Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Molecular fingerprints are not useful in large-scale search for similarly active compounds†

Vishwesh Venkatraman,*[a] Jeremiah Gaiser,[b,c], Amitava Roy,[d] and Travis J. Wheeler[b,e]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Computational approaches for small molecule drug discovery now regularly scale to consideration of libraries containing billions of candidate small molecules. One promising path to increased speed in evaluating billion-molecule libraries is to develop representations of each molecule that enable fast computation of similarity between molecules. Molecular fingerprints have long provided a mechanism for succinct representation and fast comparison of small molecules, with a large collection of competing fingerprints. Here, we explore the utility of many of these fingerprints in the context of predicting similar molecular activity. We show that fingerprint similarity enables insufficient discriminative power between active and inactive molecules for a target protein based on a known active. We also demonstrate that, 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.

1 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 such a search strategy. The SPP is simple and intuitive, and has served as the basis for predictions of biological activity [10], toxicity [11–13], aqueous solubility [13,14] (log$S$), and partition coefficient [15] (log$P$); it is, however, difficult to assess objectively, and may not necessarily reflect the chemical similarity from a medicinal chemist's perspective [16]. Furthermore, the 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 [17–19] (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 [20–22]. Such a fingerprint is computed for each molecule, then the fingerprints of molecules are compared for overlap to approximately assess molecular similarity. Fingerprint similarity has been used to effectively estimate log$S$ and log$P$ values [23]. This success is attributed to the fact that these values can largely be approximated from the small molecule itself without explicitly considering interacting partners.

Other molecular properties involve a greater 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) are unique for 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, limiting the inference power provided by

[a] *Norwegian University of Science and Technology, Department of Chemistry, 7491 Trondheim, Norway. E-mail:vishwesh.venkatraman@ntnu.no*
[b] *Department of Computer Science, University of Montana, Missoula, MT, USA.*
[c] *Department of Biochemistry and Biophysics, University of Montana, Missoula, MT, USA.*
[d] *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 59840, USA.*
[e] *Department of Pharmacy Practice and Science, University of Arizona, Tucson, AZ, USA.*

† Electronic Supplementary Information (ESI) available. See DOI: 00.0000/00000000.

similarity at the level of molecular fingerprints.

Despite previous demonstrations of the limitations in using SPP to predict similarity in biological activity[24,25], 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[2,26–28]. Martin *et al.*[24], have attributed the limitations to the lack of concurrence between the empirical and computational perception of similarity.

A variety of molecular fingerprints have been devised for use in ligand-based virtual screening (LBVS), to aid in identifying biologically active small molecules[28,29] (hereafter referred to as actives), from within a library of small molecules. LBVS begins with a small number of known 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). 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 actives in a library of mostly decoy molecules.

2 Methods and Materials

2.1 Fingerprint Representations

The palette of fingerprints evaluated in this study (see Table 1), can be broadly classified into those based on (i) path, (ii) circular features, (iii) pharmacophore, and (iv) pre-defined generic substructures/keys[30]. The circular[31] and path-based fingerprints are generated using an exhaustive enumeration of (linear/circular) fragments up to a given radius/size, 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[15] works directly with the SMILES strings (rather than the usual dependence on a molecular graph representation), by fragmenting the strings into overlapping substrings.

All fingerprints were generated using open-source software. Routines in the RDKit[45] 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[38]. The other fingerprints were calculated using custom software that makes use of the *jCompoundMapper*[34] and Chemistry Development Kit[36] libraries.

Although a number of similarity metrics have been in use[46], the most common approach to measuring fingerprint similarity is the Tanimoto coefficient[47]. The Tanimoto coefficient T_{ab} is a measure of the similarity of two fingerprints, such that $T_{ab} = |F_a \cap F_b|/|F_a \cup F_b|$, where F_a and F_b are the fingerprints of molecules a and b, respectively – this is equivalent to the Jaccard index. The value ranges between 1 (identical fingerprints, though not necessarily identical compounds) and 0 (disjoint fingerprints).

2.2 Benchmarking Data Sets

In order to evaluate VS methods, numerous benchmarking data sets have been developed over the years[48,49]. 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[50–52]. Data set composition can impact VS evaluation, such that both artificial under- and over-estimation of enrichment have been documented[51,53,54].

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[55]: DUD-E is a widely-used data set for VS benchmarking, containing data for 102 protein targets. On average, each target is represented by ∼224 active ligands, ∼90 experimental decoys, and ∼14,000 computational decoys per target. Compounds are considered active based on a 1 μM experimental cutoff, and experimental decoys are ligands with no measurable affinity up to 30 μM. Computational decoy ligands are selected from ZINC[50] 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[56]) Tanimoto coefficient $T_{ab} < 0.5$.

MUV Maximum Unbiased Validation[57] MUV data sets are based on bioactivity data available in PubChem[58]. This benchmark design strategy makes use of experimental design to select sets of 30 actives (taken from confirmation assays) and 15000 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 the 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)[59] benchmark is based on BindingDB[60] 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

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[31,36]	CIRCULAR	Extended-connectivity fingerprint of diameter 0	1024
ECFP_2[31,36]	CIRCULAR	Extended-connectivity fingerprint of diameter 2	1024
ECFP_4[31,36]	CIRCULAR	Extended-connectivity fingerprint of diameter 4	1024
ECFP_6[31,36]	CIRCULAR	Extended-connectivity fingerprint of diameter 6	1024
ESTATE[36,37]	SUBSTRUCTURE	Fingerprint based on E-State fragments	79
FCFP_0[31,36]	CIRCULAR	Feature-class fingerprint of diameter 0	1024
FCFP_2[31,36]	CIRCULAR	Feature-class fingerprint of diameter 2	1024
FCFP_4[31,36]	CIRCULAR	Feature-class fingerprint of diameter 4	1024
FCFP_6[31,36]	CIRCULAR	Feature-class fingerprint of diameter 6	1024
FP2[38]	PATH	Indexes linear fragments up to 7 atoms in length	–
FP3[38]	SUBSTRUCTURE	Based on 55 SMARTS patterns defining functional groups	–
FP4[38]	SUBSTRUCTURE	Based on SMARTS patterns defining functional groups	–
KR[36,39]	SUBSTRUCTURE	Klekota-Roth SMARTS based fingerprint	4860
LINGO[15,36]	TEXT	fragmentation of SMILES strings	–
LSTAR[34]	PATH	Local Path Environments	4096
MACCS[40]	SUBSTRUCTURE	Molecular ACCess System structural keys	166
MAP4[41]	CIRCULAR	combines substructure and atom-pair concepts	2048
MHFP[42]	CIRCULAR	encodes circular substructures	2048
P2PPHAR2D[43]	PHARMACOPHORE	Pharmacophore pair encoding	4096
P3PPHAR2D[43]	PHARMACOPHORE	Pharmacophore triplet encoding	4096
PUBCHEM[36,44]	SUBSTRUCTURE	substructure fingerprint	881
RAD2D[18]	CIRCULAR	Topological Molprint-like fingerprints	4096
RDK5[45]	SUBSTRUCTURE	Encodes substructures at most 5 bonds long	1024
RDK6[45]	SUBSTRUCTURE	Encodes substructures at most 6 bonds long	1024
RDK7[45]	SUBSTRUCTURE	Encodes substructures at most 7 bonds long	1024
SIGNATURE[32,36]	SUBSTRUCTURE	based on an array of atom signatures	–
TT[33]	PATH	based on bond paths of four non-hydrogen atoms	–

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

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 (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.

set of 30 structurally diverse decoys per active. The DEKOIS data set includes 81 targets found in the DUD-E data set.

LIT-PCBA The LIT-PCBA benchmark[61] 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. Targets in the LIT-PCBA have a variable active to decoy ratio that ranges from as low as 1:20 to 1:19000.

2.3 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[62]. 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, while the worst possible value is 0.

AUC provides a measure of the sensitivity/specificity trade-off across the full sensitivity range, but medicinal chemists are often more interested in early recognition of active molecules[63]. 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, even though the early enrichment (ranking 10% of actives above every single decoy) is excellent.

To address this shortcoming of ROC AUC, a number of other metrics have been devised to assess early enrichment[28,64,65]. Unfortunately, it can be difficult to extract an intuitive meaning from these measures[66], 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[63] and sum of log rank[28] are summarised visually in Figure F1 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 x^{th} element. 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:

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

If no decoys have score greater than the x^{th} active element, then $d_p = 0$, so that $\mathrm{DRF}_p = 0$. If $\mathrm{DRF}_p = 1$, the fraction of decoys with score above x is the same as the fraction of actives – the method is performing equivalently to a random score assignment. A $\mathrm{DRF}_p = 0.2$ indicates that only 20% of the expected number of decoys remain (there is a 5-fold reduction in decoys), while 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), 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.

3 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 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.

3.1 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. $\mathrm{DRF}_{0.1}$ 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 $\mathrm{DRF}_{0.1}$ values close to one. Performance is somewhat better for DEKOIS and DUD-E, but not particularly strong, and is offset by concerns previously expressed about these data sets. Others have highlighted issues such as artificial enrichment[49,67–71] (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[51,72]. In

response, Imrie *et al.* [70] developed a deep learning method (DeepCoy) to generate decoys that closely match the actives in terms of the physicochemical properties while simultaneously minimizing risk of introducing false negatives. By incorporating the DeepCoy decoys into DUD-E, Imrie *et al.* [70] reported a lowering of the average per-target AUC to 0.63, compared with the value of 0.70 for the original set (using AutoDock Vina [73]). The use of the DeepCoy decoys in place of the original decoys was not found to impact the enrichment values (see Table S1 in the Supplementary Information). 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.

Most circular and path-based fingerprints employ a standard length of 1024 bits. O'Boyle and Sayle [74] 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 S2 in Supplementary Information).

3.2 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 [59], 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 are quite similar – 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 sufficiently high score threshold. 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 not sufficient enrichment for useful downstream analysis, particularly considering existing concerns about bias in DEKOIS decoys (see previous section).

3.3 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. 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 [75]. Figure 2 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 suggests 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 [76].

Figure 3 presents scatter plots corresponding to three of the heatmap squares in Figure 2. The middle plot shows fingerprint similarity and AC_{50} values for the target/fingerprint pair (target=VDR, fingerprint=SIGNATURE) with median Kendall correlation, 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 from -0.30 to 0.05 and for ASP with targets other than PPARG ($\tau = -0.53$) from -0.23 to 0.18. Even in the occasional case of a specific fingerprint having a high correlation with compound potency (perhaps due to chance effects in the case of low number of actives), such properties are not generalized enough to be useful for VS studies.

3.4 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 [77] is an excellent resource to evaluate 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 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 in-

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

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.

actives (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, suggesting 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, and that <1.5% of actives show Tanimoto similarity >0.1 to another active, raising concerns about the usefulness of MAP4.

4 Discussion

There is substantial interest in the development of computational approaches to identifying good candidate small molecule drugs for specified protein binding pockets. This can be supported by high-quality, succinct representations of the molecules in a library, such that it is possible to rapidly identify ”neighbors” of known or suspected active molecules. The results of this study demonstrate that molecular fingerprints, and specifically measurement of molecular similarity based on those fingerprints, are not effective at discriminating molecules with similar binding activity from those with dissimilar activity. This suggests that the field must expand beyond fingerprint representation of molecules. 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, and will represent the compound not as a monolith, but as a collection of surface patches[78–80]. These, we believe, will enable a more context-dependent emphasis on features of importance to particular interactions, without interference from unimportant features.

Author Contributions

VV, AR and TW conceived the study. VV performed the fingerprint similarity calculations and performed primary analysis of the fingerprint performance, with contributions from TJW, AR, and JG. All authors wrote the manuscript collaboratively.

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/`.

Fig. 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.

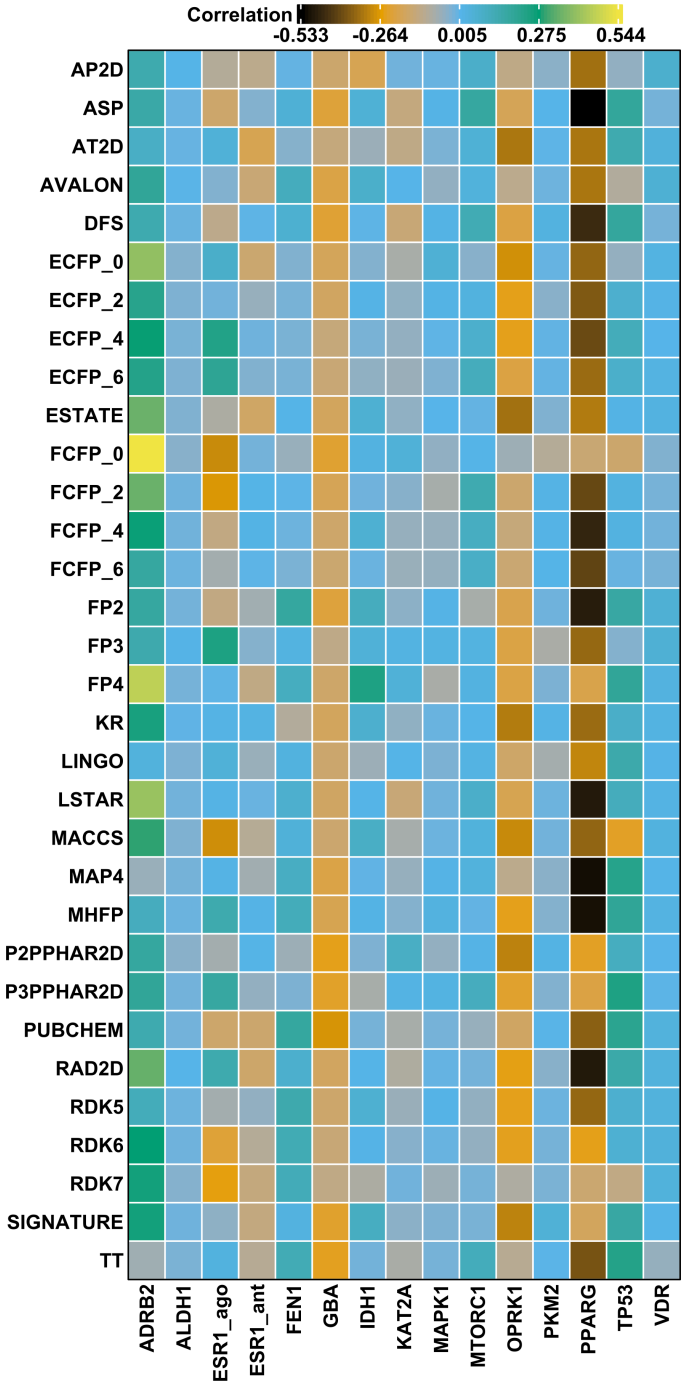

Fig. 2 Heatmap of the Kendall rank correlation (τ) between fingerprint Tanimoto (T_c) similarities calculated between the most active compound for a given target and the potency values (AC_{50}) of the actives for that target.

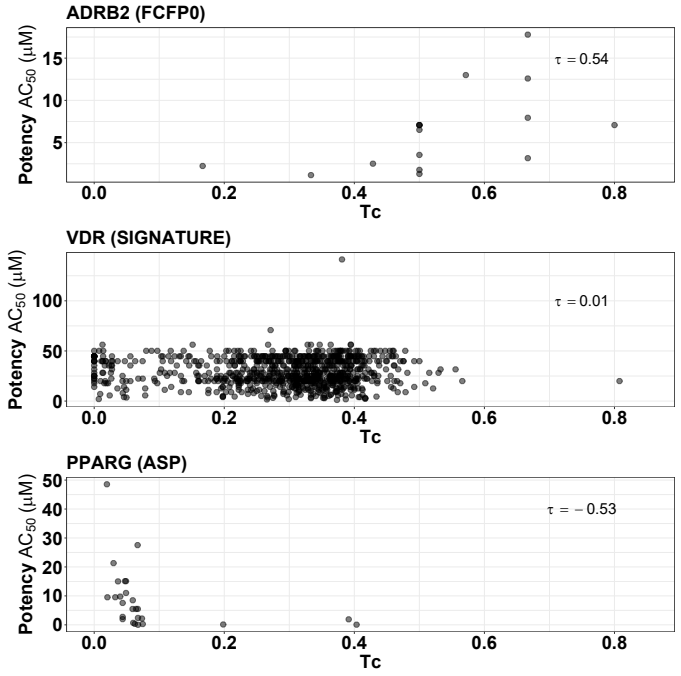

Fig. 3 Scatter plot of the fingerprint similarities (T_c) and the potencies (AC_{50}) of active compounds for ADRB2 (using FCFP0 fingerprint), VDR (SIGNATURE fingerprint) and PPARG (ASP fingerprint).

Conflicts of interest

The authors declare that they have no competing interests.

Acknowledgements

VV thanks the Research Council of Norway for financial support through grant no. 275752. AR and JLG acknowledge funding from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Department of Health and Human Services under BCBB Support Services Contract HHSN316201300006W/HHSN27200002 to MSC, Inc. TJW acknowledges funding from the National Institute of General Medical Sciences (NIGMS), National Institutes of Health (NIH), grant GM132600, and from the Biological and Environmental Research (BER) Program, Department of Energy (DOE), grant DE-SC0021216. The authors are grateful for the use of the GSCC cluster at the University of Montana, without which, these analyses could not have been performed.

Notes and references

1 A. A. Sadybekov, A. V. Sadybekov, Y. Liu, C. Iliopoulos-Tsoutsouvas, X.-P. Huang, J. Pickett, B. Houser, N. Patel, N. K. Tran, F. Tong *et al.*, *Nature*, 2022, **601**, 452–459.

2 V. Venkatraman, T. H. Colligan, G. T. Lesica, D. R. Olson, J. Gaiser, C. J. Copeland, T. J. Wheeler and A. Roy, *Front. Pharmacol.*, 2022, **13**,.

3 A. Luttens, H. Gullberg, E. Abdurakhmanov, D. D. Vo, D. Akaberi, V. O. Talibov, N. Nekhotiaeva, L. Vangeel, S. De Jonghe, D. Jochmans *et al.*, *J. Am. Chem. Soc.*, 2022, **144**, 2905–2920.

4 W. A. Warr, M. C. Nicklaus, C. A. Nicolaou and M. Rarey, *J. Chem. Inf. Model.*, 2022, **62**, 2021–2034.

5 W. P. Walters, *J. Med. Chem.*, 2018, **62**, 1116–1124.

6 A. Gimeno, M. Ojeda-Montes, S. Tomás-Hernández, A. Cereto-Massagué, R. Beltrán-Debón, M. Mulero, G. Pujadas and S. Garcia-Vallvé, *Int. J. Mol. Sci.*, 2019, **20**, 1375.

7 E. H. B. Maia, L. C. Assis, T. A. de Oliveira, A. M. da Silva and A. G. Taranto, *Front. Chem.*, 2020, **8**,.

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

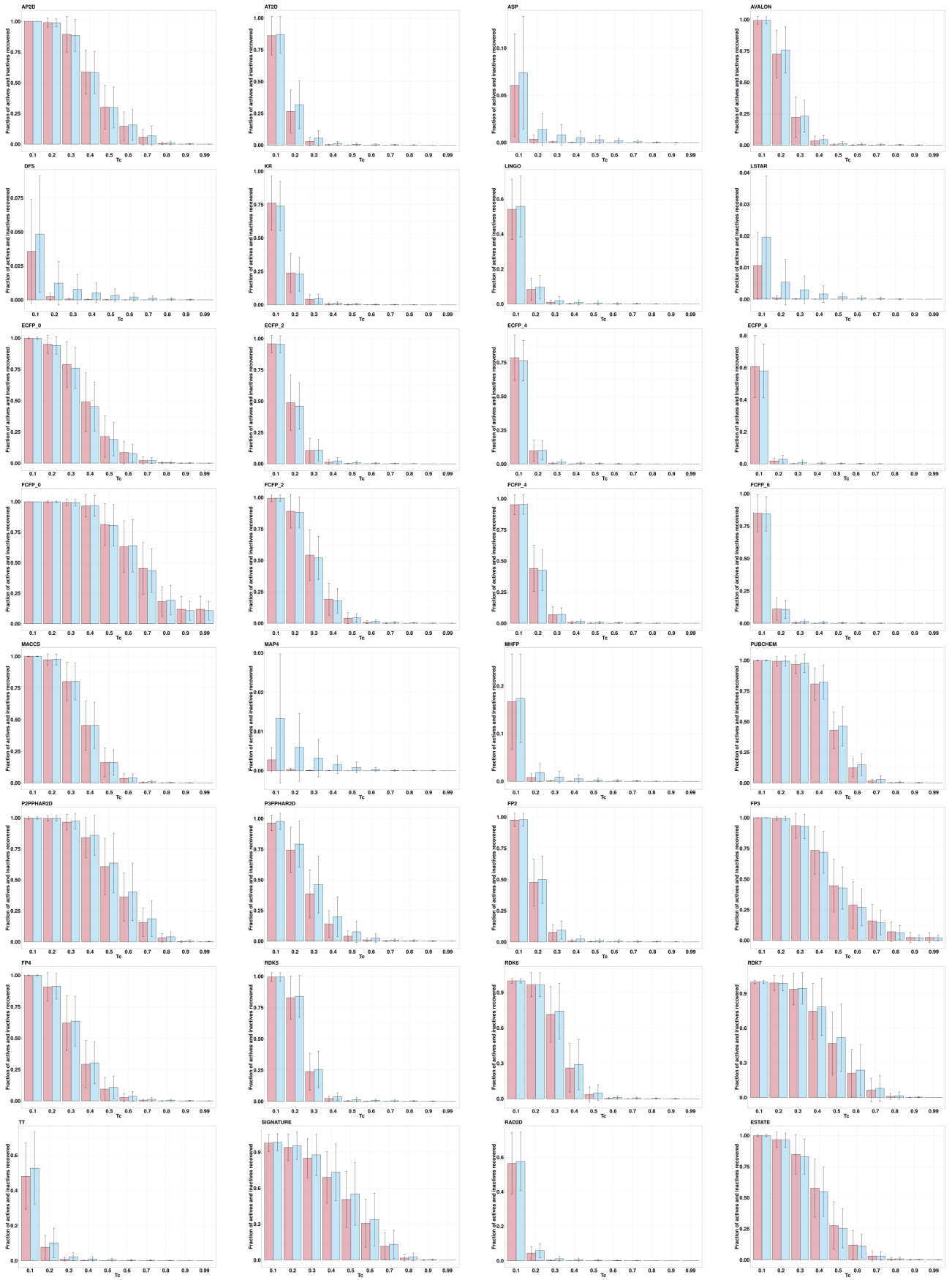

Fig. 4 For compounds in the St. Jude malaria data set, bar plots show the fraction of the inactives (in red) and actives (in blue) exceeding Tanimoto similarity cutoffs by the different fingerprints. Tanimoto similarities were calculated using each active as the query; mean and standard deviation (based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

9 G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, *J. Med. Chem.*, 2013, **57**, 3186–3204.

10 I. Cortés-Ciriano, C. Škuta, A. Bender and D. Svozil, *J. Cheminf.*, 2020, **12**,.

11 I. V. Tetko, P. Bruneau, H.-W. Mewes, D. C. Rohrer and G. I. Poda, *Drug Discov. Today*, 2006, **11**, 700–707.

12 C. Mellor, R. M. Robinson, R. Benigni, D. Ebbrell, S. Enoch, J. Firman, J. Madden, G. Pawar, C. Yang and M. Cronin, *Regul. Toxicol. Pharmacol.*, 2019, **101**, 121–134.

13 V. Venkatraman, *J. Cheminf.*, 2021, **13**,.

14 A. L. Teixeira and A. O. Falcao, *J. Chem. Inf. Model.*, 2014, **54**, 1833–1849.

15 D. Vidal, M. Thormann and M. Pons, *J. Chem. Inf. Model.*, 2005, **45**, 386–393.

16 H. Kubinyi, *Perspectives in Drug Discovery and Design*, 1998, **9**, 225–252.

17 F. Barbosa and D. Horvath, *Curr. Top. Med. Chem.*, 2004, **4**, 589–600.

18 A. Bender and R. C. Glen, *Org. Biomol. Chem.*, 2004, **2**, 3204.

19 E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chem. Soc. Rev.*, 2020, **49**, 3525–3564.

20 P. Willett, *Methods in Molecular Biology*, Humana Press, 2010, pp. 133–158.

21 D. Stumpfe and J. Bajorath, *WIRES Comput. Mol. Sci.*, 2011, **1**, 260–282.

22 A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé and G. Pujadas, *Methods*, 2015, **71**, 58–63.

23 K. Gao, D. D. Nguyen, V. Sresht, A. M. Mathiowetz, M. Tu and G.-W. Wei, *Phys. Chem. Chem. Phys.*, 2020, **22**, 8373–8390.

24 Y. C. Martin, J. L. Kofron and L. M. Traphagen, *J. Med. Chem.*, 2002, **45**, 4350–4358.

25 B. K. Shoichet, *Nature*, 2004, **432**, 862–865.

26 M. M. Hann and T. I. Oprea, *Curr. Opin. Chem. Biol.*, 2004, **8**, 255–263.

27 P. D. Leeson and B. Springthorpe, *Nat. Rev. Drug Discov.*, 2007, **6**, 881–890.

28 V. Venkatraman, V. I. Pérez-Nueno, L. Mavridis and D. W. Ritchie, *J. Chem. Inf. Model.*, 2010, **50**, 2079–2093.

29 S. Sciabola, R. Torella, A. Nagata and M. Boehm, *Mol. Inf.*, 2022, 2200103.

30 A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick and J. W. Davies, *J. Chem. Inf. Model.*, 2009, **49**, 108–119.

31 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

32 J.-L. Faulon, D. P. Visco and R. S. Pophale, *J. Chem. Inf. Model.*, 2003, **43**, 707–720.

33 R. E. Carhart, D. H. Smith and R. Venkataraghavan, *J. Chem. Inf. Model.*, 1985, **25**, 64–73.

34 G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner and A. Zell, *J Cheminf.*, 2011, **3**,.

35 L. Ralaivola, S. J. Swamidass, H. Saigo and P. Baldi, *Neural Networks*, 2005, **18**, 1093–1110.

36 E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, G. Torrance, C. T. Evelo, R. Guha and C. Steinbeck, *J Cheminf.*, 2017, **9**,.

37 L. H. Hall and L. B. Kier, *J. Chem. Inf. Model.*, 1995, **35**, 1039–1045.

38 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**,.

39 J. Klekota and F. P. Roth, *Bioinformatics*, 2008, **24**, 2518–2525.

40 J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Model.*, 2002, **42**, 1273–1280.

41 A. Capecchi, D. Probst and J.-L. Reymond, *J. Cheminf.*, 2020, **12**,.

42 D. Probst and J.-L. Reymond, *J. Cheminf.*, 2018, **10**,.

43 P. Mahé, L. Ralaivola, V. Stoven and J.-P. Vert, *J. Chem. Inf. Model.*, 2006, **46**, 2003–2014.

44 *PubChem Substructure Fingerprint*, `ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/`, 2022, Version: 1.3.

45 G. Landrum, *RDKit: Open-source cheminformatics*, `https://www.rdkit.org`, 2022, Release: 2022.03.5.

46 J. W. Raymond and P. Willett, *J. Comput. Aided Mol. Des.*, 2002, **16**, 59–71.

47 D. Bajusz, A. Rácz and K. Héberger, *J. Cheminf.*, 2015, **7**,.

48 N. Lagarde, J.-F. Zagury and M. Montes, *J. Chem. Inf. Model.*, 2015, **55**, 1297–1307.

49 M. Réau, F. Langenfeld, J.-F. Zagury, N. Lagarde and M. Montes, *Front. Pharmacol.*, 2018, **9**,.

50 J. J. Irwin, *J. Comput. Aided Mol. Des.*, 2008, **22**, 193–199.

51 M. Réau, F. Langenfeld, J.-F. Zagury, N. Lagarde and M. Montes, *Front. Pharmacol.*, 2018, **9**,.

52 V.-K. Tran-Nguyen and D. Rognan, *Int. J. Mol. Sci.*, 2020, **21**, 4380.

53 B. Nisius and J. Bajorath, *ChemMedChem*, 2010, **5**, 859–868.

54 L. Chaput, J. Martinez-Sanz, N. Saettel and L. Mouawad, *J. Cheminf.*, 2016, **8**,.

55 M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, *J. Med. Chem.*, 2012, **55**, 6582–6594.

56 I. Daylight Chemical Information Systems, *Daylight Theory Manual*, `https://www.daylight.com/dayhtml/doc/theory`, 2011, Version 4.9.

57 S. G. Rohrer and K. Baumann, *J. Chem. Inf. Model.*, 2009, **49**, 169–184.

58 D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner and E. Yaschenko, *Nucleic Acids Res.*, 2007, **36**, D13–D21.

59 M. R. Bauer, T. M. Ibrahim, S. M. Vogel and F. M. Boeckler, *J. Chem. Inf. Model.*, 2013, **53**, 1447–1462.

60 M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang and J. Chong, *Nucleic Acids Res.*, 2015, **44**, D1045–D1053.

61 V.-K. Tran-Nguyen, C. Jacquemard and D. Rognan, *J. Chem. Inf. Model.*, 2020, **60**, 4263–4273.

62 J. A. Hanley and B. J. McNeil, *Radiology*, 1982, **143**, 29–36.

63 J.-F. Truchon and C. I. Bayly, *J. Chem. Inf. Model.*, 2007, **47**, 488–508.

64 R. D. Clark and D. J. Webster-Clark, *J. Comput. Aided Mol. Des.*, 2008, **22**, 141–146.

65 J. C. D. Lopes, F. M. dos Santos, A. Martins-José, K. Augustyns and H. D. Winter, *J. Cheminf.*, 2017, **9**,.

66 W. Zhao, K. E. Hevener, S. W. White, R. E. Lee and J. M. Boyett, *BMC Bioinf.*, 2009, **10**,.

67 M. L. Verdonk, V. Berdini, M. J. Hartshorn, W. T. M. Mooij, C. W. Murray, R. D. Taylor and P. Watson, *J. Chem. Inf. Model.*, 2004, **44**, 793–806.

68 J. Sieg, F. Flachsenberg and M. Rarey, *J. Chem. Inf. Model.*, 2019, **59**, 947–961.

69 H. Li, K.-H. Sze, G. Lu and P. J. Ballester, *WIRES Comput. Mol. Sci.*, 2020, **10**,.

70 F. Imrie, A. R. Bradley and C. M. Deane, *Bioinformatics*, 2021, **37**, 2134–2141.

71 R. M. Stein, Y. Yang, T. E. Balius, M. J. O'Meara, J. Lyu, J. Young, K. Tang, B. K. Shoichet and J. J. Irwin, *J. Chem. Inf. Model.*, 2021, **61**, 699–714.

72 L. Chen, A. Cruz, S. Ramsey, C. J. Dickson, J. S. Duca, V. Hornak, D. R. Koes and T. Kurtzman, *PLoS ONE*, 2019, **14**, e0220113.

73 O. Trott and A. J. Olson, *J Comput. Chem.*, 2009, **31**, 455–461.

74 N. M. O'Boyle and R. A. Sayle, *J. Cheminf.*, 2016, **8**,.

75 M. G. Kendall, *Biometrika*, 1938, **30**, 81–93.

76 M. Vogt and J. Bajorath, *J. Chem. Inf. Model.*, 2013, **53**, 1613–1619.

77 A. Verras, C. L. Waller, P. Gedeck, D. V. S. Green, T. Kogej, A. Raichurkar, M. Panda, A. A. Shelat, J. Clark, R. K. Guy, G. Papadatos and J. Burrows, *J. Chem. Inf. Model.*, 2017, **57**, 445–453.

78 C. Hofbauer, H. Lohninger and A. Aszódi, *J. Chem. Inf. Model.*, 2004, **44**, 837–847.

79 P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. M. Bronstein and B. E. Correia, *Nature Methods*, 2019, **17**, 184–192.

80 D. Douguet and F. Payan, *Mol. Inf.*, 2020, **39**, 2000081.