A Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods

Vishwesh Venkatraman, Violeta I. Pérez-Nueno, Lazaros Mavridis, David Ritchie

A Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Dataset Reveals Limitations of Current 3D Methods

Vishwesh Venkatraman,[*,†,‡] Violeta I. Pérez-Nueno,[†,‡] Lazaros Mavridis,[†,‡] and David W. Ritchie[†]

INRIA Nancy Grand Est, LORIA, 54506, Vandoeuvre-lès-Nancy, France

E-mail: vishwesh.venkatraman@loria.fr

Abstract

In recent years, many virtual screening (VS) tools have been developed that employ different molecular representations and have different speed and accuracy characteristics. In this paper, we compare ten popular ligand-based VS tools using the publicly available Directory of Useful Decoys (DUD) dataset comprising over 100,000 compounds distributed across 40 protein targets. The DUD was developed initially to evaluate docking algorithms, but our results from an operational correlation analysis show that it is also well suited for comparing ligand-based VS tools. Although it is conventional wisdom that 3D molecular shape is an important determinant of biological activity, our results based on permutational significance tests of several commonly used VS metrics show that the 2D fingerprint-based methods generally give better VS performance than the 3D shape-based approaches for surprisingly many of the DUD targets. In order to help understand this finding, we have analysed the nature of the scoring functions used and the composition of the DUD dataset itself. We propose that in order to

[*]To whom correspondence should be addressed
[†]INRIA Nancy Grand Est, LORIA, 54506, Vandoeuvre-lès-Nancy, France
[‡]Contributed equally to this work

improve the VS performance of current 3D methods, it will be necessary to devise screening queries which can represent multiple possible conformations and which can exploit knowledge of known actives that span multiple scaffold families.

Introduction

The goal of ligand-based virtual screening (VS) is to search chemical databases to find compounds that best match a given query. Different VS tools are usually compared by assessing their ability to distinguish known active molecules from a large number of inactive compounds, or decoys, in a database. In recent years, many different VS tools have been developed[1–3] which often employ different representations of molecular properties and which often have different speed and accuracy characteristics. Hence there is a need to perform an objective comparison of currently available VS tools.

The results of VS studies depend on several factors, the two most important being the choice of representation and the matching algorithm. Although 3D molecular shape is clearly crucial for ligand binding, it has been reported previously that some 2D methods can still give better VS performance than 3D shape-based approaches.[4–6] Another important aspect is the nature of the actives and decoys that form the dataset.[7] Numerous studies to compare different VS approaches have been carried out, ranging from receptor-based docking[7–10] to ligand-based schemes[11,12] or a combination of both.[13,14] The general trend has been to show the superiority of the method being advocated with respect to the datasets analysed. It is therefore important that new algorithms be compared against standard benchmarks. Two datasets which can provide a good test for VS methods are the Directory of Useful Decoys (DUD)[15] and the Maximum Unbiased Validation (MUV)[16] have been introduced in the public domain. Here, we focus on the earlier DUD because it provides 3D coordinates for all actives and decoys and it includes a crystallographic ligand for almost every target.

The DUD is a publicly available dataset of about 100,000 compounds distributed over 40 pro-

tein targets.[15] The decoys are chosen to be physico-chemically similar to but topologically different from the actives. Though it was developed initially for evaluating docking methods, the DUD has been used in a number of ligand-based VS evaluation studies.[17,18] However, the dataset has been criticized for its intrinsic "analogue bias", which for most VS methods is expected to yield artificial enrichments.[19,20] Indeed, Irwin [19] has argued that because many of the compounds in the DUD share a common scaffold, performing ligand-based screening of this dataset should be trivially easy for both 2D and 3D approaches. Although, this premise has been quoted in several previous studies,[12,21–24] until now this expectation has not been thoroughly tested in detail. On the contrary, we believe that the DUD provides a good benchmark with which to assess screening ability. In order to test this supposition thoroughly, a comprehensive comparison of ten different 2D and 3D ligand-based tools was made using in all cases standard (*i.e.* default[25]) software parameters, and the results were evaluated using a range of commonly used metrics.

It has been reported previously that VS results depend strongly on the target family,[15] the query structure and conformation,[26] and also on the nature of the ligand and decoy sets.[25] Although analogue bias can influence the apparent utility of 2D fingerprint-based methods, we believe this is less of a concern for 3D shape-matching approaches which have to deal with the additional problems of selecting the best conformation to use as the query and finding the best 3D superposition between the query and each of the database compounds. By comparing the VS performance on the original DUD dataset with the filtered subset proposed by Good and Oprea,[20] we show here that these aspects influence shape-based VS performance to a much greater extent than analogue bias. Furthermore, although it is conventional wisdom that 3D molecular shape is one of the most important determinants of biological activity, our results show that current 2D fingerprint-based methods often give better VS performance than 3D shape-based approaches for surprisingly many of the DUD targets. In order to help understand this finding, we present analyses of the nature of the scoring functions tested and of the composition of the DUD dataset itself. We propose that in order to improve the VS performance of current 3D methods, it will be necessary to devise screening queries which can take into account multiple query or database conformations, and which can

better exploit knowledge of the structures of multiple known actives.

Materials and Methods

Virtual Screening Tools

Of the popular ligand-based VS tools studied here, five are based on 2D chemical fingerprint representations (OPENBABEL,[27] BCI,[28] MACCS,[29] DAYLIGHT,[30] and MOLPRINT2D[31]), and five use 3D molecular shape-based representations (ESHAPE3D,[29] ROCS,[32] PARAFIT,[33] SHAEP,[21] and USR[34]). Using these tools, a total of 15 scoring functions were tested among which four are variants of ROCS and EON, two are derived from ESHAPE3D and two are derived from SHAEP. Though the choice of the similarity metric and other software parameters can influence VS results, optimizing these for all the software used here is impractical. All calculations were therefore performed using the default software settings. Brief descriptions of these methods are given in the following sections.

2D Fingerprint Methods

2D fingerprint-based methods encode the structural features of molecules as bit strings, whereby each bit indicates the presence or absence of pre-defined structural and chemical patterns such as atom sequences, electronic configurations, atom pairs, and ring systems. Dictionary-based fingerprint approaches such as BCI (1052 bits), MACCS (166 bits) and MOLPRINT2D belong to this category. Alternatively, molecules can also be represented as hashed fingerprints as in DAYLIGHT (2048 bits) and OPENBABEL (1024 bits) which encode all patterns in a molecule: a pattern for each atom, a pattern representing each atom and its nearest neighbours, a pattern representing atoms and bonds connected by paths up to a pre-determined length N (typically $3 <= N <= 7$). A commonly used measure for comparing such 2D fingerprints is the Tanimoto coefficient T_{AB}, given by

$$T_{AB} = \frac{c}{a+b-c} \tag{1}$$

where a and b are the number of bits set in the fingerprints of molecules A and B, respectively, and where c is the number of bits set in both fingerprints. Thus, 2D fingerprint representations provide a very fast way to calculate molecular similarity.

3D Shape-Based Methods

3D methods use the atomic coordinates of the ligand structures to calculate shape-based similarity scores.[3] For example, ROCS[32] uses atom-centered Gaussian functions to represent molecular shape. Molecules are then superposed by maximizing the volume overlap of the structures being compared, and the 3D similarity is expressed numerically using a Tanimoto-like measure. While ROCS focuses on shape and chemical overlays, EON compares electrostatic fields calculated from the Poisson-Boltzmann equation. Here, four scoring schemes are compared: ROCS shape-only (ROCS_S), ROCS shape plus chemistry (ROCS_SC), EON shape plus electrostatics (EON_SE), and EON shape plus chemistry plus electrostatics (EON_SCE). For scoring, the default Tanimoto-combo and ET_combo scores are used here for ROCS and EON, respectively.

The PARAFIT program[35] compares and superposes spherical harmonic (SH) expansions of the molecular surface and local surface properties calculated from semi-empirical quantum mechanics theory using ParaSurf.[33] Here, molecules are ranked using the shape-Tanimoto score calculated from the PARAFIT surface overlap expression.

The SHAEP program[21] compares molecular field graph representations of the given structures and identifies maximal common subgraphs in order to perform a rigid body superposition. The similarity score (Hodgkin index) is a combination of both the shape and electrostatic potential evaluated at each field-graph vertex of the molecules. Here, two scoring functions are used, one based purely on shape (SHAPE_S) and one based on a combination of shape and electrostatics (SHAPE_SE).

The USR program uses three statistical moments (mean, standard deviation, and skewness) to

represent the molecular shape.[34] These moments are calculated from interatomic distance distributions to encode the size, compactness, and asymmetry of a molecule. The moments are calculated with respect to four reference points: the centroid, the atom closest to the centroid, the atom furthest to the centroid, and the atom furthest from it. This yields a descriptor of length 12 (three moments for each reference point). Molecular similarity is calculated as the inverse of the distance between the moment-based representations, F^A and F^B for molecules A and B, respectively:

$$S_{AB} = \frac{1}{1 + \frac{\sum_{i=1}^{12} |F_i^A - F_i^B|}{12}} \tag{2}$$

The ESHAPE3D program uses fixed length fingerprints generated from the eigenvalues of a heavy atom distance matrix.[29] A variant ESHAPE3D-HYD is calculated using the hydrophobic heavy atoms. Molecular similarities are calculated using an inverse distance metric similar to Eq. (2).

The DUD Dataset

The DUD dataset (Release 2 downloaded from http://dud.docking.org/r2) was screened for duplicates using MOE.[29] The numbers of actives and decoys are shown in Table 1. For each of the 40 targets in the dataset, the known actives and the target-specific decoys were used to compare the selected 2D and 3D ligand-based methods. For each target, the crystallographic ligand conformation was used as the query for both the fingerprint and the shape-based matching approaches. Although tautomer and ionization states can affect VS results,[36] all DUD structures were used without modification in order to perform a fair comparison.

Recently, filtered subsets of the DUD have been made[20,22] by applying lead-like Lipinski and Oprea rules, and by performing a reduced graph-based clustering of the actives (see Table 1).[12,37,38] VS results are also analysed with respect to this filtered subset.

Table 1: General statistics of the DUD dataset, showing the target name, Protein Data Bank (PDB) code, the numbers of decoys and actives for each target and the number of rotatable bonds (#rotB) in the crystallographic ligand. The targets are ordered in terms of increasing flexibility of the bound ligand. Also shown are the number of filtered actives proposed by Good and Oprea.[20]

Target	PDB code	#Decoys	#Actives	#Actives(Filtered)	#rotB
ar	2ao6	2792	74	63	3
cdk2	1ckp	2015	58	55	3
inha	1p44	3232	86	58	3
sahh	1a7a	1312	33	33	3
fgfr1	1agw	4490	120	73	4
cox2	1cx2	13158	412	250	5
gpb	1a8i	2115	52	52	5
parp	1efy	1331	35	33	5
pnp	1b8o	1017	30	30	5
tk	1kim	876	22	22	5
cox1	1q4g	908	24	23	6
er_agonist	112i	2517	67	63	6
fxa	1f0r	5549	146	64	6
trypsin	1bju	1644	46	10	6
mr	2aa2	630	15	13	7
pdgfrb	model	5904	169	136	7
ache	1eve	3867	106	101	8
ada	1ndw	904	37	37	8
gr	1m2z	2922	78	9	8
vegfr2	1vr2	2849	78	49	8
comt	1h1d	459	11	11	9
hivrt	1rt1	1495	42	35	9
pr	1sr7	1019	27	22	9
rxr-α	1mvc	744	20	18	9
src	2src	6217	159	102	9
alr2	1ah3	985	26	26	10
ampc	1xgj	781	21	21	10
pde5	1xp0	1972	76	34	12
er_antagonist	3ert	1434	39	18	13
hsp90	1uy6	965	25	24	13
egfr	1m17	15750	458	379	14
p38	1kv2	9041	353	219	14
ppar-γ	1fm9	3071	82	7	15
gart	1c2t	863	31	13	16
na	1a4g	1866	49	49	16
thrombin	1ba8	2425	68	26	16
dhfr	3dfr	8147	408	387	17
ace	1o86	1787	49	46	18
hmga	1hw8	1450	35	25	18
hivpr	1hpx	1999	62	6	23

Performance Metrics

Several metrics for assessing VS performance have been proposed,[39–42] with Receiver Operator Characteristic (ROC) curves being one of the most recommended.[43] Because only a small fraction of a database is tested experimentally in practical VS studies, it is often important to recognize actives or leads as "early" as possible. We therefore examined several metrics ranging from the commonly used enrichment factor (EF) to those that better highlight early performance. We also consider a simple new logarithmic metric which takes into account the rank of all the actives in the dataset. These metrics are described in more detail below.

The EF may be defined as the ratio of the number of actives retrieved relative to the number of database molecules tested.[38] An EF is often described with respect to a given percentage of the database screened. For example, $EF_{5\%}$ would represent the EF value obtained when 5% of the database has been screened. In general, if n represents the total number of actives and N represents the total number of molecules in the database, then if there are n_a actives among the $N_{x\%}$ molecules in the first x% of the database screened, the EF is given by

$$EF_{x\%} = \frac{n_a/N_{x\%}}{n/N} \tag{3}$$

Because EF values are easily influenced by the number of actives in the dataset, von Korff *et al.*[18] advocate the use of a relative EF defined as

$$EF_{rel} = \frac{n_a}{\min(N_{x\%}, n)} \times 100 \tag{4}$$

where the achieved enrichment for a given cut-off (say x%) is normalized by the maximum possible enrichment. Here, we report relative enrichment factors for the top 1%, 5%, and 10% of the database screened.

A ROC plot is conceptually similar to an enrichment plot in that it shows the relationship between the true positive rate (TPR, or sensitivity) and the false positive rate (FPR, or 1-specificity).[44]

The Area Under the Curve (AUC) of a ROC plot is common way to summarise the overall quality of a ROC plot. In the context of VS, the AUC is a measure of how highly a randomly selected active is ranked compared to a randomly chosen decoy.[45] The AUC is typically calculated for the whole of the ROC curve[38] using

$$AUC = \frac{1}{n}\sum_{i=1}^{n}(1 - f_i) \tag{5}$$

where f_i is the fraction of decoys ranked higher than the i^{th} active. The value of the AUC varies between 0 and 1, where 1 represents a perfect ranking (all actives ranked above the decoys) while 0.5 corresponds to a random ranking. However, EF and AUC values do not distinguish early and late performance.[39] Since early recognition is important in VS, we report values of the AUC for the first 5% and 10% of the ROC curve.

The BEDROC (Boltzmann-Enhanced Discrimination of ROC plots)[39] value is calculated as

$$BEDROC = \frac{\sum_{i=1}^{n}\exp(-\alpha\frac{r_i}{N})}{\frac{n}{N}\frac{1-\exp(-\alpha)}{\exp(\frac{\alpha}{N}-1)}} \times \frac{R_a \sinh(\frac{\alpha}{2})}{\cosh(\frac{\alpha}{2}) - \cosh(\frac{\alpha}{2} - \alpha R_a)} + \frac{1}{1-\exp(\alpha(1-R_a))} \tag{6}$$

where $R_a = \frac{n}{N}$ is the ratio of the number of actives to the total number of compounds, and α is an exponential weighting factor which controls the emphasis given to early recognition. While BEDROC scores emphasize the relative rank of an active, the simpler BAROC (Balanced ROC) metric proposed by Mackey and Melville[38] uses the fraction of decoys found before each active (f_i) to give a measure of early recognition:

$$BAROC = \frac{1}{n}\sum_{i=1}^{n}\exp(-\alpha f_i) \tag{7}$$

Both the BEDROC and BAROC metrics are bounded by 0 and 1. For the performance evaluations calculated in this study, the value of the weighting parameter α was set to 20.[39]

Finally, the sum of logarithms of ranks (SLR) metric is calculated as

$$SLR = -\sum_{i=1}^{n} log\left(\frac{r_i}{N}\right) \tag{8}$$

where r_i is the rank of the i^{th} active. The negative logarithm emphasizes early recognition. Noting that for an ideal case, a VS method would rank all actives within the first n positions, a theoretical maximum SLR may be calculated as

$$SLR_{max} = -\sum_{i=1}^{n} log\left(\frac{i}{N}\right) \tag{9}$$

This allows a normalised SLR (NSLR) to be calculated as

$$NSLR = \frac{SLR}{SLR_{max}} \tag{10}$$

This metric ranges from 0 to 1 (best achievable ranking).

Results

Enrichment Assessment and ROC Curves

We applied the above metrics to assess the performance of the selected VS methods. Table 2 shows the average relative EF values (Eq. (4)) calculated for the top 1%, 5%, and 10% of the database screened. This table shows that the 2D fingerprints BCI, BABEL, DAYLIGHT, and MOLPRINT2D give the best relative enrichments, followed by ROCS_SC (3D), MACCS (2D) and SHAEP (3D).

Table 3 shows the average relative EF values for the filtered subset of actives. Although the filtered subset gives better enrichment for all methods compared to the original unfiltered set (Table 2), the standard deviations are also significantly higher. This can be attributed to the reduced number of actives in the filtered set that inevitably enhances the calculated enrichment values, owing to the nature of the metric.[25,39] While 2D methods continue to give relatively better enrichments, ROCS_SC is the only 3D method with comparable values.

To our knowledge there is no standard way to combine EF or ROC results over multiple targets. Therefore, in order to provide a concise measure of how the methods performed across the

Table 2: Average relative enrichment rates (%) with corresponding standard deviations for the top 1%, 5%, and 10% of the database screened for all 40 DUD targets using the original actives and decoys.

METHOD	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$
BABEL	18.9 ± 10.8	8.4 ± 5.1	5.0 ± 2.6
DAYLIGHT	18.7 ± 11.0	8.4 ± 5.2	5.3 ± 2.7
MACCS	13.6 ± 11.2	6.0 ± 4.5	3.8 ± 2.3
BCI	19.6 ± 12.4	8.3 ± 5.5	4.9 ± 2.9
MOLPRINT2D	16.1 ± 11.4	7.0 ± 5.4	4.1 ± 2.8
PARAFIT_S	7.9 ± 8.5	4.3 ± 3.6	3.0 ± 2.0
ROCS_SC	15.5 ± 12.0	6.7 ± 5.2	4.1 ± 2.8
ROCS_S	10.6 ± 10.2	4.9 ± 4.2	3.2 ± 2.3
EON_SCE	10.9 ± 11.1	5.1 ± 4.8	3.4 ± 2.4
EON_SE	10.5 ± 11.3	4.8 ± 4.2	3.2 ± 2.3
SHAEP_SE	11.4 ± 10.2	4.8 ± 4.1	3.2 ± 2.2
SHAEP_S	11.2 ± 10.5	4.7 ± 4.0	3.1 ± 2.2
USR	5.3 ± 6.2	3.0 ± 2.5	2.2 ± 1.6
ESHAPE3D_HYD	6.7 ± 7.6	3.0 ± 2.7	2.3 ± 1.6
ESHAPE3D	7.0 ± 7.4	2.7 ± 1.7	1.9 ± 1.0

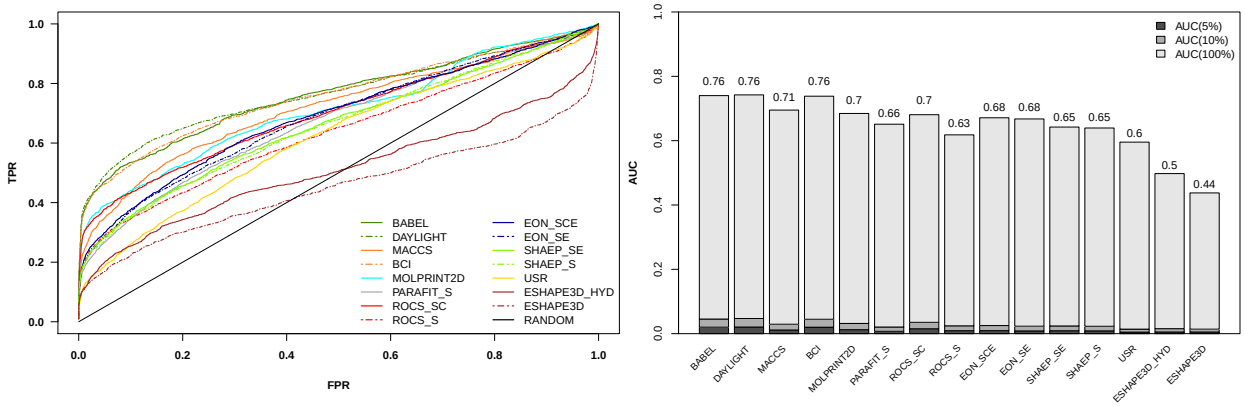
Table 3: Average relative enrichment rates (%) with corresponding standard deviations for the top 1%, 5%, and 10% of the database screened for all 40 DUD targets using the filtered set of actives prepared by Good and Oprea.[20]

METHOD	$EF_{1\%}$	$EF_{5\%}$	$EF_{10\%}$
BABEL	44.4 ± 28.4	41.1 ± 25.4	49.6 ± 26.6
DAYLIGHT	43.9 ± 28.7	41.8 ± 25.8	52.2 ± 26.7
MACCS	30.5 ± 25.7	29.7 ± 22.8	39.6 ± 23.3
BCI	46.7 ± 31.7	41.3 ± 28.5	49.1 ± 29.7
MOLPRINT2D	34.5 ± 28.3	33.8 ± 26.9	40.9 ± 30.2
PARAFIT_S	19.1 ± 20.3	24.4 ± 20.1	33.0 ± 22.4
ROCS_S	27.3 ± 25.7	27.8 ± 22.4	35.2 ± 24.1
ROCS_SC	36.8 ± 29.7	35.2 ± 27.1	44.0 ± 28.7
EON_SCE	24.2 ± 26.5	24.8 ± 24.1	33.3 ± 24.1
EON_SE	22.9 ± 25.4	24.7 ± 21.5	32.2 ± 22.8
SHAEP_SE	29.0 ± 25.5	27.2 ± 22.1	35.3 ± 23.7
SHAEP_S	28.1 ± 26.6	27.2 ± 22.1	35.5 ± 23.8
USR	12.7 ± 15.6	16.2 ± 13.9	24.3 ± 17.6
ESHAPE3D_HYD	24.0 ± 27.6	23.1 ± 20.8	27.8 ± 23.4
ESHAPE3D	14.1 ± 16.8	13.0 ± 9.8	18.6 ± 12.7

40 targets, we calculate an aggregate ROC plot for each method by vertically averaging the ROC curves.[46,47] Figure 1a shows the aggregate ROC curves for the 40 targets in the DUD dataset obtained for each of the VS methods studied. These plots show that with the exception of ESHAPE3D and ESHAPE3D_HYD, all methods perform significantly better than random (AUC > 0.5). Individual ROC curves for each of the 40 targets are available in the Supporting Information.

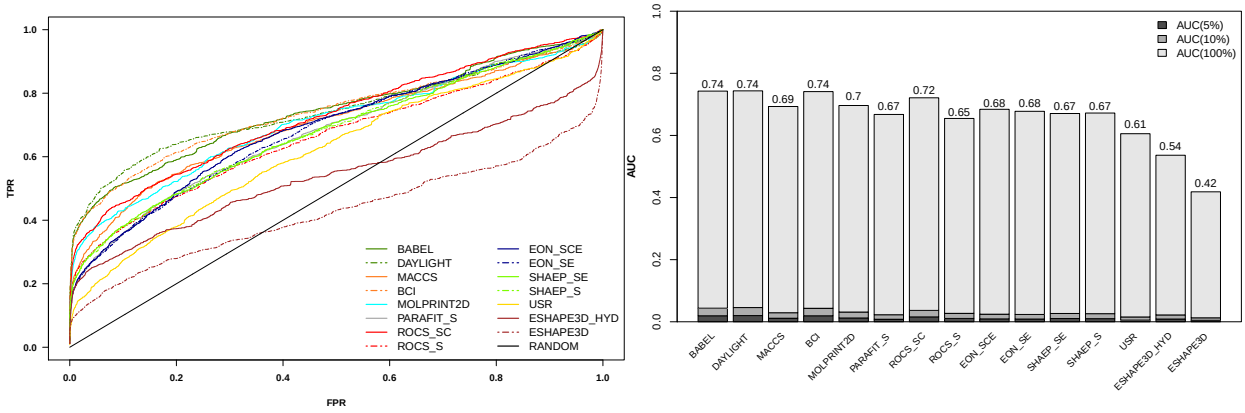
To summarise further the performance of the methods, Figure 1b shows a bar graph of the aggregate AUC values obtained for the 40 targets. This figure shows that the 2D approaches DAYLIGHT, BABEL and BCI give the highest mean AUC values of 0.76. These results are significantly better than those of the 3D ligand-based approaches, with only ROCS_SC (shape plus chemistry) achieving comparable performance (AUC=0.70). Although DAYLIGHT and ROCS_SC give the best 2D and 3D VS performance, respectively, no method achieves the maximum possible early recognition for the selected 5% ($AUC_{max} = 0.05$) and 10% ($AUC_{max} = 0.10$) thresholds. Corresponding ROC plot analyses for the filtered subset are shown in Figure 2. Overall, the general trends seen with the unfiltered (Figure 1a) and filtered datasets (Figure 2a) are highly similar. Comparing Figures 1b and 2b shows that filtering the dataset marginally improves the 3D methods (1-3%) and marginally reduces the performance for 2D methods (2%). On the other hand, ROC plots provide a more objective comparison of performance[25] which does not strongly depend on the number of actives and inactives. The overall similarity between the aggregate ROC results (Figures 1 and 2) shows that the VS performance of each method is broadly the same for both filtered and unfiltered datasets.

The relative performance of the VS methods can also be appreciated from the AUC "heat map" shown in Figure 3a. It is apparent from this figure that the 2D methods perform substantially better than the 3D methods. However, there are some targets for which both approaches are reasonably successful such as cox2, er_agonist, sahh, ar, and rxr-α. On the other hand, for the pdgfrb, p38, comt, alr2 and trypsin targets, the VS performance is relatively poor for both 2D and 3D methods.



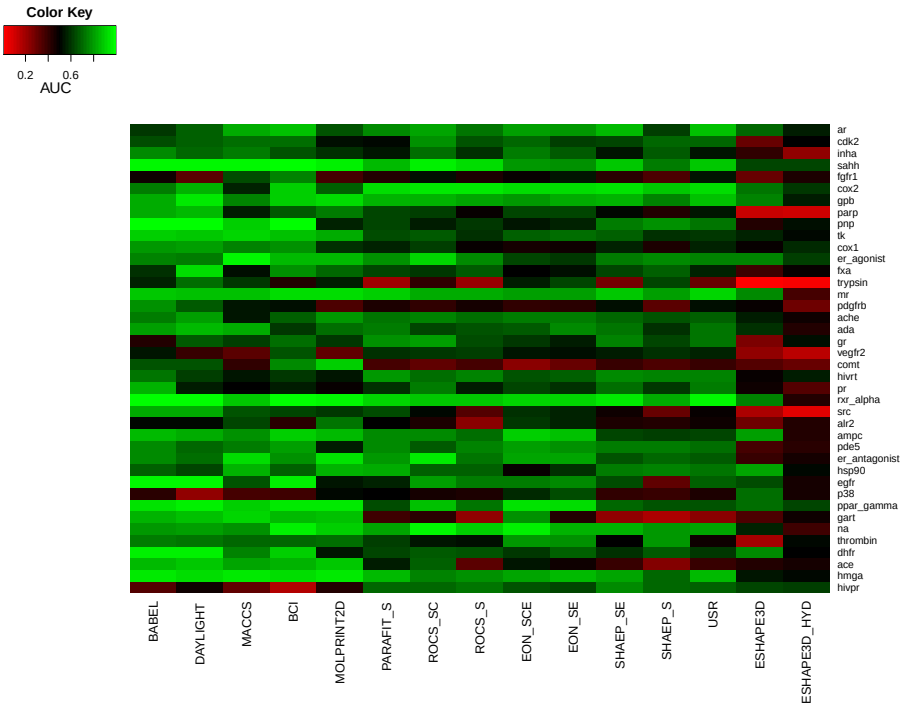
(a) Aggregate ROC plots. (b) Overall ROC performance.

Figure 1: (a) Aggregate ROC plots for the 40 DUD targets. (b) Bar chart of aggregate AUC values for the overall curve (grey), the top 10% (dark grey) and top 5% (black).

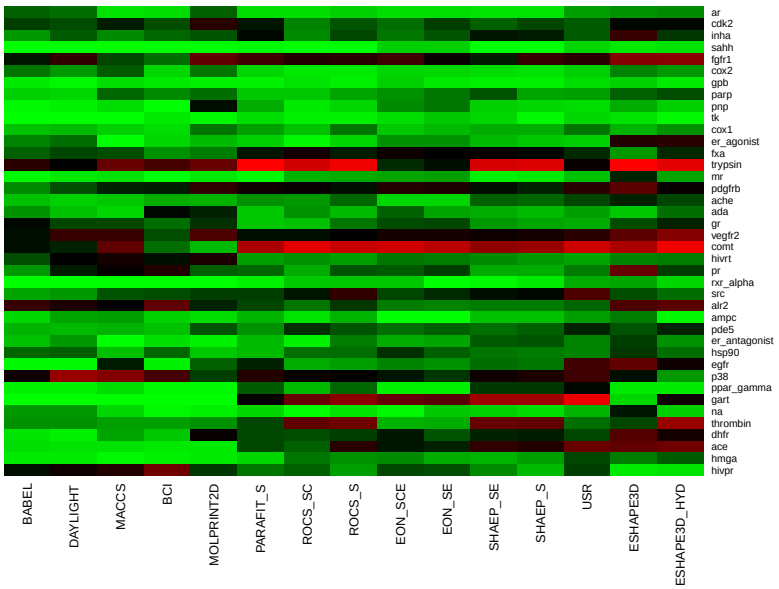


(a) Aggregate ROC plots. (b) Overall ROC performance.

Figure 2: (a) Aggregate ROC plots for the 40 DUD targets using the filtered subset proposed by Good and Oprea.[20] (b) Bar chart of aggregate AUC values for the overall curve (grey), the top 10% (dark grey) and top 5% (black).



(a)



(b)

Figure 3: Heat map plots showing the AUC for each scoring method for each of the 40 DUD targets in which (a) the AUC is calculated for each target-specific set of actives and decoys, and (b) the AUC is calculated using approximately 120,000 non target-specific decoys for each target.

Operational Correlation Analysis

The results presented so far have been calculated with respect to the specific actives and decoys for each target. In order to evaluate how similar the decoys are to the actives for each target, a large-scale VS experiment was carried out in which the decoy set for each target was assembled from the decoys of the remaining 39 targets (*i.e.* each decoy set now consists of approximately 120,000 decoys per target). This is similar to the "operational correlation analysis" advocated by Nicholls.[25] The results of this experiment are summarised in Figure 3b.

Although incorporating a large number of decoys sometimes reduces the absolute number of actives found in the first percentages, the significantly greater area of green regions (high AUC) in Figure 3b compared to Figure 3a shows that the overall VS performance of both the 2D and 3D methods improves for essentially all of the targets except comt (which has only 11 actives) and trypsin (whose performance with its own target-specific decoys is bad to begin with). This confirms that the target-specific DUD decoy sets are in fact very well constructed.

Using Multiple Database Compound Conformations

Thus far, the screening utility of the selected VS methods has been evaluated using only the crystallographic query and a single conformation for each ligand. In order to study whether using only one conformation per ligand adversely influences VS performance, an ensemble of up to a thousand conformers for each database molecule was calculated using OMEGA,[48] and the VS metrics were re-calculated for ROCS_SC. The results of this calculation are summarised in Table 4.

Compared to using a single database conformation, when using ten conformers the AUC values increase by more than 10% for twelve of the targets, and better quality superpositions with the query can be observed in those cases (details not shown). These cases correspond to small ligands or those with rigid groups for which the generated conformations span a good range of the possible structures. This is exemplified by the small ace ligands, which generally consist of two rigid ring moieties linked by from one to six rotatable bonds. In this case, it seems that using just ten database conformations can lead to better superpositions with the query, and the overall VS

retrieval AUC increases from 0.69 to 0.85. On the other hand, for the gart ligands which are larger and more flexible (typically having at least ten rotatable bonds), using multiple database conformations increases the AUC only marginally from 0.43 to 0.50.

Although using ten conformations works well for molecules with a modest number of internal degrees of freedom (e.g. up to six rotatable bonds), many more conformations need to be explored to improve the VS performance of highly flexible ligands. For example, using 1000 conformations the AUC value for gart increases signficantly from 0.43 to 0.84, but for ppar-γ only a modest improvement is seen (0.87 to 0.91). However, calculating additional conformations does not always lead to a better performance. For fifteen targets having from 3 to 23 rotatable bonds, the AUC actually decreases when 100 database conformations are used and when using 1000 conformations, the AUC decreases for fourteen targets. While for some of these cases the decrease is marginal, those for cdk2, pde5, vegfr2 and hivpr are more pronounced. Supplementary Figure XX shows a scatter plot of the data presented in Table 4. This shows that there seems to be no correlation between the flexibility of the ligand with the overall VS performance. In summary, it appears that using only ten OMEGA conformations for the database compounds gives relatively little overall improvement in VS performance and using more conformations is beneficial in some cases and worse in others.

Detailed Analysis of Selected Targets

The results presented above show that the 2D and 3D screening tools give different results for different targets. To explore this behaviour in more detail, fives cases were selected for closer examination, namely: cox2, for which both the 2D and 3D methods are successful, pdgfrb and trypsin, for which all methods perform poorly, hivpr, for which only the 3D methods give good results, and gart for which the 2D methods give better results than the 3D methods. This analysis was further extended to test the usefulness of several early recognition metrics, namely: BEDROC(Eq. (6)), BAROC (Eq. (7)), NSLR (Eq. (10)) and AUC 5%. Performance statistics for the cox2, pdgfrb, trypsin, gart, and hivpr targets are summarised in Figure 4 and Figure 5. Details of the analyses

Table 4: Summary of VS performance using ROCS_SC with and without multiple database conformations. Targets are ordered in terms of increasing flexibility of the bound ligand.

Target	AUC (1)	AUC (10)	AUC (100)	AUC (1000)
ar	0.81	0.80	0.79	0.79
cdk2	0.78	0.70	0.69	0.67
inha	0.72	0.81	0.78	0.79
sahh	0.96	0.96	0.98	0.98
fgfr1	0.53	0.61	0.51	0.45
cox2	0.95	0.94	0.95	0.95
gpb	0.84	0.93	0.94	0.94
parp	0.63	0.59	0.58	0.58
pnp	0.56	0.58	0.88	0.89
tk	0.68	0.84	0.88	0.89
cox1	0.62	0.62	0.58	0.57
er_agonist	0.92	0.93	0.94	0.94
fxa	0.61	0.49	0.66	0.64
trypsin	0.41	0.49	0.57	0.65
mr	0.83	0.86	0.86	0.85
pdgfrb	0.41	0.39	0.30	0.28
ache	0.76	0.75	0.77	0.78
ada	0.63	0.76	0.60	0.59
gr	0.81	0.81	0.77	0.76
vegfr2	0.61	0.54	0.44	0.39
comt	0.32	0.27	0.33	0.34
hivrt	0.71	0.71	0.72	0.71
pr	0.74	0.73	0.68	0.69
rxrα	0.88	0.98	0.97	0.95
src	0.51	0.50	0.39	0.34
alr2	0.45	0.47	0.49	0.50
ampc	0.77	0.86	0.88	0.88
pde5	0.68	0.58	0.59	0.56
er_antagonist	0.94	0.97	0.98	0.98
hsp90	0.69	0.71	0.66	0.64
egfr	0.81	0.82	0.95	0.95
p38	0.46	0.49	0.48	0.48
pparγ	0.87	0.68	0.74	0.91
gart	0.43	0.50	0.77	0.84
na	0.97	0.96	0.97	0.97
thrombin	0.54	0.66	0.66	0.55
dhfr	0.68	0.45	0.91	0.89
ace	0.69	0.85	0.82	0.77
hmga	0.76	0.90	0.94	0.93
hivpr	0.71	0.61	0.58	0.61

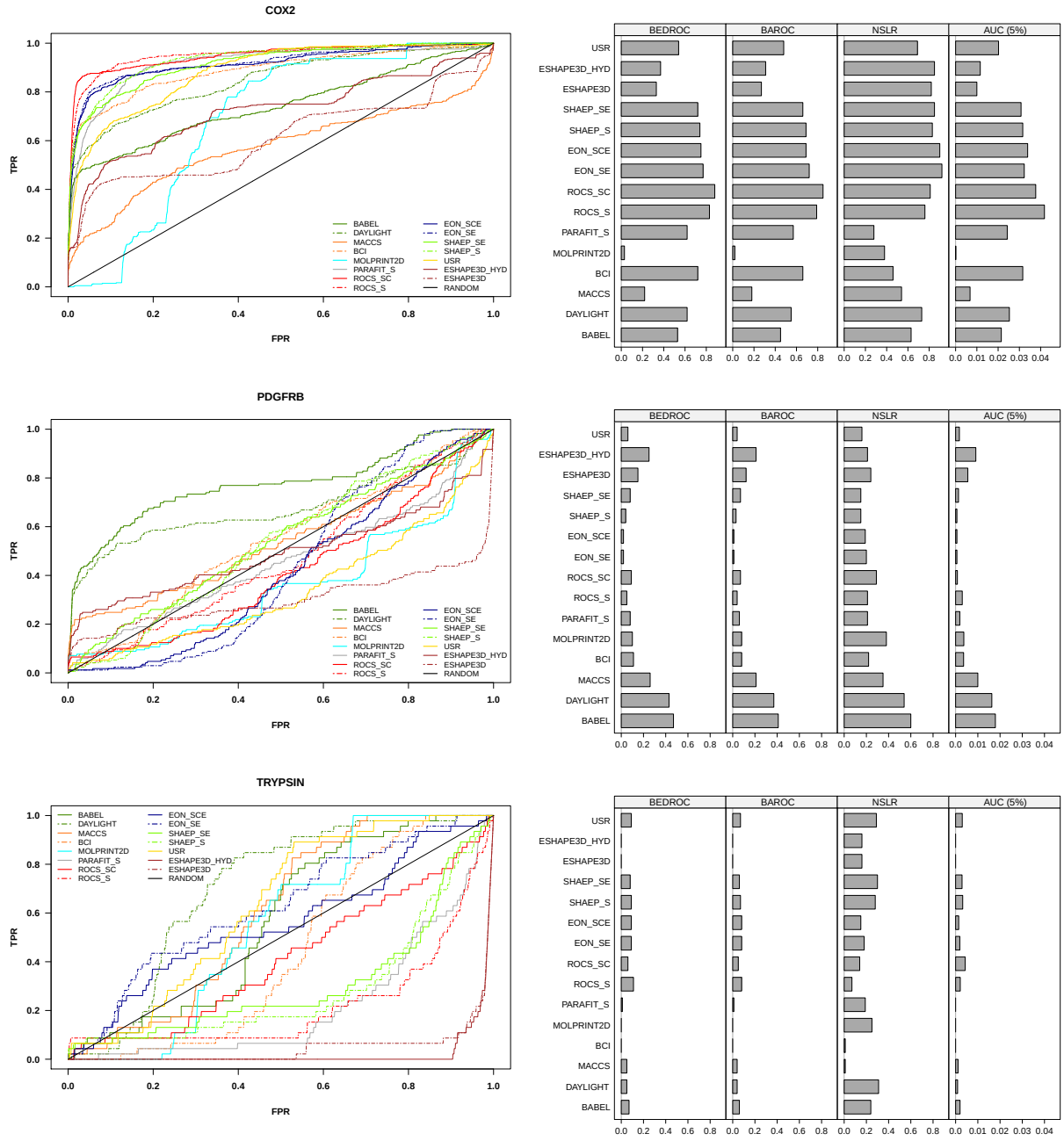


Figure 4: ROC plots and other figures of merit namely BEDROC, BAROC, NSLR and early AUC values at 5% of the database screened for the cox2, pdgfrb, and trypsin targets.

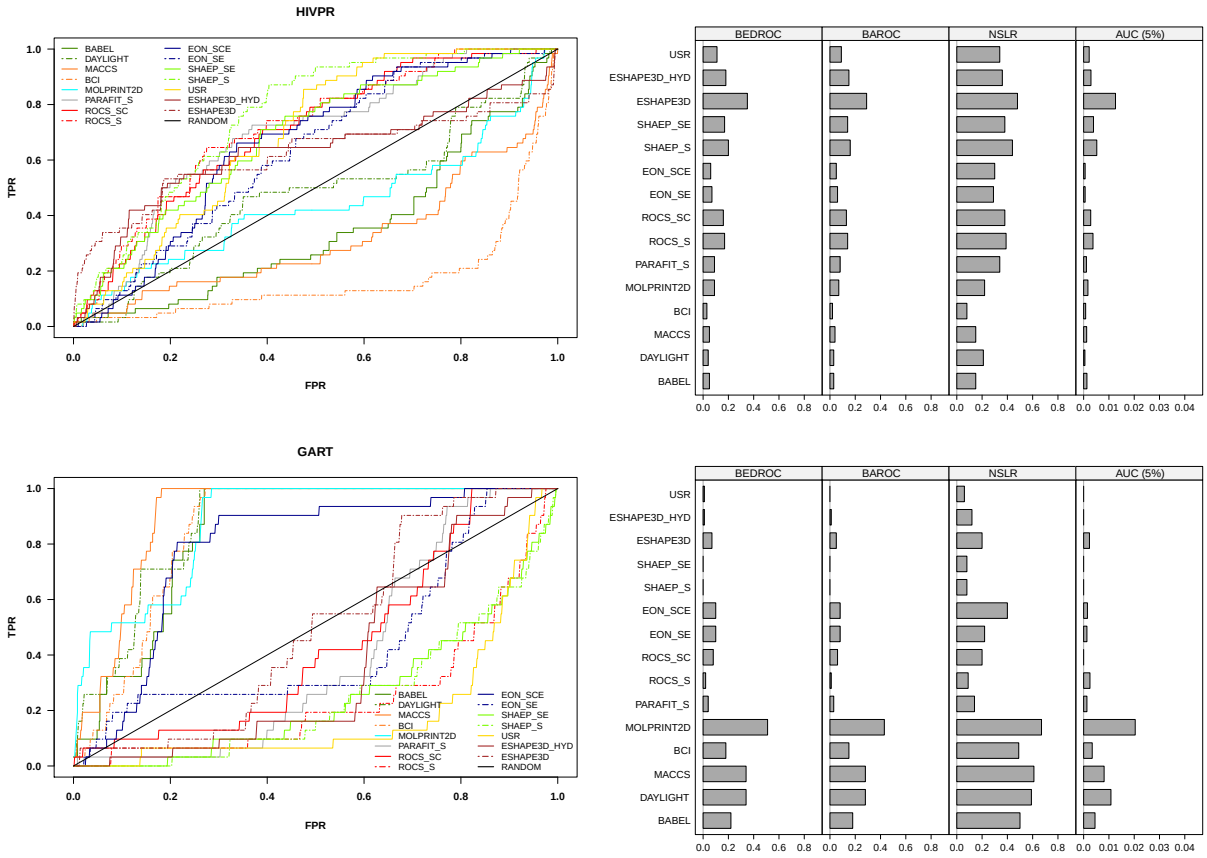


Figure 5: ROC plots and other figures of merit namely BEDROC, BAROC, NSLR and early AUC values at 5% of the database screened for the hivpr and gart targets.

for the remaining targets are available in the Supplementary Information.

Table 5 summarises the molecular weight, number of rotatable bonds, number of hydrogen-bond acceptor and donor atoms, number of hydrophobic atoms, and the octanol-water partition coefficient for the queries, actives, and decoys of the chosen targets. This table confirms that the physico-chemical properties of the actives are similar to those of the decoys but are sometimes different to those of the crystallographic query. Furthermore, some exceptions can also be seen when comparing the 3D overlays of the query with the actives and decoys, as shown below. The following sections examine these selected examples in further detail.

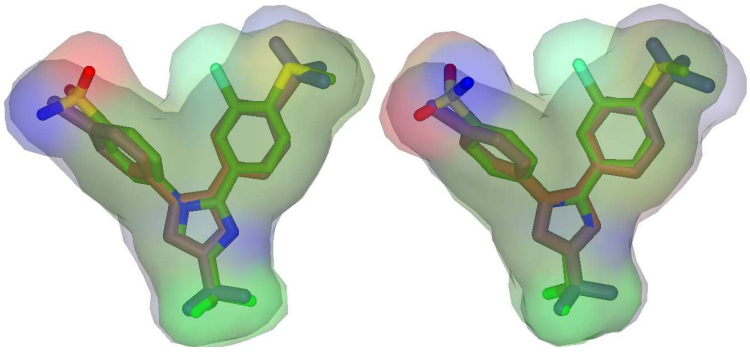
Table 5: Summary of the physico-chemical properties (number of rotable bonds (#rotB), molecular weight (MW), number of hydrogen bond acceptor (#HBA) and donor atoms (#HBD), number of hydrophobic atoms (#HYD) and the log of the octanol/water partition coefficient (SlogP) of the query (Q), decoy (D) and active (A) structures for cox2, pdgfrb, trypsin, hivpr, and gart.

Target	Q/D/A	#rotB	MW	#HBA	#HBD	#HYD	SlogP
cox2	Q	5	446.25	4	1	17	4.28
	D	5.83 ± 1.53	378.1 ± 23.96	3.69 ± 0.89	1.05 ± 0.59	15.18 ± 1.89	3.64 ± 0.69
	A	5.8 ± 1.49	383.07 ± 42.17	3.38 ± 1.17	0.51 ± 0.60	17.88 ± 2.8	4.08 ± 0.95
pdgfrb	Q	7	381.44	4	2	18	4.84
	D	7.52 ± 1.55	357.66 ± 20.14	3.57 ± 0.88	1.16 ± 0.46	15.25 ± 1.7	3.68 ± 0.66
	A	6.46 ± 2.96	356.06 ± 95.83	2.99 ± 1.21	0.91 ± 0.89	16.3 ± 4.18	3.53 ± 1.16
trypsin	Q	6	290.75	2	3	11	1.61
	D	12.2 ± 2.21	458.13 ± 36.28	5.58 ± 1.43	2.48 ± 1	18.11 ± 2.91	1.71 ± 1.32
	A	11.24 ± 3.18	481.17 ± 82.79	4.3 ± 1.41	1.13 ± 0.75	22.33 ± 4.52	0.79 ± 1.27
hivpr	Q	23	667.85	7	4	31	2.36
	D	9 ± 1.3	505.26 ± 24.7	5.42 ± 1.32	2.01 ± 0.7	24.24 ± 2.36	4.57 ± 0.88
	A	9.05 ± 1.83	519.2 ± 60.17	4.48 ± 1.17	2.18 ± 0.88	27.87 ± 4.71	5.25 ± 1.29
gart	Q	16	464.43	4	3	17	-1.72
	D	8.61 ± 1.6	496.59 ± 47.66	5.19 ± 1.89	3.4 ± 1.3	16.84 ± 3.45	-0.86 ± 1.59
	A	9.74 ± 0.89	456.32 ± 15.99	3.13 ± 0.34	4.06 ± 0.36	14.32 ± 0.98	-2.7 ± 0.64

Analysis of Cox2 VS Performance

Figure 4 shows that all methods except ESHAPE3D achieve high retrieval rates for cox2, with early performance metrics of greater than 0.8 for BEDROC, BAROC and NSLR. ROCS_S (3D) and BCI (2D) give the best results for cox2, obtaining near maximum values for the 5%AUC measure. On the other hand, the lower values for MOLPRINT2D and MACCS show that these methods have weaker screening utility for this target. As seen from Table 5, the physical properties of the query

are generally similar to those of the actives, presumably because all of the actives are derived from a central scaffold. Figure 6 shows that the query superposes well onto the scaffold (overlays using PARAFIT and EON_SCE), and therefore this example of analogue bias has a positive impact on VS performance, as might be expected.



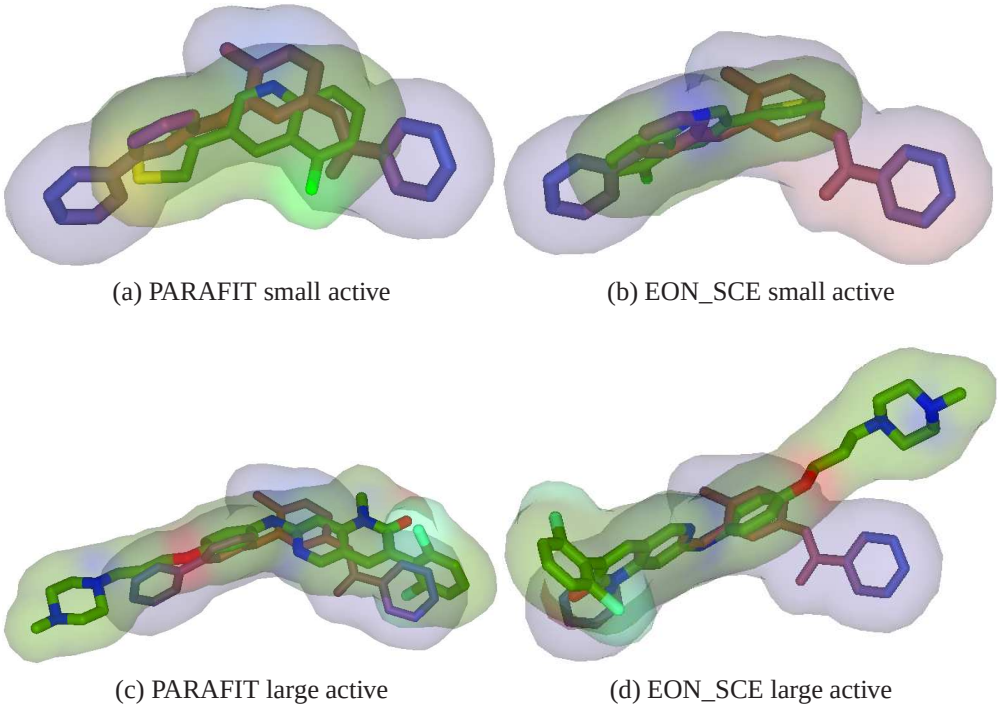
(a) PARAFIT (b) EON_SCE

Figure 6: Superpositions for the top-ranking cox2 active using (a) PARAFIT and (b) EON_SCE. The covalent structure and surface of the query is shown using a blue-red gradient and the superposed structures are coloured according to the atom type: nitrogen-blue, oxygen-red and carbon-green. All images are drawn using HEX[49] spherical harmonic surface overlays.

Analysis of Pdgfrb VS Performance

With the exception of molecular weight, the query, actives, and decoys for the pdgfrb target have similar properties (Table 5). While those of the decoys resemble the query more closely, the large standard deviation in the molecular weights of the actives indicates significant diversity in the molecular size. This probably explains why the 3D approaches give low enrichments for the pdgfr ligands (see the results for BEDROC, BAROC, NSLR, and *AUC*5% in Figure 4).

Inspecting the 3D superpositions from PARAFIT and EON_SCE clearly shows that both small (Figures 7a and 7b) and large ligands (Figures 7c and 7d) only superpose parts of the query, rather than matching the whole structure. This explains why the 3D shape-based algorithms are less successful for this target. Furthermore, pdgfrb is the only DUD target which does not have a crystal structure (it is homology modelled from the structure of c-Kit kinase[15]). Because the query structure was obtained from this template, it probably does not correspond exactly to the bound



(a) PARAFIT small active (b) EON_SCE small active

(c) PARAFIT large active (d) EON_SCE large active

Figure 7: Superpositions of pdgfrb onto the query using PARAFIT and EON_SCE for two different actives. The first set of overlays for a small active are shown in Figures (a) and (b). Figures (c) and (d) show the corresponding overlays for a large active. All molecules are coloured as in Figure 6.

ligand conformation for this target, and this may reduce the performance of both the 2D and 3D methods.

Analysis of Trypsin VS Performance

The VS results for trypsin are similar to those of pdgfrb, although the properties of the crystallographic query differ more considerably from those of the actives and decoys (Table 5). The query in this case has a much smaller molecular weight (MW=290.75) compared to the molecular weights of the other actives (average MW=481.17) and decoys (average MW=458.13). Differences are also seen for the other chemical properties such as hydrogen bond donors and acceptors, and hydrophobic atoms. These differences are mainly due to the absence of a sulphonamide group in the crystallographic query but which is present in 35 of the 46 actives and in several of the decoys.

Figure 8 shows the query superpositions calculated using PARAFIT and EON_SCE for ligands with and without the sulphonamide group. For the trypsin actives, poor quality overlays

are obtained when superposing the main scaffold onto the query. This mainly occurs for actives which have a sulphonamide group and other rigid parts such as fused rings or rings with terc-butyl substituents. Better shape overlays are seen for compounds without the sulphonamide group (Figure 8a and Figure 8b). Given that only 11 actives do not have the sulphonamide group, and by further taking into account the large differences in the chemical properties with respect to the query, it is therefore understandable that the screening performance of both 2D and 3D methods for the trypsin target is very low (see Figure 5).

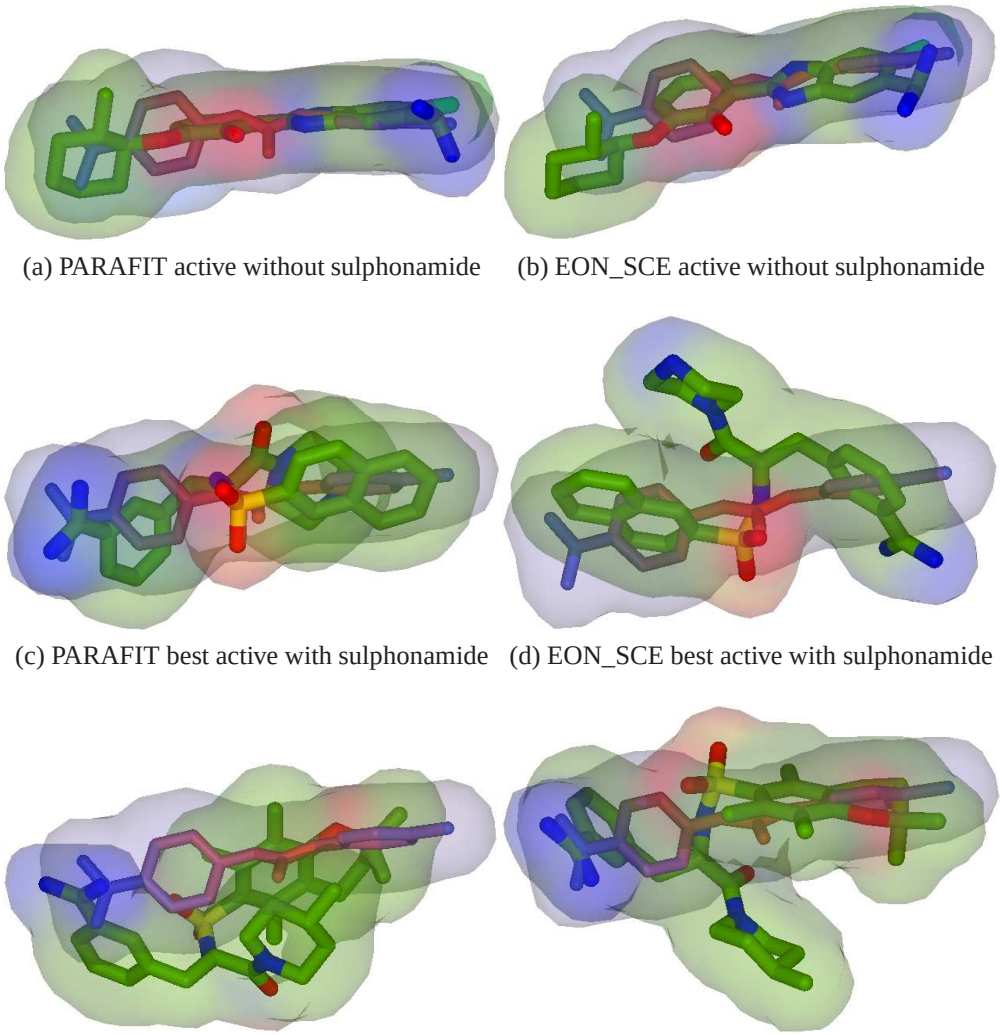
Analysis of Hivpr VS Performance

The hivpr target is an example for which the 3D methods have slightly better VS performance than the 2D methods. The data in Table 5 shows that the query and database compounds have noticeable differences in both the molecular weight (query MW=667.85, average decoy MW=505.26, average active MW=519.2) and other properties. Further analysis of the chemical structures shows that the differences in molecular weight are mainly due to the presence of different chemical groups containing heavy atoms in both ligands and decoys, but which are absent in the query. These differences cause considerable variations for the SlogP value (query SlogP=2.36, decoys SlogP=4.57 and, actives SlogP=5.25).

The alignments of the crystallographic query and the top ranking hivpr active and decoy using PARAFIT and EON_SCE are shown in Figure 9. The early performance metrics in Figure 5 show that the results for these two 3D methods are quite poor, and are at best comparable with the 2D approaches. However, slightly better results are seen for ROCS_SC, although including electrostatics worsens the original ROCS_S results in this case. For this target, ESHAPE_3D gives the best early recognition values, followed by SHAEP_S and ROCS_S.

Analysis of Gart VS Performance

For the gart target, the 2D methods give relatively high retrieval rates, with EON_SC being the only 3D method which gives comparable values. Table 5 shows that all compounds have similar

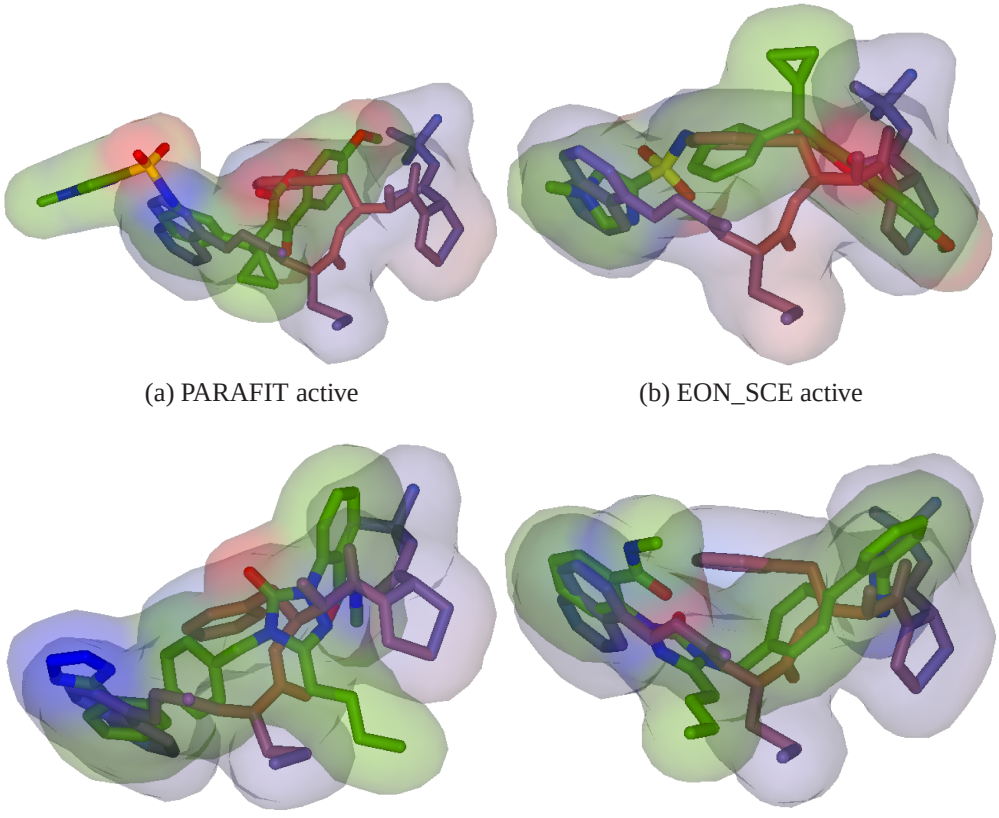


(a) PARAFIT active without sulphonamide (b) EON_SCE active without sulphonamide

(c) PARAFIT best active with sulphonamide (d) EON_SCE best active with sulphonamide

(e) PARAFIT worst active with sulphonamide (f) EON_SCE worst active with sulphonamide

Figure 8: This figure shows superpositions of the trypsin query for three different cases. Figures (a) and (b) show the overlays for an active which does not contain a sulphonamide moiety. Figures (c) and (d) show poorer overlays with the two highest scoring actives which do not contain a sulphonamide moiety, and Figures (e) and (f) show correspondingly worse overlays with the two lowest scoring actives. All molecules are coloured as in Figure 6.



(a) PARAFIT active

(b) EON_SCE active

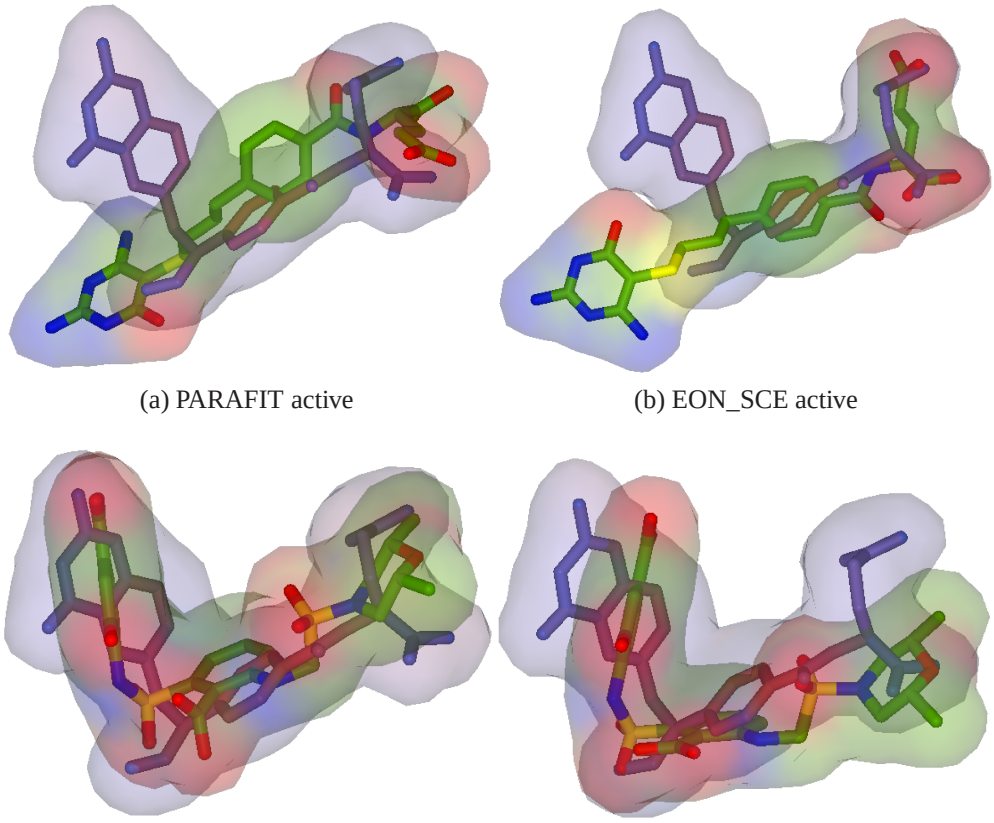
(c) PARAFIT decoy

(d) EON_SCE decoy

Figure 9: Superpositions of the hivpr query onto a top-ranking active and decoy using PARAFIT (left) and EON_SCE (right). All molecules are coloured as in Figure 6.

physico-chemical properties. Superpositions of the gart query and the top-ranking active and decoy are shown in Figure 10. This figure shows that the crystallographic query superposes the decoys relatively better than the actives, which explains the poor VS performance of the 3D approaches.

As described above, for this large and flexible (ten rotatable bonds) ligand at least 1000 database conformations need to be sampled to obtain significant performance gain (Table 4). On the other hand, it seems that the 2D approaches can find sufficient bits in common between the fingerpints of the query and the actives to be able to give good retrieval rates independently of the conformation.



(a) PARAFIT active (b) EON_SCE active

(c) PARAFIT decoy (d) EON_SCE decoy

Figure 10: Molecular superpositions of the gart query with the top-ranking active and decoy structures. All molecules are coloured as in Figure 6.

Discussion

Overall Analysis

It is generally accepted that the 3D shape of a molecule is an important factor for protein-ligand recognition. Therefore, 3D ligand-based VS methods are often expected to provide a better way to identify novel bioactive ligands than 2D methods. However, our results show that using current 3D VS tools with the bound crystallographic structure as the query often gives poorer VS results than the 2D fingerprint-based approaches. We believe this is due to factors more than simply conformational flexibility and analogue bias.

As illustrated by the gart ligands, conformational flexibility reduces the performance of a 3D shape-based virtual screen.[50] Large ligands tend to have many rotatable bonds, and hence require exhaustive sampling of their conformational space. The default settings in OMEGA were found to be suitable for the the small ace ligands, but only give a modest improvement for about 25% of the DUD targets. For the remaining targets, it seems that because the number of conformations sampled is relatively small, the chances of finding candidates that better resemble the query are low. Thus, despite being chemically similar, for many targets there is little "shape coherence"[50] between the structures of the actives and the query.

On the other hand analogue bias is often thought to enhance VS performance.[19,20] However, for the examples considered in this paper, analogue bias is found to be both beneficial (e.g. cox2) and detrimental (e.g. trypsin) for VS. Furthermore, comparison of the aggregate ROC plots between the original and filtered datasets shows almost no difference between the full and filtered sets of actives. Additionally, our operational correlation experiments show that the DUD decoys are in fact very well chosen. Hence analogue bias cannot explain the poor performance of the 3D methods observed here. Instead, the poor results for the 3D methods studied here can be explained in terms of the variety of the actives for some DUD targets. For instance, in targets where the query and the actives have dissimilar sizes, the retrieval rates are generally poor, as seen for pdgfrb. On the other hand, when the actives tend to have only minor structural differences from the query, the VS

retrieval rates are high. It is also important to note that the nature of the query molecule can influence the VS performance for both 2D and 3D approaches. Because 2D methods typically encode features of chemical groups and the topological distances between them (*i.e.* pharmacophoric and structural keys), the 2D methods can give good matches between active ligands without requiring that they superpose well.

Conversely, other targets (e.g. cox2, rxr-α and sahh) are better suited to 3D approaches. Furthermore, as illustrated by the trypsin target, the crystallographic ligand structures are not necessarily the best queries. On the other hand, it has been pointed out that different ligands may bind in different ways to their target sites (e.g. cdk2,[51] p38,[52] alr2[53]), and it is therefore unrealistic to expect that a single query can always select all of the actives. Similarly, one could argue that the 3D methods do not always perform as well as expected because they assume a global match to exist between active ligands.

Given that 3D ligand-based methods do not perform as well as expected even with multiple conformations for each database compound, it follows that one should try to improve the query. For example, Kirchmair et al. [26] used multiple active query conformations to improve VS performance. In a previous study, we calculated shape-based clusters of the conformations of multiple actives in order to identify a small number of representative queries to be used, and we are actively investigating this approach.[54]

Comparing Different VS Methods

In order to quantify VS performance, a number of metrics have been tested for their early recognition capability and as an indicator of a method's overall performance. Relative enrichment factors were used in the initial assessments but, as noted by previous authors, this measure does not emphasize early performance. On the other hand, AUCs of ROC plots are more robust and are easier to compare and interpret than EFs. However, because the AUC summarises the entire ROC plot for all thresholds, much of a ROC plot is of little practical interest. We therefore treated the first 5% and 10% of the ROC curves as distinct metrics. The BAROC and BEDROC metrics provide

closed formulae to score early retrieval while also measuring the overall performance. However, these metrics are again sensitive to the number of actives and decoys in the dataset and therefore only provide relative measures of utility.[25] On the other hand, the new NSLR metric does not require an exponential parameter or arbitrary cut-off to be defined yet it still intrinsically favours early recognition through its logarithm function.

There also exist differences in the behaviour of the evaluation metrics. In the pdgfrb target, for instance, the BEDROC, BAROC and AUC metrics suggest that DAYLIGHT and BCI give equivalent results (Figure 4). On the other hand, the NSLR values indicate that DAYLIGHT retrieves more actives earlier than BCI. Similarly for the trypsin target, the BEDROC and BAROC metrics indicate that SHAEP_S is marginally better than EON_SCE. On the other hand, inspection of the ROC plot shows that much of the SHAEP_S curve (AUC=0.30) lies below the random line which indicates that most of the actives are not recovered until very late. The corresponding ROC curve for EON_SCE gives an AUC of 0.56 ($NSLR = 0.30$) which is nearly twice that of SHAEP_S (AUC=0.16).

Nevertheless, none of the metrics can provide a direct indication of the best VS technique. In order to compare different ranking metrics in a statistically significant way, Zhao et al.[40] have proposed a technique based on random permutations. This involves repeatedly performing random reassignments of the ranks of the methods to be compared, re-calculating the performance metric for each permutation, and comparing the scores for the permuted ranks with that of the original unpermuted score. From this, a "p-value" is calculated as the proportion of scores for permuted ranks which exceed the original score. Typically, a p-value of less than 0.05 (95% confidence level) is considered to indicate a statistically significant difference.

In order to compare the 15 ranking methods used here, we computed the p-values for all possible pairs for VS ranking methods using 2,000 random permutations of the ranks. The results of the p-value comparisons for the AUC (overall, 5%, and 10%), NSLR, BAROC, and BEDROC metrics are shown in the "spider" diagrams in Figure 11. In these diagrams, each spoke and each curve corresponds to a method. The intersection of a curve and a spoke gives the number of targets

for which the spoke method performs significantly better than the curve method (as defined by the p-value test). Thus, the webs of a spider diagram will be larger between the spokes of good methods. For example, according to all of the metrics evaluated, Figure 11 shows that DAYLIGHT gives comparatively better VS performance than several other methods for a majority of the targets, whereas for many of the targets ESHAPE3D_HYD gives rather poor performance compared to the other VS tools.

As previously noted, the AUC provides an overall performance measure but it does not distinguish methods according to their ability to recognize actives at the beginning of a ranked list. The adjustable exponential parameter in the BAROC and BEDROC metrics broadly corresponds to setting an early performance AUC cut-off, but this restricts their ability to distinguishing the form of the ROC curves beyond the selected threshold.[40] On the other hand, Figure 11 shows that the simple NLSR metric described here gives a rather similar spider diagram to those of AUC 5%, BAROC and BEDROC. This suggests that NLSR provides a good *parameter-free* way to recognize both early recognition and overall performance.

In order to highlight the differences between the performance of the 2D and 3D methods, Figure 12 shows a NSLR-based spider diagram in which data for only 2D-3D and 3D-2D comparisons are plotted. This clearly shows that the 2D methods (top right quadrant) give considerable better VS performance than the 3D methods. Of the 3D methods, only the ROCS_SC, EON_SCE, and EON_SE tools give comparable performance to some of the 2D methods, *i.e.* MOLPRINT2D and MACCS.

Conclusion

Several metrics have been used to measure the performance of 15 commonly used 2D and 3D ligand-based VS tools on the 40 pharmaceutically relevant DUD targets. To our knowledge, this is the first comprehensive evaluation of ligand-based tools using this dataset. Although the validity of using the DUD as a VS benchmark has been questioned, our operational correlation analysis

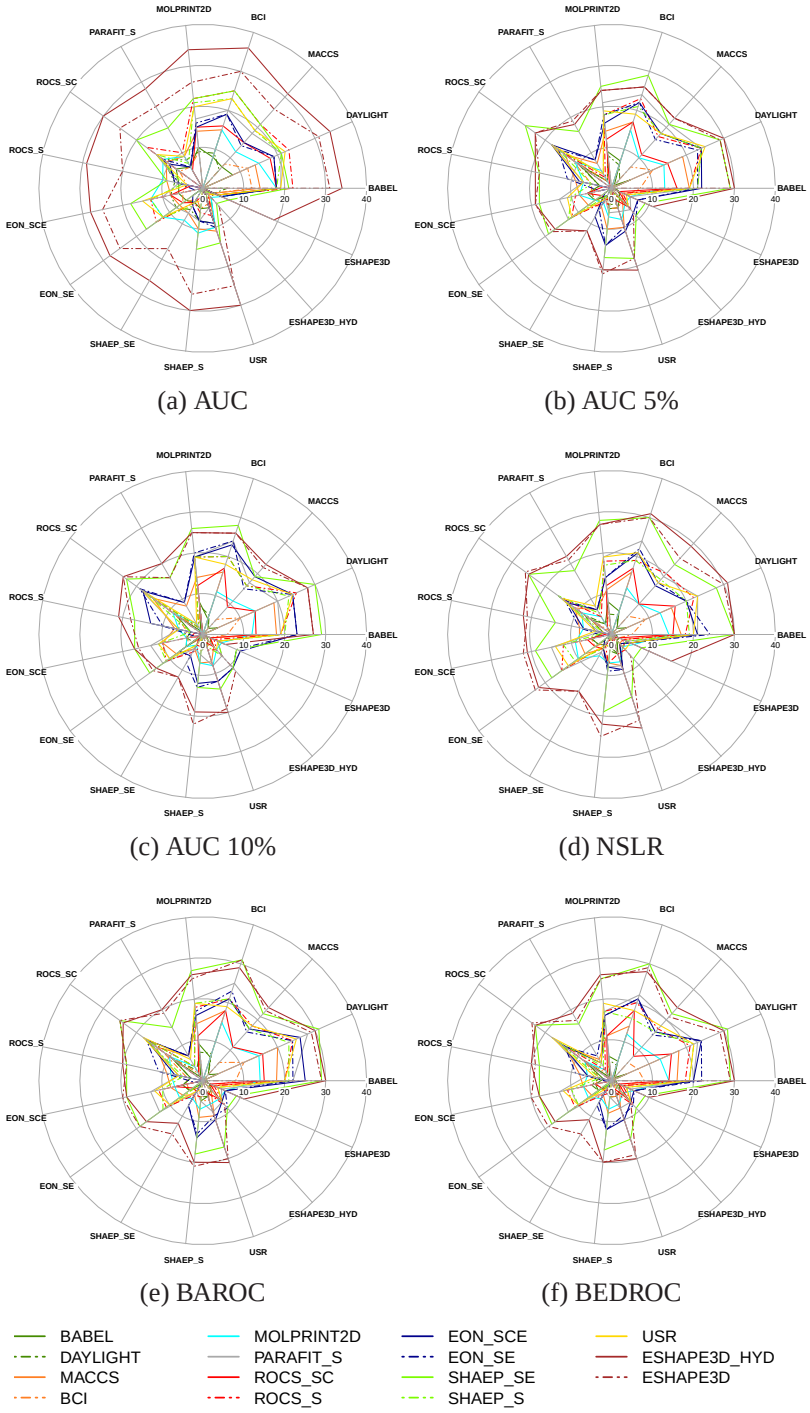


Figure 11: Spider diagrams showing the performance of the 15 VS ranking methods calculated using the overall AUC, AUC 5%, AUC 10%, NSLR, BAROC, and BEDROC metrics. Each spoke or radial line represents a method. Each colour-coded curve also corresponds to a ranking method. The intersection of a curve and a spoke shows the number of targets for which the spoke method gives significantly better VS performance than the curve method, as defined by the p-value test (see main text for details).

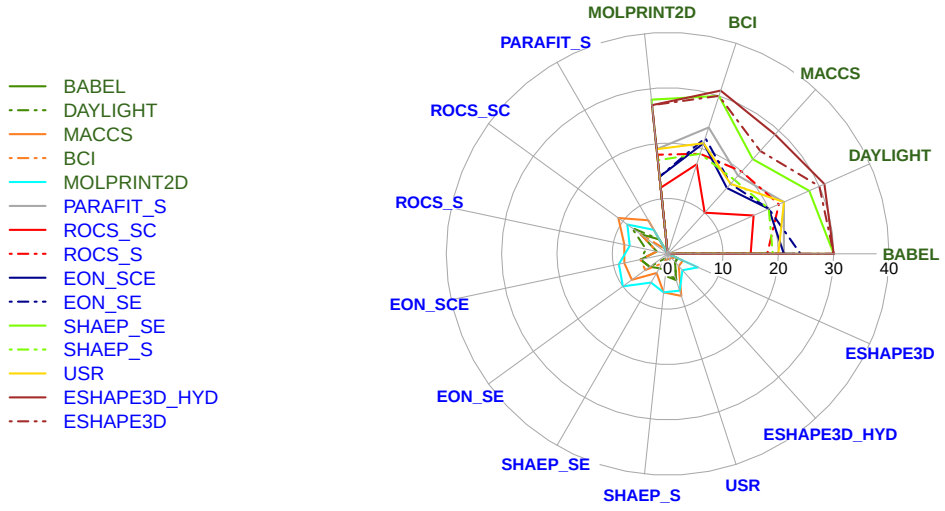


Figure 12: Spider diagram comparison of the 2D and 3D screening methods using the NSLR metric. This figures shows the same data as Figure 11d but with all 2D-2D and 3D-3D comparison points removed. This figure clearly shows that the 2D methods (green spokes) give significantly better VS performance than the 3D methods (blue spokes).

results show that the DUD is in fact well suited for ligand-based VS. Overall, we find that the 2D fingerprint-based methods give better VS performance than the 3D shape-based approaches for many of the DUD targets.

We believe that the poor results for 3D methods occur mainly because only a single conformation was used for the query and the database compounds. The ROCS-based 3D methods can use multiple different ligand conformations. We therefore used OMEGA and ROCS_SC to analyse thousand database conformations per target. However, this gave at best only a modest improvement for most of the targets. In some cases, much improved VS performances with only ten conformations were observed (ace, thrombin), while for others such as ppar-γ 1000 conformations were required. For many other targets, VS performance deteriorated when additional conformations were considered. Although, 3D methods should use a number of conformers in order to have a reasonable probability of finding good shape-based matches, we find that the VS performance has little, if any, correlation with the flexibility of the ligand.

One weakness of some 3D methods is their scoring functions. For example, the Carbo-like scoring functions in PARAFIT normalise the similarity score using the magnitudes of the surface shape descriptors, and this is effectively equivalent to scaling all molecules to a common size.

Hence, it would be useful to develop shape-based scoring functions which can better distinguish molecules of different sizes. It will also be useful to incorporate multiple properties and knowledge of multiple actives and their conformations into a small number of highly selective VS queries.[55] We are currently developing a consensus-shape based scoring scheme[54] which we believe should help improve the utility of 3D ligand-based approaches to virtual screening.

Although 3D methods are less prone to analogue bias and are better suited for scaffold hopping than the 2D approaches, they employ global representations that omit detailed atomic contributions. However, using a single 3D conformation for the query often fails to give the expected VS improvement. On the other hand, the use of "4D" multifconformer queries can be computationally expensive. As demonstrated by the ROCS_SC results, adding chemical information does improve VS performance. It would therefore be beneficial to develop methods which can encode the spatial and chemical constellations of molecular fragments, but which do not require them to be in a particular conformation. In other words, we should aim to develop more sophisticated 3D pharmacophore models which can combine the shape and chemical information from multiple active conformations, but which do not so strongly rely on global 3D shape matching techniques.

Acknowledgement

The authors thank OpenEye Scientific Software Inc., Cepos Insilico Ltd., Chemical Computing Group, DAYLIGHT Chemical Information Systems and Digital Chemistry for providing Academic Licences for ROCS, ParaSurf, MOE, DAYLIGHT, and BCI, respectively. We are also grateful to Mikko J Vainio for making the program SHAEP publicly available. This work is funded by the Agence Nationale de la Recherche, grant reference ANR-08-CEXC-017-01. V.I.P.-N. thanks the Generalitat de Catalunya-DURSI for grants from the programmes Formació de Personal Investigador (FI 2008) and Beques per a estades de recerca fora de Catalunya (BE-DGR 2009).

Supporting Information Available

Additional plots showing the performances of the methods on the different targets.

References

(1) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

(2) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **2007**, *12*, 225–233.

(3) Putta, S.; Beroza, P. Shapes of Things: Computer Modeling of Molecular Shape in Drug Discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1514–1524.

(4) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Model.* **1996**, *36*, 572–584.

(5) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Model.* **2000**, *40*, 295–307.

(6) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space: Target Fishing Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* **2006**, *49*, 6802–6810.

(7) 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.

(8) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct. Func. Bioinf.* **2004**, *57*, 225–242.

(9) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455–1474.

(10) Deng, W.; Verlinde, C. L. M. J. Evaluation of Different Virtual Screening Programs for Docking in a Charged Binding Pocket. *J. Chem. Inf. Model.* **2008**, *48*, 2010–2020.

(11) Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical Comparison of Virtual Screening Methods against the MUV Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 2168–2178.

(12) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. *J. Chem. Inf. Model.* **2008**, *48*, 2108–2117.

(13) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.

(14) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74–82.

(15) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(16) 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.

(17) Lee, H. S.; Lee, C. S.; Kim, J. S.; Kim, D. H.; Choe, H. Improving Virtual Screening Performance against Conformational Variations of Receptors by Shape Matching with Ligand Binding Pocket. *J. Chem. Inf. Model.* **2009**, *49*, 2419–2428.

(18) von Korff, M.; Freyss, J.; Sander, T. Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 209–231.

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

(20) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.

(21) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. *J. Chem. Inf. Model.* **2009**, *49*, 492–502.

(22) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. *J. Cheminf.* **2009**, *1*, 14.

(23) Kinnings, S. L.; Jackson, R. M. LigMatch: A Multiple Structure-Based Ligand Matching Method for 3D Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49*, 2056–2066.

(24) Giganti, D.; Guillemain, H.; Spadoni, J.; Nilges, M.; Zagury, J.; Montes, M. Comparative Evaluation of 3D Virtual Ligand Screening Methods: Impact of the Molecular Alignment on Enrichment. *J. Chem. Inf. Model.* **2010**, *50*, 992–1004.

(25) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.

(26) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *49*, 678–692.

(27) *The Open Babel Package version 2.2.3*, 2009. `http://openbabel.org/wiki/Main_Page/`, Accessed Date December 6, 2009.

(28) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Softwares. *J. Chem. Inf. Model.* **1997**, *37*, 141–142.

(29) *Molecular Operating Environment (MOE), version 2008.10 Release*, 2009.

(30) *Daylight Version 4.62*, 1999.

(31) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Model.* **2004**, *44*, 1708–1718.

(32) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comp. Chem.* **1996**, *17*, 1653–1666.

(33) Lin, J.; Clark, T. An Analytical, Variable Resolution, Complete Description of Static Molecules and Their Intermolecular Binding Properties. *J. Chem. Inf. Model.* **2005**, *45*, 1010–1016.

(34) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comp. Chem.* **2007**, *28*, 1711–1723.

(35) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Toward High Throughput 3D Virtual Screening Using Spherical Harmonic Surface Representations. *J. Chem. Inf. Model.* **2007**, *47*, 1787–1796.

(36) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection? What can we learn from earlier mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.

(37) Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J. Further Development of Reduced Graphs for Identifying Bioactive Compounds. *J. Chem. Inf. Model.* **2003**, *43*, 346–356.

(38) Mackey, M. D.; Melville, J. L. Better than Random? The Chemotype Enrichment Problem. *J. Chem. Inf. Model.* **2009**, *49*, 1154–1162.

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

(40) Zhao, W.; Hevener, K.; White, S.; Lee, R.; Boyett, J. A statistical framework to evaluate virtual screening. *BMC Bioinf.* **2009**, *10*, 225.

(41) Sheridan, R. P. Alternative Global Goodness Metrics and Sensitivity Analysis: Heuristics to Check the Robustness of Conclusions from Studies Comparing Virtual Screening Methods. *J. Chem. Inf. Model.* **2008**, *48*, 426–433.

(42) Swamidass, S. J.; Azencott, C.-A.; Daily, K.; Baldi, P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics* **2010**, *26*, 1348–1356.

(43) 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.

(44) Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874.

(45) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–190.

(46) Macskassy, S. A.; Provost, F.; Rosset, S. ROC confidence bands: an empirical evaluation. *ICML '05: In Proceedings of the 22nd international conference on Machine learning*, New York, NY, USA, 2005; pp 537–544.

(47) Provost, F. J.; Fawcett, T.; Kohavi, R. The Case against Accuracy Estimation for Comparing Induction Algorithms. *ICML '98: In Proceedings of the Fifteenth International Conference on Machine Learning*, San Francisco, CA, USA, 1998; pp 445–453.

(48) Böstrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics. Modell.* **2003**, *21*, 449–462.

(49) Ritchie, D. W.; Kemp, G. J. Protein docking using spherical polar Fourier correlations. *Proteins: Struct. Func. Gen.* **2000**, *39*, 178–194.

(50) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53*, 3862–3886.

(51) Sato, H.; Shewchuk, L. M.; Tang, J. Prediction of multiple binding modes of the CDK2 inhibitors, anilinopyrazoles, using the automated docking programs GOLD, FlexX, and LigandFit: an evaluation of performance. *J. Chem. Inf. Model.* **2006**, *46*, 2552–2562.

(52) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9*, 268–272.

(53) Steuber, H.; Zentgraf, M.; Motta, C. L.; Sartini, S.; Heine, A.; Klebe, G. Evidence for a Novel Binding Site Conformer of Aldose Reductase in Ligand-Bound State. *J. Mol. Biol.* **2007**, *369*, 186–197.

(54) Pérez-Nueno, V. I.; Ritchie, D. W.; Borrell, J. I.; Teixidó, J. Clustering and Classifying Diverse HIV Entry Inhibitors Using a Novel Consensus Shape-Based Virtual Screening Approach: Further Evidence for Multiple Binding Sites within the CCR5 Extracellular Pocket. *J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.

(55) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The Use of Consensus Scoring in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 277–288.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.