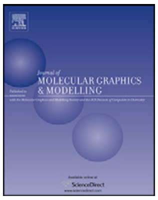
Journal of Molecular Graphics and Modelling 41 (2013) 20–30

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM



Benchmarking of HPCC: A novel 3D molecular representation combining shape and pharmacophoric descriptors for efficient molecular similarity assessments

Arnaud S. Karaboga[a,*,1], Florent Petronin[a,1], Gino Marchetti[a,1], Michel Souchet[b,*,2], Bernard Maigret[a,1]

[a] LORIA UMR 7503, CNRS-Nancy University and INRIA Nancy Grand-Est, Equipe Orpailleur, BP239, 54503 Vandoeuvre les Nancy cedex, France
[b] Harmonic Pharma, Espace Transfert, 615 rue du Jardin Botanique, 54600 Villers les Nancy, France

ARTICLE INFO

Article history:
Received 10 October 2012
Received in revised form 11 January 2013
Accepted 16 January 2013
Available online 26 January 2013

Keywords:
Molecular similarity
3D description
Ligand-centric pharmacophoric description
Spherical harmonic molecular representation
Retrospective shape-based screening

ABSTRACT

Since 3D molecular shape is an important determinant of biological activity, designing accurate 3D molecular representations is still of high interest. Several chemoinformatic approaches have been developed to try to describe accurate molecular shapes.

Here, we present a novel 3D molecular description, namely harmonic pharma chemistry coefficient (HPCC), combining a ligand-centric pharmacophoric description projected onto a spherical harmonic based shape of a ligand. The performance of HPCC was evaluated by comparison to the standard ROCS software in a ligand-based virtual screening (VS) approach using the publicly available directory of useful decoys (DUD) data set comprising over 100,000 compounds distributed across 40 protein targets.

Our results were analyzed using commonly reported statistics such as the area under the curve (AUC) and normalized sum of logarithms of ranks (NSLR) metrics. Overall, our HPCC 3D method is globally as efficient as the state-of-the-art ROCS software in terms of enrichment and slightly better for more than half of the DUD targets. Since it is largely admitted that VS results depend strongly on the nature of the protein families, we believe that the present HPCC solution is of interest over the current ligand-based VS methods.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Drug discovery remains a lengthy and costly process in which in silico approaches have become an important component to help improving the overall process leading to the identification of new lead compounds and thus, leveraging drug development [1,2]. A wide range of computational screening methods are available for mining chemical and biological data in order to identify novel chemotypes with desired properties. They are commonly classified as either ligand- or structure-based approaches but they are both based on a common central requirement linked to molecular similarity.

Indeed, this resemblance-based principle assumes that compounds are more likely to show similar biological activity if they share common molecular features. Numerous algorithms and methods have been developed and are currently used in drug discovery with the aim to find original molecules close to known actives with regard to their biological effects and/or physicochemical properties [3–7].

The first developed similarity search algorithms mostly considered chemical 1D and/or 2D information [8]. They perform very well [9], but they do not take into account the 3D conformation of the molecules which is known to be a critical parameter driving the protein–ligand recognition process. Thus, the next generation of algorithms shifted toward more complex methods with the aim of identifying molecules sharing common 3D patterns [10–20]. In this context, we [21–25] and other teams [26,27] developed shape matching approaches using spherical harmonic based representations of the molecular surface. These approaches mostly differ in the way they fit the spherical harmonic polynomials on the molecular surface description, obtained usually from conventional methods such as Conolly's one. Our method is particularly different as we fit the spherical harmonic polynomials onto a molecular surface obtained by deflating a preliminary meshed ellipsoid embracing the molecular structure of interest [28].

The use of molecular shape and its applicability to virtual screening for drug design has been emphasized and validated in numerous papers [29–51] and it is commonly admitted that ROCS [52] is one of the best performing 3D shape based methods according to its wide distribution in both academia and private research organizations.

Nonetheless it has been reported that molecular shape alone can not cover all the requirements necessary to obtain an efficient and selective measure of the molecular similarity between sets of compounds. Several attempts were reported describing the

[*] Corresponding author. Tel.: +33 354 958 520; fax: +33 383 278 319.
E-mail addresses: arnaud.karaboga@loria.fr (A.S. Karaboga), souchet@harmonicpharma.com (M. Souchet), bernard.maigret@loria.fr (B. Maigret).
[1] Tel.: +33 354 958 520; fax: +33 383 278 319.
[2] Tel.: +33 354 958 604; fax: +33 383 275 652.

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2013.01.003

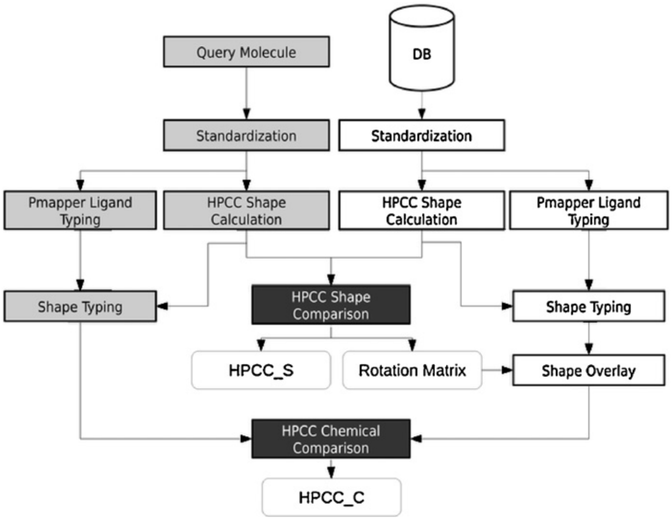


Fig. 1. Flowchart of HPCC similarity assessment (HPCC_S: shape coefficient; HPCC_C: chemistry coefficient).

addition of physicochemical criteria [53–56] like electrostatic potential to complement molecular descriptions when measuring molecular similarities.

Here, we present our HPCC algorithm, combining for the first time spherical harmonic based shapes and pharmacophoric features in a unique framework. The present method was assessed in a retrospective study using the DUD dataset [57] in order to evaluate the advantages and drawbacks with regard to ROCS software which is known to be a very successful 3D method in the VS context.

2. Methods

2.1. HPCC similarity approach

HPCC [24] uses spherical harmonic polynomial functions derived from atomic coordinates of a ligand to represent its molecular shape. These harmonic based surfaces are discretized as triangle meshes. The main idea of the HPCC algorithm is to assign a chemotype-like property to these triangles by using pharmacophoric features in a way that similarity calculation between molecules A and B consists in comparing the chemotype-like distribution associated to their respective shapes. Fig. 1 summarizes the steps involved in the HPCC similarity calculation procedure:

(1) *Standardization*: each chemical structure of the molecules of the data set is standardized using the *Standardizer* toolkit [58] from ChemAxon. This procedure is optional but recommended because the performance of the chemistry based similarity calculation depends on atomic type definition.

(2) *HPCC shape calculation*: a spherical harmonic shape is computed for each molecule of the data set using HPCC software as described in Ref. [21].

(3) *Pmapper ligand typing*: in parallel, for each of the ligands, each atom is associated to one or multiple pharmacophoric features using the default parameters of the *Pmapper* toolkit [59] from ChemAxon. We added a new tag, namely “u” for “undefined”, when *Pmapper* failed to characterize an atom type.

(4) *Shape typing*: for each of the ligands and its associated shape, triangles constituting the spherical harmonic based surface were associated with the same pharmacophoric property as the nearest ligand atom.

(5) *HPCC shape comparison (HPCC_S)*: each molecular shape is compared to the query molecule using HPCC. Coordinates of the query are kept fixed. The shape-only HPCC_S score is obtained for each molecule and the Euler angles alpha, beta, and gamma values of the rotation matrix are stored for the next step (for details see Ref. [24]).

(6) *Shape overlay*: the coordinates of the rotated molecule are updated by applying the stored HPCC_S rotation matrix. A translation is then applied to overlay the rotated and the query molecules at the same origin by using geometric centers.

(7) *Chemical overlay (HPCC_C)*: the pharmacophoric feature of each triangle of the rotated molecule surface is compared to those of the nearest triangle of the query molecule surface based on the previous shape overlay. A chemistry-only score, namely HPCC_C, is then calculated according to the similarity matrix defined in Supplementary Table 1 using a simple Tanimoto equation. The scoring ignores a triangle with an undefined feature. The HPCC_SC combo score (shape plus chemistry) is eventually calculated as the sum of shape-only HPCC_S and the chemistry-only HPCC_C scores.

2.2. ROCS method

ROCS program [52] uses the atomic coordinates of the ligand structures to calculate shape based similarity scores and the atom-centered Gaussian functions to define 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. ROCS focuses on shape and chemical overlays and therefore three types of scoring are compared: ROCS shape-only (ROCS_S), ROCS chemistry-only (ROCS_C) and ROCS shape plus chemistry or combo (ROCS_SC). The default Tanimoto metric is used for similarity calculation.

2.3. DUD data set

The active and decoy molecules for each of the 40 targets in the DUD data set Release 2 were downloaded from http://dud.docking.org/r2 and used to compare HPCC and ROCS methods. Table 1 gives a statistical overview of the data set.

We used the crystallographic ligand conformation provided by Huang et al. [57] as the query for the shape-based matching approaches, except for the following five targets: AR, PDGFRb, VEGFR2, ADA, COX-1, and Thrombin, for which the original query was replaced by a more suitable query molecule according to DUD errata or a more recent higher resolution crystal structure solved (see Supplementary Table 2). We removed duplicate from the actives and the decoys in our calculations. We also standardized and ionized at pH = 7.4 the whole DUD data set with Chemaxon Standardizer toolkit. Although tautomers [60] can affect VS results, we chose not to calculate the diverse tautomeric forms not to alter the 3D conformations of each of the DUD molecules and therefore to preserve the shape of the structures. The computation of the 40 DUD targets on a single 2.4 GHz Intel Xeon CPU takes about 9 min and 158 min for ROCS and HPCC, respectively. The HPCC computation time could be easily reduced by using multiple CPUs and/or GPU technology.

2.4. Performance metrics

There are several metrics for assessing VS performance [61–64] with receiver operator characteristic [65] (ROC) curves being one of the most commonly used. In practical VS studies, only a small fraction of a database can be tested experimentally, it is therefore crucial for a VS method to be able to recognize actives or leads as “early” as possible. According to this, we used the NSLR metric [66], a validated logarithmic metric which takes into account only the rank of active compounds in the data set, and is thus able to

Table 1
General statistics of the DUD data set showing the target name, the Protein Data Bank (PDB) code of the crystallographic ligand used as query, the number of decoys and actives for each target with and without duplicates, respectively.

Target	PDB code	#Decoys	#Decoys without duplicates	#Actives	#Actives without duplicates
GART	1c2t	879	863	40	31
HIVPR	1hpx	2038	1998	62	62
P38	1kv2	9141	9041	454	366
HMGA	1hw8	1480	1450	35	35
HSP90	1uy6	979	965	37	25
COX-2	1cx2	13,289	13,161	426	412
FGFR1	1agw	4550	4490	120	120
SRC	2src	6319	6217	159	159
PARP	1efy	1351	1331	35	35
EGFR	1m17	15,996	15,753	475	458
CDK2	1ckp	2074	2015	72	58
PDGFRb	1t46	5980	5904	170	169
PR	1sr7	1041	1019	27	27
PNP	1b8o	1036	1016	50	30
GR	1m2z	2947	2924	78	78
RXRa	1mvc	750	744	20	20
ER_ag	1l2i	2570	2517	67	67
VEGFR2	1fgi	2906	2849	88	78
ADA	1ndw	927	905	39	37
GPB	1a8i	2140	2114	52	52
INHA	1p44	3266	3232	86	86
Thrombin	3biu	2456	2425	72	68
MR	2aa2	636	630	15	15
COX-1	1q4g	911	908	25	25
ACE	1o86	1797	1789	49	49
COMT	1h1d	468	459	11	11
Trypsin	1bju	1664	1644	49	46
ALR2	1ah3	995	985	26	26
AR	2ao6	2854	2791	79	74
FXA	1f0r	5745	5550	146	146
PDE5	1xp0	1978	1972	88	76
ER_ant	3ert	1448	1434	39	39
DHFR	3dfr	8367	8146	410	408
HIVRT	1rt1	1519	1495	43	42
ACHE	1eve	3892	3867	107	106
SAHH	1a7a	1346	1311	33	33
NA	1a4g	1874	1866	49	49
TK	1kim	891	875	22	22
PPARg	1fm9	3127	3071	85	82
AMPC	1xgj	786	784	21	21
Total	40	124,413	122,510	3961	3743

assess both the early and the overall performance of a VS method. Details of ROC curves and NSLR are given below.

A ROC curve [67] is the plot of the true positive rate (TPR, or sensitivity) versus the false positive rate (FPR, or 1-specificity). Considering a rank i, two rates can be written as following:

$$\mathrm{TPR} = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} = \mathrm{Se}_i \tag{1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}_i}{\mathrm{TN}_i + \mathrm{FP}_i} = 1 - \mathrm{Sp}_i \tag{2}$$

where TP_i, FN_i, TN_i and FP_i are respectively the true positive, false negative, true negative, false positive numbers at threshold i and Se_i, Sp_i are the sensitivity and the specificity.

It may feel redundant but it helps to highlight that $(\mathrm{TP}_i + \mathrm{FN}_i)$ and $(\mathrm{TN}_i + \mathrm{FP}_i)$ are constant for any i and correspond to the data set number of actives and decoys, respectively.

The area under the curve (AUC) of a ROC plot is a scalar representation of the overall quality of the 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 [68]. The AUC is typically calculated for the whole of the ROC curve [69] using

$$\mathrm{AUC} = \frac{1}{n_q}\sum_{i=1}^{n_a} 1 - f_i = 1 - \frac{1}{n_q}\sum_{i=1}^{n_a} 1 - f_i \tag{3}$$

where f_i is the fraction of decoys ranked higher than the ith active and n_a is the total number of actives in the dataset. 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, AUC values do not distinguish early and late performance [69]. To have some information on the early recognition with this metric, we also report values of the AUC for the first 10% of the ROC curve.

The sum of logarithms of ranks (SLR) metric [66] is calculated as

$$\mathrm{SLR} = \sum_{i=1}^{n_a} \log \frac{r_i}{N} \tag{4}$$

where r_i is the rank of the ith 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

$$\mathrm{SLR}_{\mathrm{max}} = -\sum_{i=1}^{n_a} \log \frac{i}{N} \tag{5}$$

This allows a Normalized SLR (NSLR) to be calculated as:

$$\mathrm{NSLR} = \frac{\mathrm{SLR}}{\mathrm{SLR}_{\mathrm{max}}} \tag{6}$$

The above metric ranges from 1 (best achievable ranking) to a minimum value dependent on the considered set population

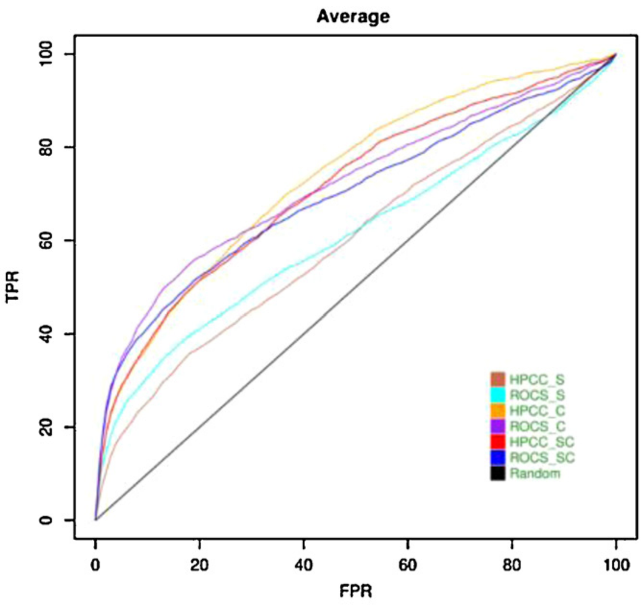


Fig. 2. Aggregate ROC plots for the 40 DUD targets.

through Eq. (4). In the present work, this value is so small that it can be approximated to zero, independently of the dataset. Consequently, our NSLR values can be considered absolute.

The consensus ranking score is calculated as

$$\mathrm{CRS}(i) = \frac{\mathrm{Avg}(r_i^m)}{\mathrm{Avg}(N, \max(r_i^m))} \tag{7}$$

Where r_i^m is the rank assigned by the mth method to the ith compound, $\mathrm{Avg}(r_i^m)$ gives the average rank to the ith compound considering all m methods, $\max(r_i^m)$ is the maximal (or highest) rank characterizing the ith compound considering all m methods, and N is the total number of compounds ranked. This metric ranges from $2/(N+1)$ when all methods rank the compound at the first place, to the maximal score of 1 when all methods rank the compound last.

3. Results

3.1. Assessment of the overall performance and early recognition by ROC curves and NSLR metric

We applied the above metrics to assess the performance of our new HPCC method compared to the standard ROCS approach using the DUD dataset. To provide the overall measure of how the methods performed across the 40 targets, we first calculated an aggregate ROC plot [70] for each of the six scoring methods by vertically averaging the individual ROC curves (see the Supporting Information for the 40 individual ROC curves). Fig. 2 shows that all methods perform significantly better than random (AUC > 0.5). To summarize further the overall performance of the six scoring schemes, Table 2 indicates the aggregate AUC values (Eq. (3)) obtained for the 40 targets. We observe that profiles of the chemistry-only HPCC_C and combo HPCC_SC scorings are globally similar to their corresponding counterpart ROCS_C and ROCS_SC with mean AUC values of 0.75, 0.73, 0.74 and 0.71, respectively. The same trend can be observed, with lower mean AUC values of 0.61 and 0.63, when comparing the two shape-only scoring methods HPCC_S and ROCS_S, respectively. It is worth noting that neither ROCS nor HPCC methods can reach the best achievable values of 0.10 for assessing early recognition on the first 10% of data set parsed (Table 2).

Table 2
Average VS performance for the DUD dataset: average AUC10%, AUC100% and NSLR rates for all 40 DUD targets using the standardized actives and decoys without duplicates.

Method	AUC10%	AUC100%	NSLR
HPCC_S	0.02	0.61	0.37
ROCS_S	0.02	0.63	0.41
HPCC_C	0.03	0.75	0.50
ROCS_C	0.03	0.74	0.53
HPCC_SC	0.03	0.73	0.49
ROCS_SC	0.03	0.71	0.52

To better discriminate the two methods according to both their overall performance and their ability to recognize actives at the very beginning of the ranked list, we calculated the NLSR metric (Eq. (6)) described by Venkatraman et al. [66]. With respective mean NSLR values of 0.50, 0.49, 0.53 and 0.52, the chemistry-only and combo methods HPCC_C, HPCC_SC, ROCS_C, ROCS_SC have similar NSLR scores when comparing between both HPCC and ROCS. The shape-only HPCC_S and ROCS_S follow the same tendency, but with lower NLSR scores (values of 0.37 and 0.41, respectively).

3.2. Target-centric assessment of VS performance

Fig. 3 represents three spiderdiagrams showing the comparative VS performance of the combo HPCC_SC and ROCS_SC (Fig. 3a), the chemistry-only HPCC_C and ROCS_C (Fig. 3b), and the shape-only HPCC_S and ROCS_S methods (Fig. 3c) for the 40 DUD targets. A consensus ranking score (Eq. (7)) is calculated by considering HPCC_SC and ROCS_SC individual ranks (see Table 3) transformed into a consensus rank that reflects both HPCC_SC and ROCS_SC performances. In order to visualize HPCC and ROCS respective performance, targets are ordered clockwise using decreasing consensus from FXA to COX-2 followed by increasing consensus ranks from SAHH to SRC. Each radial line represents a target. The intersection between a curve and a spoke gives the NSLR scores for the corresponding target.

Considering the shape plus chemistry score (Fig. 3a), NSLR scores are higher for HPCC_SC than ROCS_SC for 21 out of 40 targets, as displayed on the right-hand side of the spiderdiagram from FXA to COX-2. Thus, ROCS_SC gives better VS performance for the remaining 19 targets gathered in the left-hand side of the same diagram from SRC to SAHH. Targets positioned at the top of the Fig. 3a such as SRC, Trypsin, CDK2, PDGFRb, Thrombin, FXA, ALR2, FGFR1, GART and COMT, are those for which both ROCS_SC and HPCC_SC have poor VS performance. On the contrary, targets positioned at the bottom of the diagram, exemplified by COX-2, GPB, MR, HMGA, AR, SAHH, ER_ag, RXRa and NA, are those for which both HPCC_SC and ROCS_SC give a high VS performance.

Fig. 3b and c show that the chemistry-only methods achieve globally higher scores than the shape-only scoring schemes, for both HPCC and ROCS, emphasizing the idea that adding the pharmacophoric description to the shape is valuable and improves significantly the VS performance. Moreover, a qualitative comparison by scoring category shows that (i) when ROCS_SC gives better results than HPCC_SC (left-hand side of Fig. 3a), this relative performance is conserved for both the chemistry (left-hand side of Fig. 3b) and the shape (left-hand side of Fig. 3c) for all the targets if we do not consider the minor difference observed for some targets like PNP, (ii) similarly, for the 21 out of 40 targets for which HPCC_SC performs better than ROCS_SC (right-hand side of Fig. 3a), the same relative performance is retrieved with the chemistry-only (right-hand side of Fig. 3b) and shape-only (right-hand side of Fig. 3c) scoring schemes, except for HMGA, for which HPCC_S performs significantly worse than ROCS_S. This result confirms that the combo score is a good indicator of the global VS performance because it

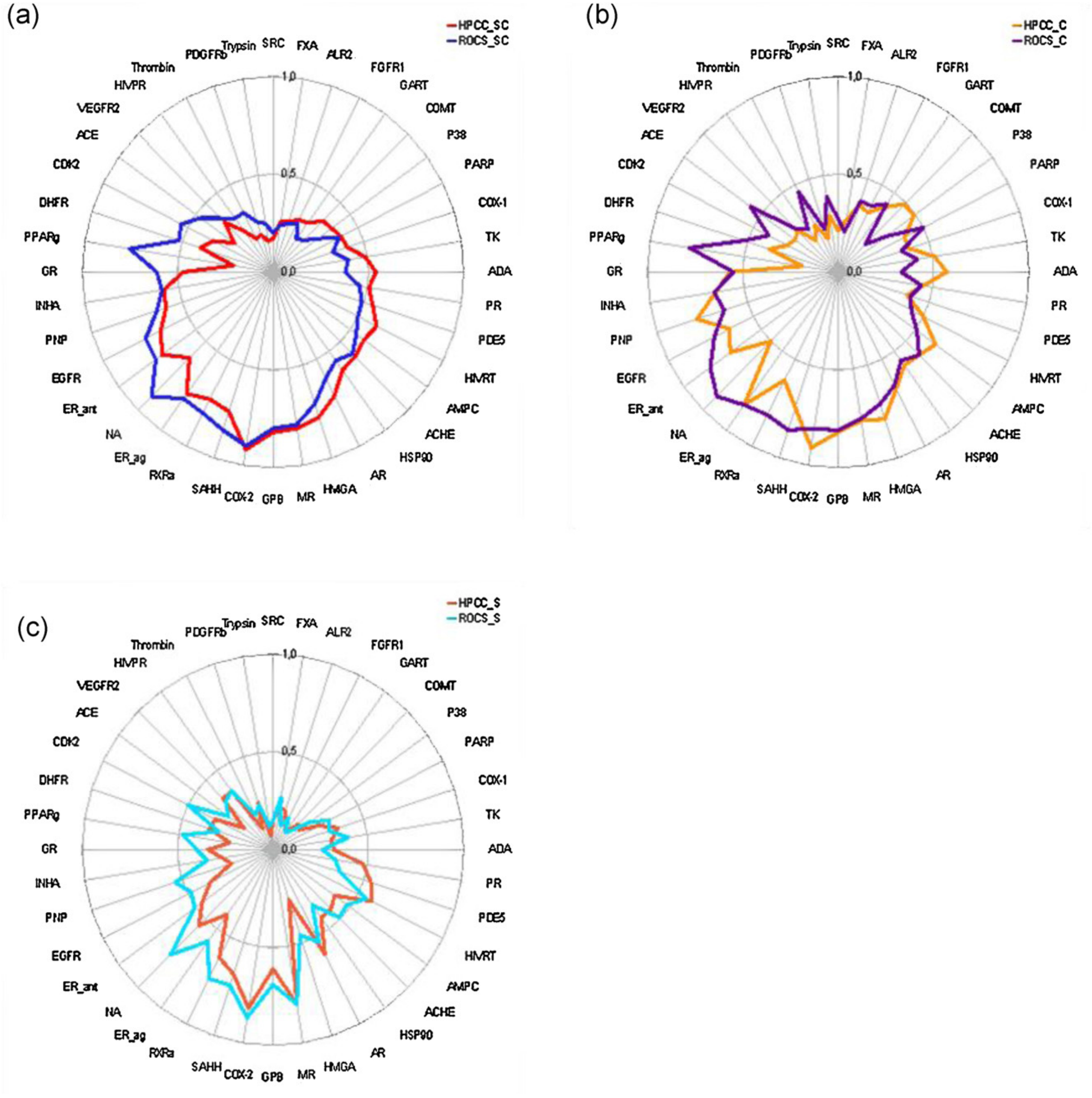


Fig. 3. Spiderdiagrams showing the comparative performance of (a) the combo HPCC_SC and ROCS_SC (b) the chemistry-only HPCC_C and ROCS_C and (c) the shape-only HPCC_S and ROCS_S methods for the 40 DUD targets. Each radial line represents a target. The intersection between a curve and a spoke gives the NSLR scores for the corresponding target.

averages the contribution of the shape-only plus the chemistry-only.

Focusing on the area between the two curves in the three spider plots, the difference observed on the left-hand side is globally similar to that on the right-hand side suggesting that when ROC_SC performs better than HPCC_SC, this difference is relatively similar to that observed for the targets for which HPCC_SC gives better VS performance than ROCS_SC. Nevertheless, this relative performance is more pronounced for few targets like PPARg and NA (left-hand side of Fig. 3a) and ADA (right-hand side of Fig. 3a).

To understand why HPCC or ROCS performed either similarly or differently in some cases, several representative targets were selected and are analyzed in more detail in the next section. For instance, COX-2 is a representative target for which the two methods show reasonably high VS performance with regard to both shape-only and chemistry scorings; Similarly, SRC is a representative target where both shape-only and chemistry scoring values are low; The target ADA illustrates the cases for which HPCC gives better results than ROCS thanks to a better shape-only and/or the chemistry-only contribution. Finally, NA is an example for which ROCS gives better VS performance than HPCC because of a better shape and/or pharmacophoric characterization.

3.3. *Detailed analysis of selected targets*

To analyze the respective early recognition and overall performance of each HPCC and ROCS scoring methods, individual ROC plots were generated (Fig. 4) and the individual AUC10%, AUC100% and NSLR rates were calculated for the chosen targets, i.e. COX-2, SRC, ADA and NA (Table 4). To help our analysis, Table 5a and b

Table 3
Individual NSLR values deriving from the shape-only (HPCC_S and ROCS_S), chemistry-only (HPCC_C and ROCS_C) and combo (HPCC_SC and ROCS_SC) methods and obtained with the 40 DUD targets. The consensus rank results from the consensus scoring of individual combo HPCC_SC and ROCS_SC ranks (Eq. (7)).

Targets	HPCC_S	ROCS_S	HPCC_C	ROCS_C	HPCC_SC	ROCS_SC	HPCC_SC Rank	ROCS_SC Rank	Consenus Rank
SRC	0.15	0.15	0.21	0.27	0.17	0.2	39	40	40
Trypsin	0.07	0.12	0.29	0.39	0.16	0.25	40	36	39
PDGFRb	0.25	0.24	0.18	0.23	0.2	0.27	37	34	38
Thrombin	0.12	0.2	0.27	0.46	0.19	0.34	38	30	33
HIVPR	0.37	0.37	0.2	0.27	0.26	0.34	34	29	32
VEGFR2	0.37	0.35	0.3	0.31	0.36	0.39	28	27	30
ACE	0.19	0.29	0.28	0.57	0.25	0.47	35	23	28
CDK2	0.34	0.5	0.3	0.41	0.35	0.54	29	19	24
DHFR	0.35	0.3	0.38	0.52	0.4	0.52	24	20	21
PPARg	0.23	0.48	0.19	0.79	0.21	0.76	36	9	19
GR	0.34	0.4	0.57	0.55	0.47	0.61	22	14	18
INHA	0.27	0.35	0.65	0.66	0.58	0.59	17	16	17
PNP	0.23	0.54	0.78	0.63	0.59	0.64	16	12	13
EGFR	0.37	0.48	0.64	0.73	0.66	0.75	10	10	10
ER_ant	0.46	0.51	0.7	0.83	0.72	0.76	8	8	9
NA	0.55	0.76	0.5	0.9	0.62	0.9	12	2	7
ER_ag	0.41	0.58	0.84	0.84	0.77	0.8	5	6	6
RXRa	0.62	0.74	0.62	0.82	0.73	0.81	7	4	5
SAHH	0.67	0.73	0.73	0.85	0.75	0.85	6	3	3
COX-2	0.82	0.87	0.91	0.81	0.92	0.9	1	1	1
GPB	0.61	0.69	0.82	0.81	0.82	0.8	2	5	2
MR	0.8	0.8	0.77	0.76	0.81	0.79	3	7	4
HMGA	0.27	0.46	0.79	0.71	0.78	0.71	4	11	8
AR	0.6	0.53	0.67	0.65	0.71	0.61	9	13	11
HSP90	0.43	0.35	0.59	0.56	0.62	0.56	11	17	12
ACHE	0.43	0.49	0.6	0.6	0.61	0.59	13	15	14
AMPC	0.4	0.48	0.63	0.51	0.59	0.54	15	18	15
HIVRT	0.58	0.55	0.5	0.43	0.61	0.5	14	21	16
PDE5	0.54	0.37	0.38	0.4	0.55	0.49	18	22	20
PR	0.48	0.33	0.45	0.44	0.51	0.46	20	24	22
ADA	0.32	0.26	0.57	0.33	0.54	0.38	19	28	23
TK	0.31	0.4	0.51	0.42	0.49	0.4	21	25	25
COX-1	0.36	0.31	0.39	0.35	0.41	0.33	23	31	26
PARP	0.27	0.33	0.39	0.5	0.4	0.39	25	26	27
P38	0.17	0.24	0.49	0.31	0.38	0.29	26	32	29
COMT	0.18	0.14	0.49	0.2	0.37	0.23	27	38	31
GART	0.11	0.11	0.41	0.43	0.31	0.21	30	39	34
FGFR1	0.15	0.18	0.34	0.38	0.3	0.28	31	33	35
ALR2	0.21	0.13	0.36	0.38	0.27	0.26	32	35	36
FXA	0.22	0.27	0.27	0.21	0.26	0.24	33	37	37

summarize the physicochemical and pharmacophoric properties of the queries, actives and decoys of the selected targets.

3.4. Analysis of COX-2 VS performance

The ROC plot for COX-2 (Fig. 4a) represents the best situation where all methods perform successfully with high retrieval rates for all metrics. For instance, AUC100% values span from 0.97–0.93 for HPCC and from 0.94 to 0.92 for ROCS; AUC10% is equal or greater than 0.05 for all methods implying a good overall and early recognition performance. Indeed, this conclusion is also confirmed by NSLR values that were not lower than 0.8 with any method analyzed.

Table 4
Individual AUC10%, AUC100% and NSLR rates for the selected targets COX2, SRC, ADA and NA with regard to the three HPCC and ROCS scoring methods.

Method	AUC10%	AUC100%	NSLR	AUC10%	AUC100%	NSLR
	COX-2			ADA		
HPCF_S	0.05	0.93	0.82	0.01	0.62	0.32
ROCS_S	0.06	0.94	0.87	0.00	0.55	0.26
HPCF_C	0.06	0.96	0.91	0.02	0.84	0.57
ROCS_C	0.05	0.92	0.81	0.02	0.54	0.33
HPCF_SC	0.06	0.97	0.92	0.02	0.76	0.54
ROCS_SC	0.06	0.94	0.90	0.01	0.65	0.38
	SRC			NA		
HPCF_S	0.00	0.42	0.15	0.03	0.85	0.55
ROCS_S	0.00	0.40	0.15	0.05	0.94	0.76
HPCF_C	0.00	0.54	0.21	0.03	0.63	0.50
ROCS_C	0.01	0.58	0.27	0.07	0.96	0.90
HPCF_SC	0.00	0.47	0.17	0.03	0.85	0.62
ROCS_SC	0.01	0.45	0.20	0.07	0.97	0.90

Visual inspection of COX-2 actives shows that they derive from several but very similar scaffolds, which explains why their physicochemical and pharmacophoric properties are close to those of the query (Table 5). Moreover, actives and decoys are sterically and chemically homogeneous and similar to the query, as seen with their similar average volume and average statistics, for both HPCC or ROCS pharmacophoric features. Consequently, Fig. 5a and b highlight the good shape overlays obtained using HPCC and ROCS for the top-ranked molecule ZINC03814719, which are very close to the query molecule except for a chlorine replacing a bromine atom and a methyl on the pyrazole ring. Thus, COX-2 is a target for which analog bias has a positive impact on VS performance, as already discussed [66].

3.5. Analysis of SRC VS performance

Fig. 4b shows the ROC curves for SRC, which illustrates a “difficult” case, where all methods are worse than random. The low values for AUC10%, AUC100% and NSLR retrieved for SRC are shown in Table 4. For instance, NSLR rates are lower than 0.30 (0.15–0.21 for HPCC and 0.15–0.27 for ROCS), highlighting and confirming that a low NSLR value is an indicator of a poor overall and early recognition performance.

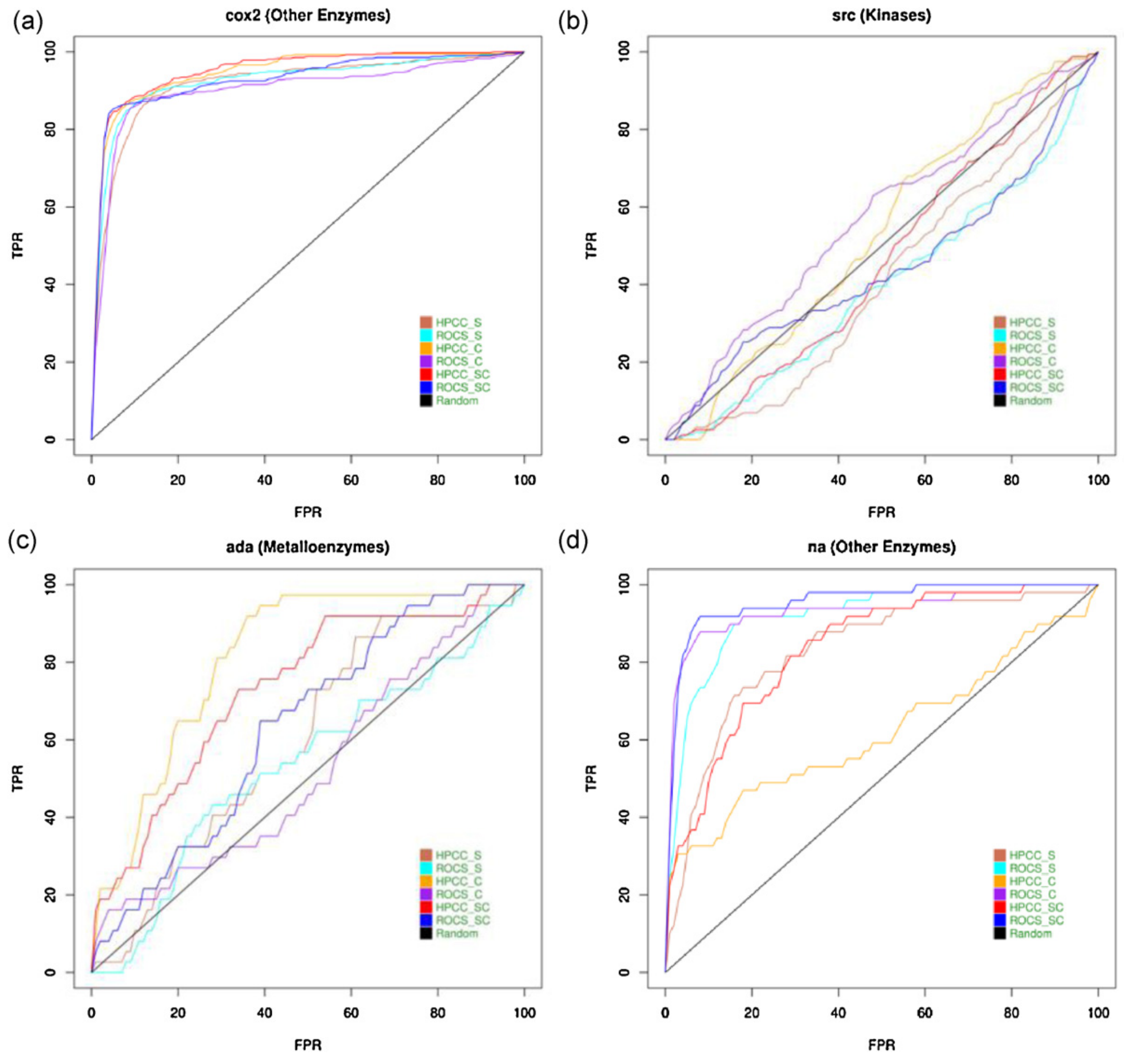


Fig. 4. ROC plots for the COX-2, SRC, ADA and NA targets.

As indicated in Table 5a and b, the actives for the SRC target show a high standard deviation regarding the surface (±60.17 and ±65.73 Å^2 for ROCS and HPCC, respectively) and the volume (±62.49 and ±69.29 Å^3 for ROCS and HPCC, respectively) meaning that the size of the active molecules are very different from that of the query molecule. Hence, the SRC query molecule is chemically not representative of the actives. As a consequence, pharmacophoric descriptions deriving from both HPCC and ROCS highlight a higher number of HBA on the query molecule compared to the respective actives and decoys; furthermore HPCC generated 12 negative charges for the query whereas the average charge is zero for actives and decoys. Altogether, the observed discrepancy between the SRC query molecule and the associated actives plus decoys could explain the poor 3D overlay obtained with both HPCC and ROCS, as exemplified in Fig. 5c and d.

3.6. Analysis of ADA VS performance

Fig. 4c shows the ROC curves for ADA, for which HPCC has a better global VS performance than ROCS, although the shape-only deriving scores of HPCC_S and ROC_S have a similar low VS performance reflected by NSLR values of 0.32 and 0.26, respectively. Adding the chemistry features results in a two-fold improvement of the performance for HPCC_C and has almost no impact for ROCS_C, with respective NSLR values of 0.57 and 0.33 (Table 4).

It is noteworthy that the main difference between HPCC and ROCS for this particular target concerns the aromatic, hydrophobic and ring features (Table 5a and b). Fig. 5e and f illustrate the impact of the different pharmacophoric description on the final overlay with the query molecule. Whereas HPCC maps the terminal benzyl ring of the query with an aromatic feature, ROCS uses the ring pattern, which is not fully suited for terminal or non-terminal aliphatic moieties. Consequently, HPCC_C allows aliphatic overlays on aromatic features according to our chemistry scoring scheme (see scoring in Supplementary Table 1), whereas ROCS_C seems not to align so well aliphatic or aromatic moieties onto a ring pattern leading to penalty on the chemistry score. This may explain why the HPCC performs better than ROCS for this particular target.

3.7. Analysis of NA VS performance

Considering the combo scores, NA is an example of target for which ROCS performs better than HPCC as the ROC curves show in Fig. 4d. The both shape-only ROCS_S and HPCC_S methods give somewhat similar medium performance, as shown by NSLR values of 0.76 and 0.55, respectively (Table 4). The main difference in the

Table 5
a Summary of the surface, volume and pharmacophoric features for ROCS: number of positive charge (POS) and negative charge (NEG) atoms, number of hydrogen bond acceptor (HBA) and hydrogen bond donor (HBD) atoms, number of aromatic (ARO) and hydrophobic (HYD) atoms is given for the Query (Q), decoy (D), and active (A) ligands associated with COX-2, SRC, ADA and NA targets. RINGS feature is specific to ROCS which also combines ARO and HYD features in single feature. b Summary of the surface, volume and pharmacophoric features for HPCC: for both methods, number of positive charge (POS) and negative charge (NEG) atoms, number of hydrogen bond acceptor (HBA) and hydrogen bond donor (HBD) atoms, number of aromatic (ARO) and hydrophobic (HYD) atoms is given for the query (Q), decoy (D), and active (A) ligands associated with COX-2, SRC, ADA and NA targets.

Target	Q/D/A	Surface ROCS	Volume ROCS	POS	NEG	HBA	HBD	RINGS	ARO/HYD
COX-2	Q	313.67	281.3	0	0	3	1	3	1
	D	299.54 ± 11.75	283.15 ± 13.62	0.07 ± 0.28	0.09 ± 0.29	3.47 ± 0.88	1.04 ± 0.62	3.35 ± 0.65	0.29 ± 0.50
	A	303.04 ± 26.45	280.11 ± 26.79	0.32 ± 0.47	0.07 ± 0.26	3.07 ± 0.89	0.62 ± 0.71	3.18 ± 0.53	0.20 ± 0.46
SRC	Q	349.34	343.76	1	1	14	3	3	0
	D	331.37 ± 14.10	313.50 ± 16.25	0.14 ± 0.39	0.09 ± 0.29	4.06 ± 1.03	1.91 ± 0.65	3.29 ± 0.69	0.36 ± 0.54
	A	333.06 ± 60.17	318.23 ± 62.49	0.41 ± 0.65	0.08 ± 0.30	4.13 ± 1.25	2.12 ± 1.14	3.75 ± 0.62	0.16 ± 0.43
ADA	Q	238.18	210.29	1	0	3	2	2	0
	D	231.73 ± 14.00	207.80 ± 13.74	0.22 ± 0.46	0.12 ± 0.38	4.77 ± 1.44	2.48 ± 0.98	1.93 ± 0.67	0.21 ± 0.48
	A	219.60 ± 28.21	196.85 ± 27.49	1.22 ± 0.62	0.00 ± 0.00	4.38 ± 1.62	3.38 ± 1.40	2.43 ± 0.89	0.59 ± 0.91
NA	Q	263.92	241.34	1	1	7	6	1	0
	D	262.25 ± 19.08	237.61 ± 19.61	0.46 ± 0.59	0.41 ± 0.56	4.98 ± 1.34	2.65 ± 1.16	1.86 ± 0.79	0.34 ± 0.57
	A	253.28 ± 42.27	226.58 ± 39.66	0.76 ± 0.43	1.00 ± 0.00	4.47 ± 1.26	2.63 ± 1.30	1.16 ± 0.42	0.92 ± 0.94

Target	Q/D/A	Surface HPCC	Volume HPCC	POS	NEG	HBA	HBD	ARO	HYD
COX-2	Q	307.49	287.38	0	0	3	1	17	5
	D	298.22 ± 11.77	281.79 ± 14.61	0.20 ± 0.56	0.00 ± 0.06	4.10 ± 1.19	1.24 ± 0.67	13.73 ± 2.69	6.81 ± 2.11
	A	297.55 ± 26.77	285.10 ± 27.88	0.17 ± 0.42	0.04 ± 0.19	3.89 ± 1.13	0.67 ± 0.69	16.19 ± 3.18	5.56 ± 2.54
SRC	Q	335.99	324.33	0	12	15	4	9	5
	D	331.83 ± 14.92	314.26 ± 18.93	0.37 ± 0.73	0.00 ± 0.07	4.60 ± 1.21	2.20 ± 0.63	14.51 ± 3.01	7.23 ± 2.20
	A	334.55 ± 65.73	324.07 ± 69.29	0.43 ± 0.65	0.08 ± 0.30	3.96 ± 1.36	2.15 ± 1.19	16.21 ± 3.51	7.75 ± 4.03
ADA	Q	233.85	211.9	0	0	3	2	11	5
	D	226.60 ± 14.34	203.74 ± 14.85	0.69 ± 1.03	0.02 ± 0.14	4.51 ± 1.28	2.63 ± 0.92	6.75 ± 2.99	6.18 ± 2.47
	A	214.82 ± 28.37	193.62 ± 27.84	0.73 ± 0.83	0.00 ± 0.00	4.38 ± 1.55	3.54 ± 1.54	6.57 ± 2.92	7.27 ± 2.20
NA	Q	254.36	236.96	3	1	7	6	0	12
	D	256.74 ± 19.47	234.63 ± 21.48	0.63 ± 0.83	0.48 ± 0.84	5.21 ± 1.39	2.92 ± 1.11	5.36 ± 3.11	8.96 ± 2.75
	A	248.20 ± 44.95	228.38 ± 45.58	1.06 ± 0.91	1.06 ± 0.42	4.49 ± 1.30	2.69 ± 1.30	1.96 ± 3.06	11.71 ± 4.54

performance between the two methods seems to be linked to the chemistry-only contribution, as pointed by NLSR values of 0.50 and 0.90 for HPCC_C and ROCS_C, respectively.

Looking at the molecular descriptors in Table 5a and b, it seems that the pharmacophoric description of the query molecule by ROCS is comparable to those of both actives and decoys, according to the low standard deviation values. On the other hand, the pharmacophoric description derived from HPCC seems to retrieve less homogeneous statistics when comparing the query molecule with the actives and decoys. For instance, HPCC description of NA actives has higher average of aromatic features and lower positively ionizable atoms than the query description, and also high standard deviations linked to aromatic, hydrophobic and positively ionizable atom features from the query description. Thus, this difference in pharmacophoric typing might explain the relative better performance of ROCS versus HPCC for NA target, which is illustrated in Fig. 5g and h by a better overlay with the query molecule when using ROCS.

4. Discussion

It is well admitted that the 3D molecular shape is a determinant factor driving the protein–ligand recognition process. Therefore, there is a continuous need of new methods providing improved accuracy in 3D description with the aim of identifying novel biologically active molecules in a ligand-based VS context.

Our new HPCC approach retrieves very satisfactory overall performances, similar to ROCS method, although its relative success varies with regard to the kind of biological targets considered. In the following sections we discuss our results according to the impact of the shape and pharmacophoric contributions in the VS performance.

4.1. Impact of the shape description in the VS performance

Both Gaussian (ROCS) and spherical harmonic based (HPCC) shape descriptions performed satisfactorily although poor performances were retrieved for a few DUD targets. Our results suggest that limitations could be linked to intrinsic biases of the DUD data set itself. Indeed, previous studies already highlighted these “difficult” targets for their poor VS performances using either 2D or 3D methods [66,71].

As a matter of fact the results obtained with the “difficult” cases could be improved by addressing the following issues:

First, the present new shape-based approach also found like in previous studies [66] that for some targets like SRC the query molecule is not well-suited. Indeed, the query shape is so different from the actives/decoys that it is quite impossible to have a good retrieval rate in a VS experience using 3D methodologies. Hence, we observed that for targets showing bad performance using ROCS or HPCC, one could (i) change the query molecule (ii) define new active/decoy data sets having 3D conformations (and therefore 3D shapes) resembling those of the current crystallographic ligand (iii) consider alternative approaches for similarity assessment (like 2D approaches) for those particular targets.

Secondly, the active molecule used as query is not sufficiently representative of all the actives present in the data set, due to possible several binding modes for this particular target. Therefore, one should ensure that the query molecule and the associated actives adopt a similar binding mode otherwise one should use multiple query molecules to cover the multiplicity of binding modes. This point was also argued by Kirchmairet al. [72] who clustered the DUD ligand sets by maximizing their dissimilarity based on EFCP4 fingerprints and consequently obtained 14% improved VS performance using the center of the clusters as query instead of the crystallographic ligands recommended in the DUD. The very recent work by Perez-Nueno et al. [73] corroborated Kirchmair’s findings.

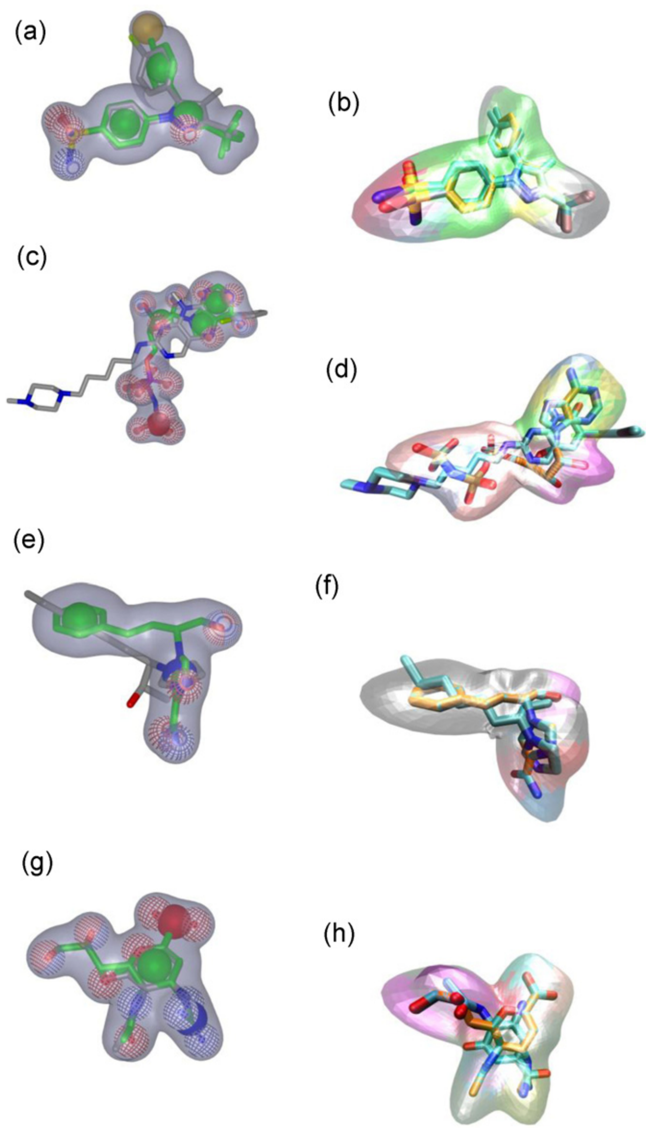


Fig. 5. Superposition of ZINC03814719, ZINC03815535, ZINC03814301 and ZINC03581100 molecules on their respective COX-2, SRC, ADA and NA query molecules using ROCS (panels a, c, e, g) and HPCC (panels b, d, f, h). For ROCS: the shape of the query molecule (C: green; N: dark blue; O: red; S: yellow, P: purple) is depicted in light blue; pharmacophoric features are represented as spheres (green: ring; blue: HBD; red: HBA; yellow: HYD); the superposed molecule is displayed with gray carbons (N: dark blue; O: red; S: yellow; halogen: light green). For HPCC: the surface of the shape is colored according to pharmacophoric features of the query molecule (a: dark red; d: dark blue; r: green; h: gray; a/r: yellow; a/d: magenta; +: red; −: blue; −/a: pink; +/d: tan; −/a/d: cyan; u: white); the superposed molecule is represented with carbons in cyan (N: dark blue; O: red; S: yellow; halogen: cyan).

Using consensus shape-clustering to encode several known high affinity ligands in a single representative pseudomolecular-shape, they showed better VS performance with targets for which using the single conformation of the crystallographic ligand failed. They applied the pseudomolecular-shape representation to two examples of targets, namely P38 and ALR2, that we also highlighted as “difficult” targets and which are characterized by a large binding pocket. Perez-Nueno et al. [73] pointed out that using a consensus-shape based query derived from the actives set improves the overall VS enrichment and helps to detect targets with multiple binding modes.

Beside the use of a shape-only approach for assessing VS performance, one of the aims of the present study was to demonstrate that combining shape-based molecular representations with our new pharmacophore based distribution improves significantly the retrieval rate of HPCC, as is also the case between ROCS shape-only and ROCS shape plus chemistry scoring. This is discussed in the next section.

4.2. *Impact of the pharmacophoric description in the VS performance*

First of all, although obvious it is worth mentioning that our new pharmacophore-based scoring provided high performances with targets showing already good shape-based performance, as illustrated by COX-2 (Fig. 5a and b). This observation confirmed and validated our new pharmacophoric-based scoring (Supplementary Table 1) as an appropriate function for similarity assessment in the same extent than those of ROCS.

Secondly, for targets associated with low shape-only VS performance adding the chemistry could either improve (GR, INHA, PNP and HMGA) or have no impact (SRC, PPARg) in the VS performance. The targets with improved VS performance proved that adding the chemistry helps to better distinguish false positives – with the same shape but a different pharmacophoric features distribution – from true positives – sharing both similar shape and chemistry – and therefore, underlines that the proposed HPCC chemistry-based scoring function is beneficial in a ligand-based VS context. The targets for which adding chemical properties had no impact on VS performance revealed the aforementioned limitations when using 3D methods with the DUD dataset. For example, as mentioned above, SRC is an example of target for which the molecular shape alone was not able to retrieve actives among decoys. Adding the pharmacophoric features did not substantially improve the VS performance, because the query molecule chemical description is quite different from the actives and decoys set, and therefore seems not to be an appropriate query to chemically discriminate actives from inactives. Finally, it seems clear that the chemistry scoring schemes for both HPCC and ROCS depend on the quality of the prior geometric overlay. Hence, introducing an additional re-alignment of the molecules based on the chemistry could lead to better VS performance of our method. However, this would involve additional computing time which seems not to be worthy given the already good results retrieved without the re-alignment step.

Altogether, it seems obvious that the quality of the chemistry-based similarity assessment is highly related to the accuracy of atom typing and the force-field used to define the atom-based pharmacophoric features of the ligands. The present work shows that the chemistry-based similarity assessment depends on the method used, as illustrated by the relative VS performance observed for ADA and NA targets. Indeed, although HPCC and ROCS use comparable description of the query shape, they gave different chemistry-based VS performance for those two targets. This result highlights the impact of the respective standardization process on the description of ligands, i.e. the query molecule, actives, and decoys. Hence, it is worth mentioning the importance of chemically standardizing properly any screening data set for achieving success in VS.

Since it is not possible to know in advance which methods is the most suitable for new targets, it is always necessary to use diverse and alternative VS approaches. As the present benchmark study highlighted that HPCC is globally as efficient as ROCS regarding the critical enrichment issue for the DUD dataset and even slightly better for some targets, we believe that HPCC is contributing to the progress in performance of the 3D shape-based methods.

5. Conclusion

We have developed a new efficient 3D ligand-centric descriptor, named harmonic pharma chemistry coefficient (HPCC), that combines the shape and pharmacophoric information of ligands, which can be used to perform VS. HPCC has been assessed using the 40 pharmaceutically DUD targets and results have been compared with ROCS approach. HPCC outperfomes ROCS for 21 out of the 40 DUD targets in terms of enrichment. To analyze further the target-dependent behavior of HPCC and ROCS, we have discussed the results obtained with four representative DUD targets.

Overall, the present work validates the mapping of pharmacophoric information onto the ligand surface as a way to improve significantly the VS performance for some targets compared to a shape-only approach. Our results show that there are many factors impacting the quality of the query molecules in terms of their ability to retrieve actives from decoys, and there is a need to continue exploring new original approaches which can tackle the still existent limitations of the current 3D shape-based methods. We think that the present HPCC approach contributes to this need.

Acknowledgements

The authors thank Dr. Violeta Perez-Nueno for her careful reading of the manuscript. The authors thank OpenEye Scientific Software Inc., for providing a license for ROCS and also ChemAxon for the license for the Marvin and Jchem toolkits. This work has been funded by the European Fund for Economic and Regional Development (FEDER), Communauté urbaine du Grand Nancy, and the Agence de Mobilization Economique de la Région Lorraine.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2013.01.003.

References

[1] Y.T. Tang, G.R. Marshall, Virtual screening for lead discovery, Methods in Molecular Biology 716 (2011) 1–22.

[2] A.L. Liu, G.H. Du, Research progress of virtual screening aided drug discovery, Yao Xue Xue Bao 44 (2009) 566–570.

[3] Y.C. Martin, J.L. Kofron, L.M. Traphagen, Do structurally similar molecules have similar biological activity? Journal of Medicinal Chemistry 45 (2002) 4350–4358.

[4] H. Eckert, J. Bajorath, Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches, Drug Discovery Today 12 (2007) 225–233.

[5] C.G. Wermuth, Similarity in drugs: reflections on analogue design, Drug Discovery Today 11 (2006) 348–354.

[6] A.G. Maldonado, J.P. Doucet, M. Petitjean, B.T. Fan, Molecular similarity and diversity in chemoinformatics: from theory to applications, Molecular Diversity 10 (2006) 39–79.

[7] J. Auer, J. Bajorath, Molecular similarity concepts and search calculations, Methods in Molecular Biology 453 (2008) 327–347.

[8] P. Willett, Similarity searching using 2D structural fingerprints, Methods in Molecular Biology 672 (2010) 133–158.

[9] E.J. Gardiner, J.D. Holliday, C. O'Dowd, P. Willett, Effectiveness of 2D fingerprints for scaffold hopping, Future Medicinal Chemistry 3 (2011) 405–414.

[10] A. Badel, J.P. Mornon, S. Hazout, Searching for geometric molecular shape complementarity using bidimensional surface profiles, Journal of Molecular Graphics 10 (1992) 205–211.

[11] B.K. Shoichet, I.D. Kuntz, Matching chemistry and shape in molecular docking, Protein Engineering 6 (1993) 723–732.

[12] M. Rosen, S.L. Lin, H. Wolfson, R. Nussinov, Molecular shape comparisons in searches for active sites and functional similarity, Protein Engineering 11 (1998) 263–277.

[13] R. Norel, H.J. Wolfson, R. Nussinov, Small molecule recognition: solid angles surface representation and molecular shape complementarity, Combinatorial Chemistry and High Throughput Screening 2 (1999) 223–237.

[14] D.A. Cosgrove, D.M. Bayada, A.P. Johnson, A novel method of aligning molecules by local surface shape similarity, Journal of Computer-Aided Molecular Design 14 (2000) 573–591.

[15] T. Kotani, K. Higashiura, Rapid evaluation of molecular shape similarity index using pairwise calculation of the nearest atomic distances, Journal of Chemical Information and Computer Science 42 (2002) 58–63.

[16] P.K. Agarwal, N.H. Mustafa, Y. Wang, Fast molecular shape matching using contact maps, Journal of Computational Biology 14 (2007) 131–143.

[17] V. Venkatraman, P.R. Chakravarthy, D. Kihara, Application of 3D Zernike descriptors to shape-based ligand similarity searching, Journal of Chemical Information 1 (2009) 19.

[18] Y.S. Liu, Y. Fang, K. Ramani, IDSS: deformation invariant signatures for molecular shape comparison, BMC Bioinformatics 10 (2009) 157.

[19] Y.S. Liu, Q. Li, G.Q. Zheng, K. Ramani, W. Benjamin, Using diffusion distances for flexible molecular shape comparison, BMC Bioinformatics 11 (2010) 480.

[20] P.J. Ballester, W.G. Richards, Ultrafast shape recognition to search compound databases for similar molecular shapes, Journal of Computational Chemistry 28 (2007) 1711–1723.

[21] W. Cai, X. Shao, B. Maigret, Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening, Journal of Molecular Graphics and Modelling 20 (2002) 313–328.

[22] W. Cai, J.W. Xu, X. Shao, B. Maigret, Molecular simulations using spherical harmonics, Chinese Journal of Chemistry 21 (2003) 1252–1255.

[23] M.E. Yamagishi, N.F. Martins, G. Neshich, W. Cai, X. Shao, A. Beautrait, B. Maigret, A fast surface-matching procedure for protein–ligand docking, Journal of Molecular Modeling 12 (2006) 965–972.

[24] W. Cai, J. Xu, X. Shao, V. Leroux, A. Beautrait, B. Maigret, SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces, Journal of Molecular Modeling 14 (2008) 393–401.

[25] A. Beautrait, V. Leroux, M. Chavent, L. Ghemtio, M.D. Devignes, M. Smail-Tabbone, W. Cai, X. Shao, G. Moreau, P. Bladon, J. Yao, B. Maigret, Multiple-step virtual screening using VSM-G: overview and validation of fast geometrical matching enrichment, Journal of Molecular Modeling 14 (2008) 135–148.

[26] R.J. Morris, R.J. Najmanovich, A. Kahraman, J.M. Thornton, Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons, Bioinformatics 21 (2005) 2347–2355.

[27] D.W. Ritchie, G.J.L. Kemp, Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces, Journal of Computational Chemistry 20 (1999) 383–395.

[28] W. Cai, M. Zhang, B. Maigret, New approach for representation of molecular surface, Journal of Computational Chemistry 19 (1998) 1805–1815.

[29] A.Y. Meyer, W.G. Richards, Similarity of molecular shape, Journal of Computer-Aided Molecular Design 5 (1991) 427–439.

[30] G.W. Bemis, I.D. Kuntz, A fast and efficient method for 2D and 3D molecular shape description, Journal of Computer-Aided Molecular Design 6 (1992) 607–628.

[31] B.S. Duncan, A.J. Olson, Shape analysis of molecular surfaces, Biopolymers 33 (1993) 231–238.

[32] A.C. Good, T.J. Ewing, D.A. Gschwend, I.D. Kuntz, New molecular shape descriptors: application in database screening, Journal of Computer-Aided Molecular Design 9 (1995) 1–12.

[33] J. Liang, H. Edelsbrunner, P. Fu, P.V. Sudhakar, S. Subramaniam, Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape, Proteins: Structure Function and Bioinformation 33 (1998) 1–17.

[34] M. Randic, Novel shape descriptors for molecular graphs, Journal of Chemical Information and Computer Science 41 (2001) 607–613.

[35] F. Weinhold, Chemistry. A new twist on molecular shape, Nature 411 (2001) 539–541.

[36] M.L. Mansfield, D.G. Covell, R.L. Jernigan, A new class of molecular shape descriptors. 1. Theory and properties, Journal of Chemical Information and Computer Science 42 (2002) 259–273.

[37] W.H. Sauer, M.K. Schwarz, Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity, Journal of Chemical Information and Computer Science 43 (2003) 987–1003.

[38] F. Fontaine, M. Pastor, F. Sanz, Incorporating molecular shape into the alignment-free grid-independent descriptors, Journal of Medicinal Chemistry 47 (2004) 2805–2815.

[39] Y. Zyrianov, Distribution-based descriptors of the molecular shape, Journal of Chemical Information and Modeling 45 (2005) 657–672.

[40] S. Putta, P. Beroza, Shapes of things: computer modeling of molecular shape in drug discovery, Current Topics in Medicinal Chemistry 7 (2007) 1514–1524.

[41] L. Mavridis, B.D. Hudson, D.W. Ritchie, Toward high throughput 3D virtual screening using spherical harmonic surface representations, Journal of Chemical Information and Modeling 47 (2007) 1787–1796.

[42] V.I. Perez-Nueno, D.W. Ritchie, J.I. Borrell, J. Teixido, 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, Journal of Chemical Information and Modeling 48 (2008) 2146–2165.

[43] V.I. Perez-Nueno, D.W. Ritchie, O. Rabal, R. Pascual, J.I. Borrell, J. Teixido, Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand–receptor docking, Journal of Chemical Information and Modeling 48 (2008) 509–533.

[44] J.A. Wilson, A. Bender, T. Kaya, P.A. Clemons, Alpha shapes applied to molecular shape characterization exhibit novel properties compared to established shape descriptors, Journal of Chemical Information and Modeling 49 (2009) 2231–2241.

[45] S. Kortagere, M.D. Krasowski, S. Ekins, The importance of discerning shape in molecular pharmacology, Trends in Pharmacological Sciences 30 (2009) 138–147.

[46] A. Nicholls, G.B. McGaughey, R.P. Sheridan, A.C. Good, G. Warren, M. Mathieu, S.W. Muchmore, S.P. Brown, J.A. Grant, J.A. Haigh, N. Nevins, A.N. Jain, B. Kelley,

Molecular shape and medicinal chemistry: a perspective, Journal of Medicinal Chemistry 53 (2010) 3862–3886.

[47] J.O. Ebalunode, W. Zheng, Molecular shape technologies in drug discovery: methods and applications, Current Topics in Medicinal Chemistry 10 (2010) 669–679.

[48] P.J. Ballester, I. Westwood, N. Laurieri, E. Sim, W.G. Richards, Prospective virtual screening with ultrafast shape recognition: the identification of novel inhibitors of arylamine *N*-acetyltransferases, Journal of the Royal Society, Interface 7 (2009) 335–342.

[49] H. Li, J. Huang, L. Chen, X. Liu, T. Chen, J. Zhu, W. Lu, X. Shen, J. Li, R. Hilgenfeld, H. Jiang, Identification of novel falcipain-2 inhibitors as potential antimalarial agents through structure-based virtual screening, Journal of Medicinal Chemistry 52 (2009) 4936–4940.

[50] S.W. Muchmore, A.J. Souers, I. Akritopoulou-Zanze, The use of three-dimensional shape and electrostatic similarity searching in the identification of a melanin-concentrating hormone receptor 1 antagonist, Chemical Biology and Drug Design 67 (2006) 174–176.

[51] E. Naylor, A. Arredouani, S.R. Vasudevan, A.M. Lewis, R. Parkesh, A. Mizote, D. Rosen, J.M. Thomas, M. Izumi, A. Ganesan, A. Galione, G.C. Churchill, Identification of a chemical probe for NAADP by virtual screening, Nature Chemical Biology 5 (2009) 220–226.

[52] J.A. Grant, M.A. Gallardo, B.T. Pickup, A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape, Journal of Computational Chemistry 17 (1996) 1653–1666.

[53] A. Nicholls, J.A. Grant, Molecular shape and electrostatics in the encoding of relevant chemical information, Journal of Computer-Aided Molecular Design 19 (2005) 661–686.

[54] A. Jennings, M. Tennant, Selection of molecules based on shape and electrostatic similarity: proof of concept of "electroforms", Journal of Chemical Information and Modeling 47 (2007) 1829–1838.

[55] M.J. Vainio, J.S. Puranen, M.S. Johnson, ShaEP: molecular overlay based on shape and electrostatic potential, Journal of Chemical Information and Modeling 49 (2009) 492–502.

[56] M.S. Armstrong, G.M. Morris, P.W. Finn, R. Sharma, L. Moretti, R.I. Cooper, W.G. Richards, ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics, Journal of Computer-Aided Molecular Design 24 (2010) 789–801.

[57] N. Huang, B.K. Shoichet, J.J. Irwin, Benchmarking sets for molecular docking, Journal of Medicinal Chemistry 49 (2006) 6789–6801.

[58] Standardizer, JChem version 5.4, 2011, Chemaxon, Budapest, Hungary.

[59] Pmapper, JChem version 5.4, 2011, Chemaxon, Budapest, Hungary.

[60] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, T. Langer, Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection – what can we learn from earlier mistakes? Journal of Computer-Aided Molecular Design 22 (2008) 213–228.

[61] J.F. Truchon, C.I. Bayly, Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem, Journal of Chemical Information and Modeling 47 (2007) 488–508.

[62] W. Zhao, K.E. Hevener, S.W. White, R.E. Lee, J.M. Boyett, A statistical framework to evaluate virtual screening, BMC Bioinformatics 10 (2009) 225.

[63] R.P. Sheridan, Alternative global goodness metrics and sensitivity analysis: heuristics to check the robustness of conclusions from studies comparing virtual screening methods, Journal of Chemical Information and Modeling 48 (2008) 426–433.

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

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

[66] V. Venkatraman, V.I. Perez-Nueno, L. Mavridis, D.W. Ritchie, Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods, Journal of Chemical Information and Modeling 50 (2010) 2079–2093.

[67] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861–874.

[68] P.C. Hawkins, G.L. Warren, A.G. Skillman, A. Nicholls, How to do an evaluation: pitfalls and traps, Journal of Computer-Aided Molecular Design 22 (2008) 179–190.

[69] M.D. Mackey, J.L. Melville, Better than random? The chemotype enrichment problem, Journal of Chemical Information and Modeling 49 (2009) 1154–1162.

[70] F.J. Provost, T. Fawcett, R. Kohavi, The case against accuracy estimation for comparing induction algorithms, in: J.W. Shavlik (Ed.), Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., Madison, Wisconsin, USA, 1998, pp. 445–453.

[71] D. Giganti, H. Guillemain, J.L. Spadoni, M. Nilges, J.F. Zagury, M. Montes, Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment, Journal of Chemical Information and Modeling 50 (2010) 992–1004.

[72] J. Kirchmair, S. Distinto, P. Markt, D. Schuster, G.M. Spitzer, K.R. Liedl, G. Wolber, How to optimize shape-based virtual screening: choosing the right query and including chemical information, Journal of Chemical Information and Modeling 49 (2009) 678–692.

[73] V.I. Perez-Nueno, D.W. Ritchie, Using consensus-shape clustering to identify promiscuous ligands and protein targets and to choose the right query for shape-based virtual screening, Journal of Chemical Information and Modeling 51 (2011) 1233–1248.