaviruses. *J. Am. Chem. Soc.* **2022**, *144*, 2905–2920.

(4) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of ultralarge compound collections for drug discovery. *J. Chem. Inf. Model.* **2022**, *62*, 2021–2034.

(5) Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **2018**, *62*, 1116–1124.

(6) Gimeno, A.; Ojeda-Montes, M.; Tomás-Hernández, S.; Cereto-Massagué, A.; Beltrán-Debón, R.; Mulero, M.; Pujadas, G.; Garcia-Vallvé, S. The Light and Dark Sides of Virtual Screening: What Is There to Know? *Int. J. Mol. Sci.* **2019**, *20*, 1375.

(7) Maia, E. H. B.; Assis, L. C.; de Oliveira, T. A.; da Silva, A. M.; Taranto, A. G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front. Chem.* **2020**, *8*.

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

(9) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2013**, *57*, 3186–3204.

(10) Cortés-Ciriano, I.; Škuta, C.; Bender, A.; Svozil, D. QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. *J. Cheminf.* **2020**, *12*.

(11) Venkatraman, V. FP-MAP: an extensive library of fingerprint-based molecular activity prediction tools. *Frontiers in Chemistry* **2023**, *11*.

(12) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME–Tox predictions? *Drug Discov. Today* **2006**, *11*, 700–707.

(13) Mellor, C.; Robinson, R. M.; Benigni, R.; Ebbrell, D.; Enoch, S.; Firman, J.; Madden, J.; Pawar, G.; Yang, C.; Cronin, M. Molecular fingerprint-derived similarity mea-

sures for toxicological read-across: Recommendations for optimal use. *Regul. Toxicol. Pharmacol.* **2019**, *101*, 121–134.

(14) Venkatraman, V. FP-ADMET: a compendium of fingerprint-based ADMET prediction models. *J. Cheminf.* **2021**, *13*.

(15) Teixeira, A. L.; Falcao, A. O. Structural Similarity Based Kriging for Quantitative Structure Activity and Property Relationship Modeling. *J. Chem. Inf. Model.* **2014**, *54*, 1833–1849.

(16) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386–393.

(17) Kubinyi, H. Similarity and dissimilarity: a medicinal chemist's view. *Perspectives in Drug Discovery and Design* **1998**, *9*, 225–252.

(18) Barbosa, F.; Horvath, D. Molecular Similarity and Property Similarity. *Curr. Top. Med. Chem.* **2004**, *4*, 589–600.

(19) Bender, A.; Glen, R. C. Molecular similarity: a key technique in Mol. Inf. *Org. Biomol. Chem.* **2004**, *2*, 3204.

(20) Muratov, E. N. et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525–3564.

(21) Willett, P. *Methods in Molecular Biology*; Humana Press, 2010; pp 133–158.

(22) Stumpfe, D.; Bajorath, J. Similarity searching. *WIRES Comput. Mol. Sci.* **2011**, *1*, 260–282.

(23) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63.

(24) Gao, K.; Nguyen, D. D.; Sresht, V.; Mathiowetz, A. M.; Tu, M.; Wei, G.-W. Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **2020**, *22*, 8373–8390.

(25) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.

(26) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

(27) Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255–263.

(28) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **2007**, *6*, 881–890.

(29) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079–2093.

(30) Sciabola, S.; Torella, R.; Nagata, A.; Boehm, M. Critical Assessment of State-of-the-Art Ligand-Based Virtual Screening Methods. *Mol. Inf.* **2022**, 2200103.

(31) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49*, 108–119.

(32) Faulon, J.-L.; Visco, D. P.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Model.* **2003**, *43*, 707–720.

(33) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Model.* **1985**, *25*, 64–73.

(34) Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Zell, A. jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints. *J Cheminf.* **2011**, *3*.

(35) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093–1110.

(36) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(37) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminf.* **2017**, *9*.

(38) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Model.* **1995**, *35*, 1039–1045.

(39) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*.

(40) Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **2008**, *24*, 2518–2525.

(41) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Model.* **2002**, *42*, 1273–1280.

(42) Capecchi, A.; Probst, D.; Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminf.* **2020**, *12*.

(43) Probst, D.; Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *J. Cheminf.* **2018**, *10*.

(44) Mahé, P.; Ralaivola, L.; Stoven, V.; Vert, J.-P. The Pharmacophore Kernel for Virtual Screening with Support Vector Machines. *J. Chem. Inf. Model.* **2006**, *46*, 2003–2014.

(45) PubChem Substructure Fingerprint. `ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/`, 2022; Version: 1.3.

(46) Landrum, G. RDKit: Open-source cheminformatics. `https://www.rdkit.org`, 2022; Release: 2022.03.5.

(47) Raymond, J. W.; Willett, P. *J. Comput. Aided Mol. Des.* **2002**, *16*, 59–71.

(48) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*.

(49) Lagarde, N.; Zagury, J.-F.; Montes, M. Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives. *J. Chem. Inf. Model.* **2015**, *55*, 1297–1307.

(50) Réau, M.; Langenfeld, F.; Zagury, J.-F.; Lagarde, N.; Montes, M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **2018**, *9*.

(51) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput. Aided Mol. Des.* **2008**, *22*, 193–199.

(52) Réau, M.; Langenfeld, F.; Zagury, J.-F.; Lagarde, N.; Montes, M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **2018**, *9*.

(53) Tran-Nguyen, V.-K.; Rognan, D. Benchmarking Data Sets from PubChem BioAssay Data: Current Scenario and Room for Improvement. *Int. J. Mol. Sci.* **2020**, *21*, 4380.

(54) Nisius, B.; Bajorath, J. Rendering Conventional Molecular Fingerprints for Virtual Screening Independent of Molecular Complexity and Size Effects. *ChemMedChem* **2010**, *5*, 859–868.

(55) Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. *J. Cheminf.* **2016**, *8*.

(56) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *Journal of medicinal chemistry* **2006**, *49*, 6789–6801.

(57) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.

(58) Daylight Chemical Information Systems, I. Daylight Theory Manual. `https://www.daylight.com/dayhtml/doc/theory`, 2011; Version 4.9.

(59) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169–184.

(60) Wheeler, D. L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2007**, *36*, D13–D21.

(61) Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public Library of Challenging Docking Benchmark Sets. *J. Chem. Inf. Model.* **2013**, *53*, 1447–1462.

(62) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2015**, *44*, D1045–D1053.

(63) Tran-Nguyen, V.-K.; Jacquemard, C.; Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4263–4273.

(64) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.

(65) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(66) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput. Aided Mol. Des.* **2008**, *22*, 141–146.

(67) Lopes, J. C. D.; dos Santos, F. M.; Martins-José, A.; Augustyns, K.; Winter, H. D. The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability. *J. Cheminf.* **2017**, *9*.

(68) Zhao, W.; Hevener, K. E.; White, S. W.; Lee, R. E.; Boyett, J. M. A statistical framework to evaluate virtual screening. *BMC Bioinf.* **2009**, *10*.

(69) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Model.* **2004**, *44*, 793–806.

(70) Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947–961.

(71) Li, H.; Sze, K.-H.; Lu, G.; Ballester, P. J. Machine-learning scoring functions for structure-based drug lead optimization. *WIRES Comput. Mol. Sci.* **2020**, *10*.

(72) Imrie, F.; Bradley, A. R.; Deane, C. M. Generating property-matched decoy molecules using deep learning. *Bioinformatics* **2021**, *37*, 2134–2141.

(73) Stein, R. M.; Yang, Y.; Balius, T. E.; O'Meara, M. J.; Lyu, J.; Young, J.; Tang, K.; Shoichet, B. K.; Irwin, J. J. Property-Unmatched Decoys in Docking Benchmarks. *J. Chem. Inf. Model.* **2021**, *61*, 699–714.

(74) Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS ONE* **2019**, *14*, e0220113.

(75) O'Boyle, N. M.; Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminf.* **2016**, *8*.

(76) Baldi, P.; Nasr, R. When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *Journal of chemical information and modeling* **2010**, *50*, 1205–1222.

(77) O Hagan, S.; Swainston, N.; Handl, J.; Kell, D. B. A 'rule of 0.5'for the metabolite-likeness of approved pharmaceutical drugs. *Metabolomics* **2015**, *11*, 323–339.

(78) Vogt, M.; Bajorath, J. Modeling tanimoto similarity value distributions and predicting search results. *Molecular Informatics* **2017**, *36*, 1600131.

(79) Verras, A.; Waller, C. L.; Gedeck, P.; Green, D. V. S.; Kogej, T.; Raichurkar, A.; Panda, M.; Shelat, A. A.; Clark, J.; Guy, R. K.; Papadatos, G.; Burrows, J. Shared Consensus Machine Learning Models for Predicting Blood Stage Malaria Inhibition. *J. Chem. Inf. Model.* **2017**, *57*, 445–453.

(80) Kendall, M. G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93.

(81) Vogt, M.; Bajorath, J. Similarity Searching for Potent Compounds Using Feature Selection. *J. Chem. Inf. Model.* **2013**, *53*, 1613–1619.

(82) Yang, Y.; Yao, K.; Repasky, M. P.; Leswing, K.; Abel, R.; Shoichet, B. K.; Jerome, S. V. Efficient exploration of chemical space with docking and deep learning. *Journal of Chemical Theory and Computation* **2021**, *17*, 7106–7119.

(83) Gorantla, R.; Kubincova, A.; Weiße, A. Y.; Mey, A. S. From Proteins to Ligands:

Decoding Deep Learning Methods for Binding Affinity Prediction. *Journal of Chemical Information and Modeling* **2023**,

(84) Gentile, F.; Yaacoub, J. C.; Gleave, J.; Fernandez, M.; Ton, A.-T.; Ban, F.; Stern, A.; Cherkasov, A. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nature Protocols* **2022**, *17*, 672–697.

(85) Pérez-Nueno, V. I.; Venkatraman, V.; Mavridis, L.; Clark, T.; Ritchie, D. W. Using Spherical Harmonic Surface Property Representations for Ligand-Based Virtual Screening. *Molecular Informatics* **2010**, *30*, 151–159.

(86) Pérez-Nueno, V. I.; Venkatraman, V.; Mavridis, L.; Ritchie, D. W. Detecting Drug Promiscuity Using Gaussian Ensemble Screening. *Journal of Chemical Information and Modeling* **2012**, *52*, 1948–1961.

(87) Hofbauer, C.; Lohninger, H.; Aszódi, A. SURFCOMP: A Novel Graph-Based Approach to Molecular Surface Comparison. *J. Chem. Inf. Model.* **2004**, *44*, 837–847.

(88) Gainza, P.; Sverrisson, F.; Monti, F.; Rodolà, E.; Boscaini, D.; Bronstein, M. M.; Correia, B. E. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **2019**, *17*, 184–192.

(89) Douguet, D.; Payan, F. sensaas: Shape-based Alignment by Registration of Colored Point-based Surfaces. *Mol. Inf.* **2020**, *39*, 2000081.