Electron Density Fingerprints (EDprints): Virtual Screening Using Assembled Information of Electron Density

Albert J. Kooistra,[†,‡,¶] Thomas W. Binsl,[‡,¶] Johannes H. G. M. van Beek,[§] Chris de Graaf,[†] and Jaap Heringa*[,‡]

Leiden/Amsterdam Center for Drug Research (LACDR), Division of Medicinal Chemistry, Department of Chemistry and Pharmaceutical Sciences, Faculty of Sciences, VU University Amsterdam, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands, Centre for Integrative Bioinformatics VU (IBIVU), Faculty of Sciences and Faculty of Earth & Life Sciences, VU University Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands, and Department of Clinical Genetics, Section Medical Genomics, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands

Received July 8, 2010

We have designed a method to encode properties related to the electron densities of molecules (calculated ^{1}H and ^{13}C NMR shifts and atomic partial charges) in molecular fingerprints (EDprints). EDprints was evaluated in terms of their retrospective virtual screening accuracy against the Directory of Useful Decoys (DUD) and compared to the established ligand-based similarity search methods MOLPRINT 2D and FCFP-4. Although there are no significant differences in the overall virtual screening accuracies of the three methods, specific examples highlight interesting differences between the new EDprints fingerprint method and the atom-centered circular fingerprint methods of MOLPRINT 2D and FCFP-4. On one hand, EDprints similarity searches can be biased by the molecular protonation state, especially when reference ligands contain multiple ionizable groups. On the other hand, EDprints models are more robust toward subtle rearrangements of chemical groups and more suitable for screening against reference molecules with fused ring systems than MOLPRINT 2D and FCFP-4. EDprints is furthermore the fastest method under investigation in comparing fingerprints (average 56−233-fold increase in speed), which makes it highly suitable for all-against-all similarity searches and for repetitive virtual screening against large chemical databases of millions of compounds.

INTRODUCTION

Virtual screening (VS) is an increasingly important method for the identification of novel bioactive molecules, complementary to experimental high-throughput screening methods.[1] While protein-based VS requires (structural) knowledge of the ligand binding pocket of the target, ligand-based VS searches are based on chemical similarity to known reference ligands. The growing size of chemical databases[2,3] and the need for all-against-all molecule comparisons for chemogenomics profiling[4] require faster and faster methods for chemical similarity comparisons. To detect molecular similarity, molecular features can be explicitly aligned[5,6] or compared by encoding them in molecular fingerprints.[7]

Four classes of molecular fingerprint descriptors can be distinguished:[8] (i) circular fingerprints considering the absence/presence of features, (ii) circular fingerprints considering counts, (iii) path-based/keyed fingerprints, and (iv) pharmacophore-based fingerprints. Circular fingerprints (like MOLPRINT 2D[9,10] and Pipeline Pilot's circular fingerprints)[11] are based on the absence/presence (i) or the number (ii) of unique atom-centered circular substructures consisting of a specific combination of predefined atom types (each bit in the fingerprint representing a structural key). Path fingerprints (iii) (e.g., Pipeline Pilot's FPFP)[11] are based on atom-type pairs and the atom count of the shortest path separating them, while pharmacophore fingerprints (e.g., MOE)[12] are derived from combinations of three or four pharmacophoric points and their interindividual distances. We currently present a new fingerprint descriptor based on numerical values reflecting the molecular electron density distribution, rather than on molecular topology. Chemical shifts (expressed as the relative frequency of NMR spectroscopy signals) and partial atomic charges are selected as atom-based descriptors of molecular electron density and combined in our novel EDprints method. Unlike most molecular fingerprints, EDprints uses numerical values derived from the molecules instead of explicitly encoding substructures, paths, or atom features. Although EDprints is related to circular fingerprint type i, it should be noted that our method does not include explicit information on atom types nor circular connectivity. NMR shifts and partial atomic charges nevertheless depend on molecular topology and implicitly reflect the presence and relative location of functional groups in molecules.

We have tested and cross-validated our method on the 40 data sets of the Directory of Useful Decoys (DUD),[13] which contains for every active ligand a set of 36 inactive decoys with similar physical properties but different topology. The DUD data set has previously been used to validate several ligand- and structure-based VS methods.[13−17] This makes DUD a suitable database for comparing VS approaches.[14,18]

* Corresponding author. E-mail: heringa@few.vu.nl.
[†] LACDR.
[‡] IBIVU.
[§] VU University Medical Center.
[¶] These authors contributed equally to this work.

10.1021/ci1002608
Published on Web 09/14/2010

Table 1. Transformation Rules for ^{13}C and ^{1}H Chemical Shifts and Partial Charges into Non-Negative Integers

value (x)	range	conversion	new range
^{13}C chemical shift	$x \in [-20, 230)$	$x + 20$	[0, 250)
^{1}H chemical shift	$x \in [0.0, 15.0)$	$x \cdot 10$	[0, 150)
partial charge	$x \in [-2.00, 2.00)$	$\lvert x \rvert < 1 \rightarrow (x + 1) \cdot 100 \cdot 2$ $x \geq 1.00 \rightarrow x \cdot 100 \cdot 2$ $x \leq -1.00 \rightarrow (x + 2) \cdot 100 \cdot 2$	[0, 400)

We compared the results of EDprints to the results of two other two-dimensional (2D) ligand-based similarity methods, namely MOLPRINT 2D[9,10] and the FCFP-4 circular fingerprint in Pipeline Pilot,[11] programs which have been frequently used in VS applications[19−22] and belong to the circular fingerprint descriptor class i (related to our new EDprints method).[8] Our study shows that EDprints is able to achieve VS accuracies as high as other ligand-based 2D similarity methods. Specific examples however highlight interesting differences between the new EDprints fingerprint method and the atom-centered circular fingerprint methods of MOLPRINT 2D and FCFP-4.

EDprints is furthermore the fastest method under investigation for comparing fingerprints, which makes it highly suitable for all-against-all similarity searches and for repetitive VS against large chemical databases of millions of compounds.

METHODS

EDprints: Encoding of Electron Density Information. Chemical shifts in frequency of NMR spectroscopy signals are determined by the electron density present at the particular H and C atoms as well as partial atomic charges within the molecule, which are caused by the asymmetric distribution of electrons in chemical bonds. Therefore, both values represent their particular molecular environment and are useful representations of the chemical and structural properties of molecules.

Chemical shifts of ^{13}C and ^{1}H atoms and Merck molecular force field (MMFF94)-type nonpolarized partial atomic charges were calculated for all molecules in the DUD database[13] using BatchNMRPrediction[23−25] and Balloon,[26] respectively. MMFF94 atomic partial charges,[27] have been successfully used for the construction of 3D quantitative structure−activity relationship (QSAR) modeling[28,29] and ligand-based VS[30,29] and are considered as good alternatives to more computationally expensive semiempirical and electrostatic potential fit charge models.[28,29] MMFF94 charges were furthermore found to be more suitable than Gasteiger−Marsili[31] for EDprints-based VS (data not shown).

Both BatchNMRPrediction and Balloon use molecule information stored in the structure data format (SDF), which is a common format for compound databases like DUD. We found that ^{13}C and ^{1}H chemical shifts and partial charges are mostly present in ranges of [−20, 230), [0, 15), and [−2, 2) ppm, respectively (see Supporting Information, Figure 1). Transforming these value ranges into non-negative integers (see Table 1), we were able to derive a binary substring for each of the electron density descriptors (^{13}C and ^{1}H chemical shifts and partial charges), where each bit position reflects a particular descriptor value. The transformation of the descriptor values into corresponding positions in the bit string was done differently for all three descriptors and is explained.

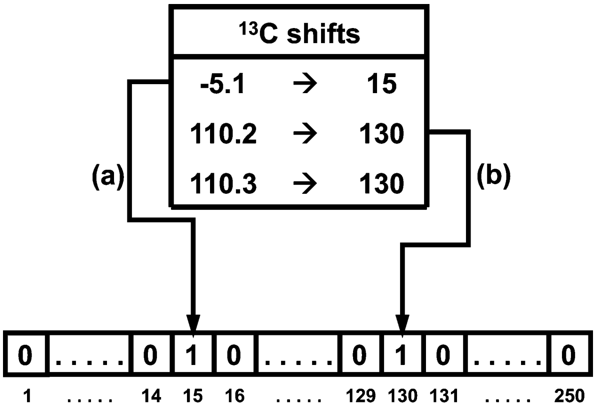


Figure 1. Encoding of a molecules ^{13}C chemical shift values into a bit string. The shift values (in ppm) are transformed according to the rule given in Table 1, and the corresponding bit position is set to '1' (see a and b). Multiple appearances of a shift value are ignored.

Encoding molecular properties into bit strings representing molecules is not a trivial task, and there are endless different encoding algorithms possible.[32] The major obstacles are equal molecular properties (with equal electron density information) that appear multiple times within the same molecule (e.g., multiple carboxyl groups). These multiple appearances compete for the same position in the bit string and can either be ignored or given special treatment. Possible solution scenarios comprise the reservation of multiple bits for a particular molecular property or use hash functions which are mathematical constructions to treat such competition.[33] During the analysis of the DUD database, we found that the value ranges and the number of unique ^{13}C and ^{1}H chemical shifts and partial charges per molecule are different. Hence, the individual fingerprints are encoded in different ways, as presented in the following paragraph.

An analysis of the ^{13}C chemical shifts showed that due to the huge range of possible shift values (−20 and 230 ppm), the shifts of carbons in different chemical environments mainly differ in the integer part. Hence, a value transformation by rounding the shift value and adding a value of 20 was sufficient to cover all bit positions of a bit string with length 250 (Table 1). For instance, a ^{13}C chemical shift value of −15 corresponds with the bit string position 5. During the encoding process of a molecule for each appearing value, the corresponding bit position is set to '1' indicating the presence of the particular chemical shift (Figure 1). However, multiple appearances of a particular ^{13}C chemical shift turned out to have no, or only a small positive, or sometimes even a small negative effect on the enrichment achieved in the VS process (data not shown) and are therefore ignored in the encoding process.

In contrast to the range of ^{13}C shift values, ^{1}H chemical shifts have a shorter range than ^{13}C shifts; they vary roughly between 0 and 15.0 ppm. Hence, the main differences among ^{1}H shift values in different chemical environments were discovered up to the first decimal. Therefore, a value transformation of the ^{1}H chemical shifts into corresponding bit positions includes a multiplication by 10 (Table 1), and the transformed values range from 0−150. However, a molecule is more likely to contain identical ^{1}H shifts than ^{13}C shifts, due to the much higher number of ^{1}H atoms in a molecule on average. This made it necessary to provide more

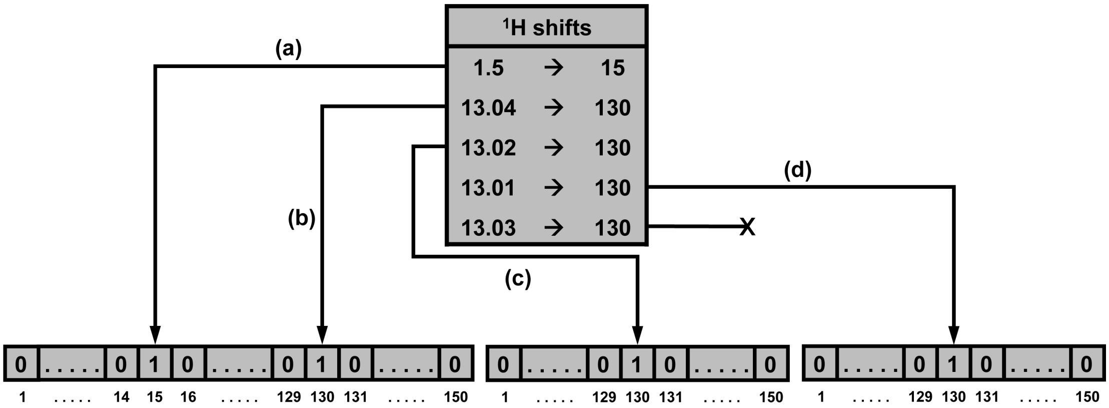


Figure 2. Encoding of a molecules ^{1}H chemical shift values into three bit strings. The shift values (in ppm) are transformed according to the rule given in Table 1, and the corresponding bit position is set to '1' (see a and b). Multiple appearances of a shift value are encoded in the second and third bit string, respectively (see c and d). Values appearing more than three times are discarded.

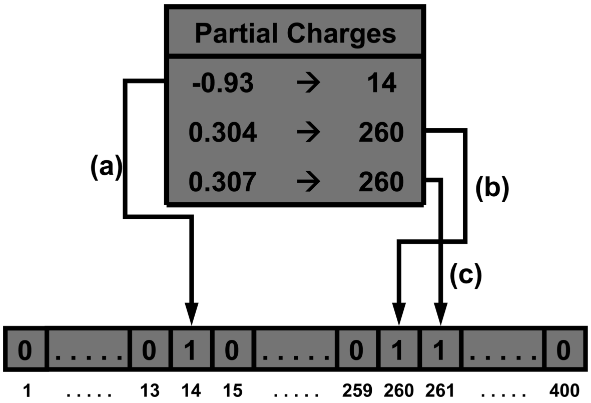


Figure 3. Encoding of a molecule's partial charge values into a bit string. The charge values are transformed according to the rule given in Table 1, and the corresponding bit position is set to '1' (see a and b). Multiple appearance of a charge value is encoded in the next upstream bit (see c) up to a maximum of five appearances. Values appearing more than five times are discarded.

than one bit for each ^{1}H shift value. Hence, three bit strings are used to encode ^{1}H shift values providing in total three alternative positions for each value (Figure 2). Whenever a bit is already set to '1' due to a particular shift value and this value appears a second or even a third time, the corresponding bit in the second and third bit string is set, respectively. However, values appearing more than three times are discarded.

The partial charges found among the molecules of the DUD database were in a range between −2 and 2, and the main differences in these values were discovered up to the second decimal. Therefore, a value transformation into integers includes a multiplication of 100 (Table 1). Additionally, partial charges show a standard distribution around the mean value 0. Because partial atomic charges above 1 and below −1 are rare and reserving unique bit positions for these values would increase the fingerprint size (increasing the overhead and decreasing the performance), these values are shifted to the −1 to 1 range (Table 1). Like for ^{1}H chemical shifts, identical values for partial charges appear more often within a particular molecule than identical ^{13}C chemical shifts. Although using the same encoding procedure as developed for ^{1}H shifts would have been possible, a slightly modified algorithm was used that made it feasible to encode even more identical values at unique bit positions (Figure 3). In particular, whenever a bit has already been set due to a partial charge value, the first available upstream bit is set when the same value appears another time. However, not more than a maximum of five upstream bits are set in this way. Due to the properties of the partial charges, several partial charges are present in almost every compound. To value the bit positions according to their relevance, the bit positions were assigned weights. Based on partial charge fingerprints for every molecule contained in the "everything" subset of the ZINC database,[2] weights were calculated for each position. We used a slightly adapted version of the "term frequency-inverse document frequency" formula[34] (eqs 1 and 2), which is widely used in information retrieval and text mining, to calculate these weights. The term frequency (n/N) was inverted to rate less frequently occurring bit positions higher than more common bit positions, thereby weighting their corresponding partial charge values higher. Instead of document and term frequency, the bit frequency was used in this formula, therefore n is the number of times the bit at this position is set, and N is the total number of fingerprints. A variable k, containing the additive inverse of the minimum value rounded up, was used to convert the weights into positive numbers only. Where the few bit positions that had a 0 count, a bit count of 1 was set in order to calculate its weight. The sum of the weights is used in the McConnaughey similarity measure (see Scoring Similarity Section) instead of the bit counts.

$$f(x) = \log_2(m) \Leftarrow (2^{m-1} < x \leq 2^m) \tag{1}$$

$$(1 - n/N) \cdot (f(N) - f(n) + 1 + k) \tag{2}$$

During a VS the final similarity score between two compounds is computed by weighting the individual similarities of the three different descriptors (Figure 4) described above. The weighting factors were determined in a 10-fold cross-validation (data not shown) on the DUD data sets (25% ^{13}C shifts, 25% ^{1}H shifts, and 50% partial charges). In this way, for comparison of compounds A and B, three different similarities $^{13C}\mathrm{Sim}_{A,B}$, $^{1H}\mathrm{Sim}_{A,B}$ and $^{PC}\mathrm{Sim}_{A,B}$ can be calculated for the particular molecular descriptors ^{13}C and ^{1}H shifts and partial charges, respectively. The total similarity between A and B is then given as the weighted sum of the three individual similarities:

$$\mathrm{Sim}_{A,B} = 0.25 \cdot {}^{13C}\mathrm{Sim}_{A,B} + 0.25 \cdot {}^{1H}\mathrm{Sim}_{A,B} + 0.5 \cdot {}^{PC}\mathrm{Sim}_{A,B} \tag{3}$$

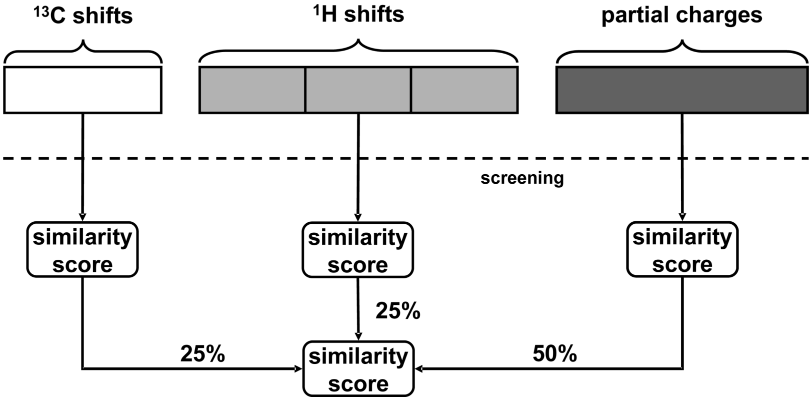
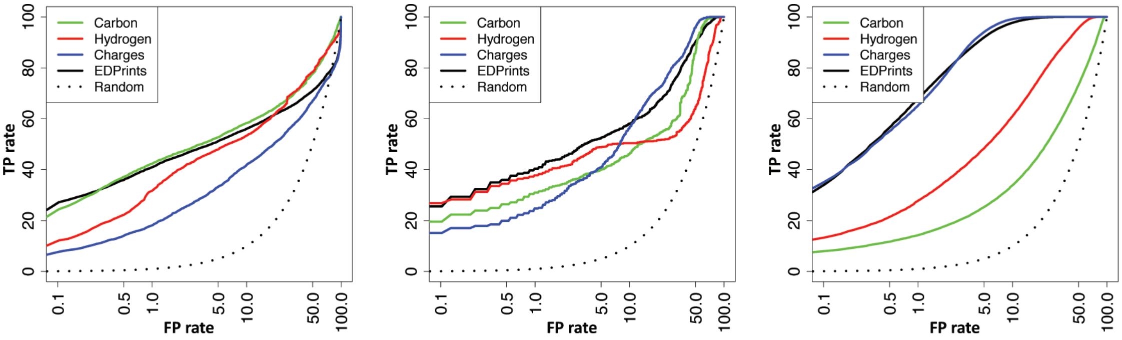


Figure 4. Combination of the individual fingerprint similarities into a final similarity score.



(a) ACHE dataset (b) ADA dataset (c) DHFR dataset

Figure 5. VS enrichment curves of individual ^{1}H and ^{13}C NMR shifts and atomic partial charge fingerprints vs EDprints models against ACHE (a), ADA (b), and DHFR (c) ligand data sets.

Scoring Similarity. VSs were performed on all 40 data sets of the DUD database (3209 molecules on average per data set) using EDprints and compared to screening results achieved with MOLPRINT 2D and FCFP-4. Each method was applied on each data set in the following way: Each active compound of a particular data set was screened against both the remaining active compounds and the decoys contained in this data set. The resulting similarity scores were sorted in decreasing order and subsequently used to compute the average enrichment factor for the data set. The similarity measure used for the similarity calculation between two bit strings was the one developed by McConnaughey[35] and given as

$$\frac{(a+c)\cdot c+(b+c)\cdot c-(a+c)\cdot(b+c)}{(a+c)\cdot(b+c)}=\frac{c^2-a\cdot b}{(a+c)\cdot(b+c)} \qquad (4)$$

where a and b are the number of bits uniquely present in the first and second bit string, respectively, and c the number of bits commonly set in both bit strings. The similarity measure ranges between −1 and 1 and particularly emphasizes the number of bits concurrently set in both bit strings with respect to the number of unique bits and the total number of active bits. Hence, this similarity measure is suitable for comparison of molecules of different sizes.

VS Accuracy Analysis. VS accuracies were first determined in terms of area under the curve (AUC) of receiver−operator characteristic (ROC) plots and its 95% confidence interval computed using the R-package DiagnosisMed.[36] AUC values are a measure for the overall VS accuracy. Enrichment EF_x in true positives (TP) is reported at different false-positive (FP) rates (x) as follows:

$$EF_x = \frac{TP}{FP_x} \qquad (5)$$

Early enrichments[37] at 0.5, 1, 2, and 5% FP rates were computed for each virtual screen.

RESULTS AND DISCUSSION

To validate EDPrints (consisting of individual ^{1}H and ^{13}C NMR shifts and atomic partial-charge fingerprints (Figure 5)), we analyzed the retrospective VS results of all 40 targets of the DUD data set[13] and compared them to the VS accuracies of two other VS methods: MOLPRINT 2D and FCFP-4 (Figures 6−12 and Table 2 and Supporting Information, Tables 1−3). In addition, we compared the average computation time of each screening method to perform a similarity calculation between two compounds.

EDprints Combines Electron Density Descriptors into a Robust VS Tool. EDprints similarity scores are composed of the similarity scores of the ^{13}C and ^{1}H NMR shift and atomic partial charge fingerprints with 0.25, 0.25, and 0.50 weighting factors, respectively (as determined by a 10-fold cross-validation, see Methods Section). Combination of the three different electron density descriptors results in a more robust EDprints model, as the performance of individual fingerprints varies per target. This is exemplified for the acetylcholinesterase (ACHE), adenosine deaminase (ADA), and dihydrofolate reductase (DHFR) targets (Figure 5). ^{13}C NMR shift fingerprints are significantly better descriptors

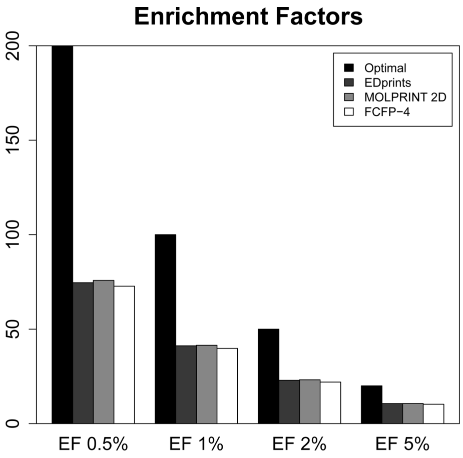


Figure 6. Average enrichment factors among all 40 DUD data sets is given in for the first 0.5, 1, 2, and 5% FP rates.

than ^{1}H NMR shift and partial charge fingerprints for VS of ACHE ligands (Figure 5), while for the retrieval of ADA ligands, ^{1}H NMR shift fingerprints are better descriptors than ^{13}C NMR shift and partial charge fingerprints (Figure 5b). For VS of DHFR ligands on the other hand, partial charge fingerprints clearly outperform NMR shift fingerprint models (Figure 5c).

Similar Overall VS Accuracy of Molecular Similarity Tools. Figure 6 shows a comparison of the average enrichment factors among all 40 DUD data sets at 0.5, 1, 2, and 5% FPs, while the AUC values and enrichment factors at 1% FP rate for individual targets is presented in Table 2. AUC values (including 95% confidence interval) and enrichment factors at 0.5, 1, 2, and 5% for all 40 DUD ligand data sets are presented in Supporting Information, Tables 1−3. Figure 6 shows that EDprints has a similar overall VS performance in terms of early enrichment compared to MOLPRINT 2D and FCFP-4. This demonstrates that electron density information is a suitable descriptor for ligand-based VS.

Table 2 presents the AUC values and early enrichments for all 40 DUD data sets and demonstrates that the differences in VS accuracy between the methods are also small when comparing the results for individual targets. Like most other ligand-based VS approaches, EDprints has a good global VS accuracy (AUC > 0.80) and early enrichment (EF1% > 30) for most targets: ACE, ADA, AmpC, COX-2, DHFR, EGFr, ER-agonist, ER-antagonist, FXA, GART, GPB, HMGA, HSP90, NA, P38, PARP, PNP, PPARg, PR, RXRa, SAHH, thrombin, and trypsin.

For MR and TK the global VS accuracy of EDprints is good (AUC > 0.80), while early enrichment is mediocre (EF1% = 21−22). For few targets, global VS accuracy is mediocre (AUC < 0.80), while early enrichment is high (EF 1% > 30): ACHE, COMT, FGFR1, GR, HIVPR, INHA, and SRC. As expected, EDprints (as well as the other ligand-based 2D similarity methods) performs poorly (AUROC < 0.7 and EF1% < 30) for more challenging ligand sets with high chemical scaffold diversity: ALR2, AR, CDK2, COX-1, HIVRT, PDE5, PDGFRB, and VEGFr2 (Table 2). While the average VS performance of all ligand-based methods is comparable for the different targets, it should be noted that ligand-based VS results for the DUD data set can be highly reference ligand-dependent (as reported by Ebalunode et al.).[17] The HIVPR ligand set, for example, contains three distinct scaffold classes which are not equally represented: oxo-pyran/oxan sulfonamides (40 entries, including compound (cpd) **1**), carboxamides (4 entries, including cpd **2**), and diazepan-2-ones (13 entries, including cpd **3**). Logically, ligand-based VS runs against an oxo-pyran/oxan sulfonamide reference (e.g., cpd **1**) enable retrieval of a significantly higher number of true HIVPR actives than screening against a representative of one of the other two ligand types (Figure 7).

Target-specific variations in VS accuracies of the evaluated methods are generally in line with previous retrospective 2D ligand-based VS exercises on the DUD data set,[15−17] Like MOLPRINT 2D and FCFP-4, as well as several other VS approaches previously validated with the DUD data set,[15−17] EDprints performs well for targets containing conserved scaffolds in the ligand data set but performs poorly for more challenging targets with high ligand scaffold diversity (Table 2). For some targets however, EDprints performed significantly better (ACE, AmpC, COX-2, Fxa, PDE5, and trypsin)

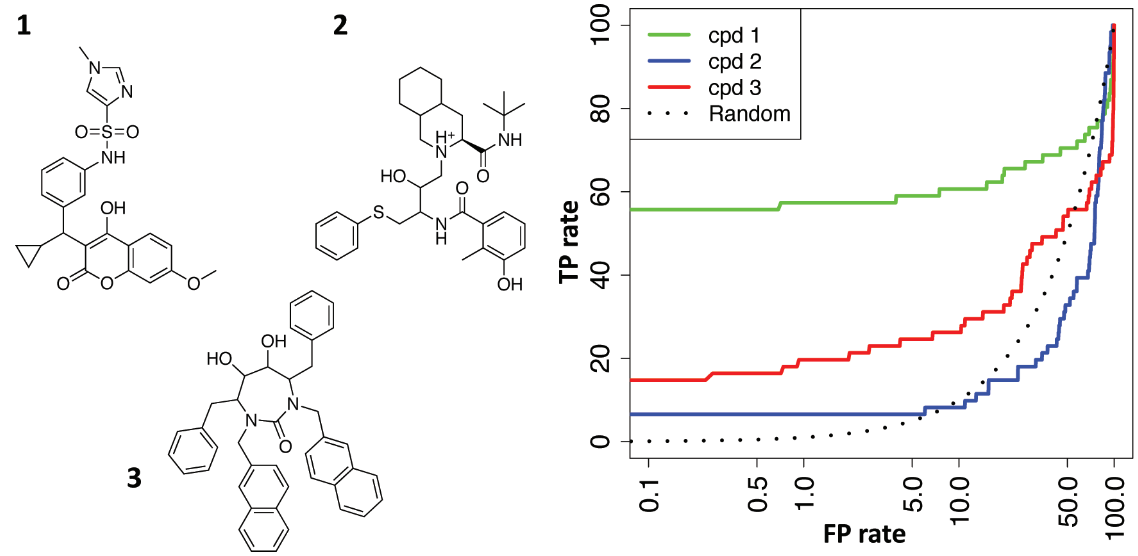


Figure 7. VS enrichment plots of known HIVPR ligands (FP) and corresponding decoys (TP) from the DUD data set[13] using EDprints based on HIVPR reference compounds **1**, **2**, and **3** (green, blue, and red lines, respectively).

Table 2. Area Under the ROC Curve (AUROC) and EF at 1% False-Positive Rate for EDprints, MOLPRINT 2D, and FCFP-4 Observed on all 40 DUD Target-Specific Data Sets

	EDprints		MOLPRINT 2D		FCFP-4	
data set	AUROC	EF 1%	AUROC	EF 1%	AUROC	EF 1%
ACE	0.83	31	0.85	26	0.81	25
ACHE	0.70	41	0.72	44	0.72	35
ADA	0.84	40	0.88	34	0.88	41
ALR2	0.61	14	0.66	16	0.55	12
AmpC	0.81	50	0.80	47	0.86	42
AR	0.75	26	0.63	26	0.67	28
CDK2	0.56	13	0.51	13	0.54	12
COMT	0.78	35	0.75	35	0.75	32
COX-1	0.54	8	0.60	10	0.62	7
COX-2	0.87	46	0.82	43	0.77	34
DHFR	0.99	68	0.98	72	0.92	53
EGFr	0.94	56	0.93	64	0.89	61
ER agonist	0.80	32	0.85	38	0.81	35
ER antagonist	0.83	34	0.86	39	0.83	34
FGFr1	0.77	45	0.74	46	0.66	40
Fxa	0.88	58	0.85	55	0.86	50
GART	0.96	45	0.93	39	0.96	50
GPB	0.89	45	0.89	32	0.90	45
GR	0.69	42	0.63	46	0.69	46
HIVPR	0.69	40	0.78	38	0.68	42
HIVRT	0.48	8	0.50	9	0.51	8
HMGA	0.86	73	0.90	78	0.84	71
HSP90	0.87	53	0.87	56	0.78	54
InhA	0.64	36	0.69	36	0.63	36
MR	0.81	21	0.80	22	0.79	30
NA	0.84	45	0.90	46	0.79	48
P38	0.81	51	0.84	61	0.82	49
PARP	0.82	69	0.93	69	0.92	71
PDE5	0.65	20	0.56	17	0.55	14
PDGFrb	0.62	22	0.61	22	0.53	22
PNP	0.92	68	0.87	65	0.88	68
PPARg	0.95	77	0.95	79	0.95	76
PR	0.86	56	0.69	53	0.66	56
RXRa	0.98	72	0.99	75	0.98	73
SAHH	0.95	36	0.95	39	0.92	39
SRC	0.73	37	0.69	36	0.60	31
thrombin	0.84	45	0.72	43	0.70	44
TK	0.91	22	0.90	23	0.84	17
trypsin	0.86	56	0.75	54	0.73	53
VEGFr2	0.60	11	0.54	10	0.54	9

or worse (EGFr, ER-agonist, GR, and SAHH) than other previously tested 2D ligand-based methods[16] in terms of early enrichment.

Earlier retrospective VS studies on the DUD data set also show that 3D ligand-[15−17] or docking-based[13,16] methods are in few cases more suitable for identifying: (i) ligands of varying chemical topology (ALR2 (imidazolidine-2,4-dione vs carboxylate-containing ligands) and TK (purine vs pyrimidine-containing ligands)), (ii) varying size (ACE and COMT), or (iii) ligands containing few heteroatoms (MR). The overall VS accuracy of 2D-ligand based approaches is comparable to that of 3D-ligand based methods[15−17] and superior to docking-based approaches.[13] A possible approach to overcome scoring problems[38] with docking-based VS is the use of protein−ligand interaction fingerprint scoring functions to rank docking poses.[39] Finally, it should be noticed that the complementary use of different ligand- and structure-based VS methods increases the chance of finding more (and more diverse) active compounds.[40]

Dissimilarities between Molecular Similarity Methods in Specific VS Cases. While MOLPRINT 2D and FCFP-4 explicitly encode molecular topology, EDprints uses numerical values reflecting the electron density distribution within molecules. Although on average there are no significant differences in the (target and reference ligand-dependent) VS accuracies of EDprints, MOLPRINT 2D, and FCFP-4 (Figure 6 and Table 2), specific examples highlight interesting differences between the three methods (Figures 7−12). First of all, EDprints is more suitable for screening against reference molecules with fused ring systems than atom-centered circular fingerprint methods MOLPRINT 2D and FCFP-4 (Figures 8 and 9). For example, pyrazolo/triazolopyrimidinone containing ligands of PDE5 (like cpd **5**) can be identified with the fused ring system containing cpd **4** (Figure 8), while the fused phenylpyrrolopyridine reference cpd **6** is a suitable screening reference for phenylimidazolpyridine ligands of P38 (like cpd **7**, Figure 9), and the fused phenylthiadiazatricyclotetraene scaffold of COX-2 cpd **8** can be used to efficiently retrieve diphenyl-containing ligands (like cpd **9**, Figure 10). Finally, EDprints models are more robust toward subtle rearrangements of functional groups. The thiophene sulfonamide ligand of ACMP is a more efficient reference for VS of thiophene sulfomayl ligands with EDprints than with MOLPRINT 2D or FCFP-4 (Figure 11).

It should be noted, however, that EDprints similarity searches can be biased by differences in the molecular protonation and tautomeric states (yielding different partial charges and NMR shifts and therefore different fingerprints) of reference and target ligands. While EDprints-based VS accuracy is not significantly affected by small differences in the number of ionizable groups in reference and target

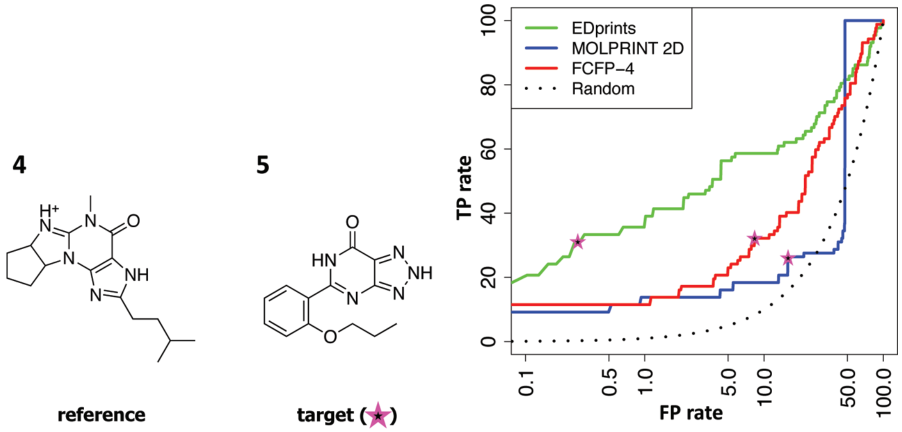


Figure 8. VS enrichment plots of known PDE5 ligands (FP) and corresponding decoys (TP) from the DUD data set[13] using EDprints (green), MOLPRINT 2D[9,10] (blue), and FCFP-4[11] (red) based on PDE5 reference compound **4**. Rankings of specific search target compound **5** in the enrichment curves of the three different screening methods are indicated with an asterisk.

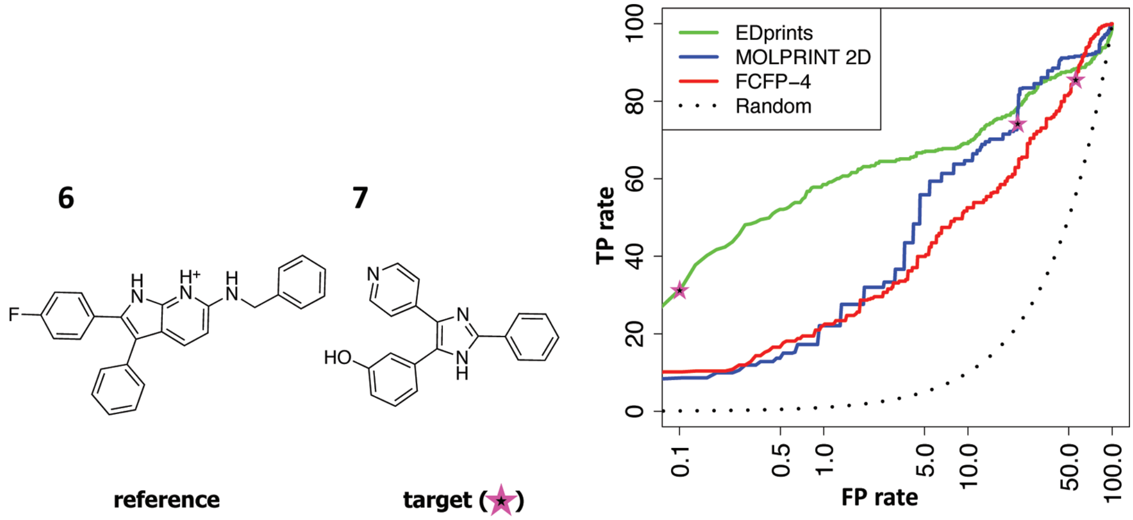


Figure 9. VS enrichment plots of known p38 ligands (FP) and corresponding decoys (TP) from the DUD data set[13] using EDprints (green), MOLPRINT 2D[9,10] (blue), and FCFP-4[11] (red) based on p38 reference compound **6**. Rankings of specific search target compound **7** in the enrichment curves of the three different screening methods are indicated with an asterisk.

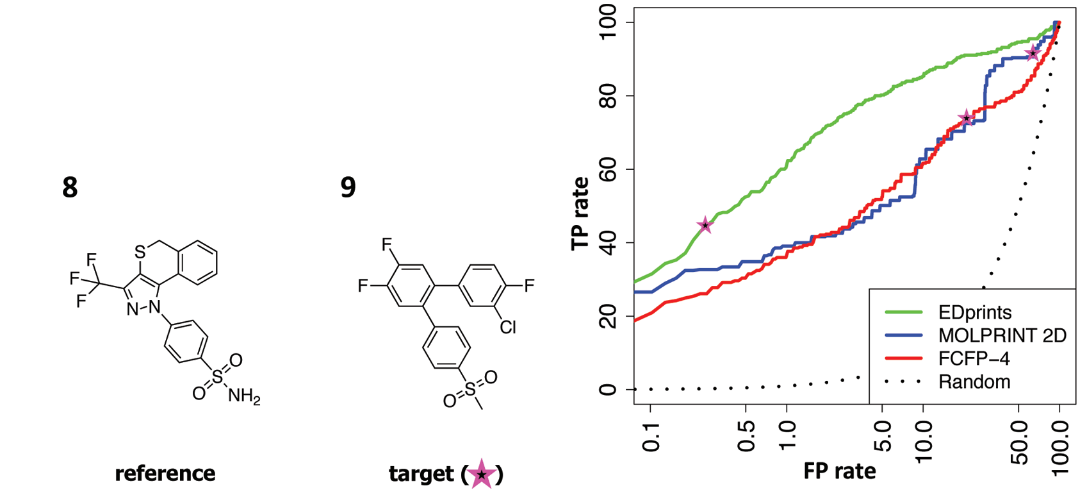


Figure 10. VS enrichment plots of known COX-2 ligands (FP) and corresponding decoys (TP) from the DUD data set[13] using EDprints (green), MOLPRINT 2D[9,10] (blue), and FCFP-4[11] (red) based on COX-2 reference compound **8**. Rankings of specific search target compound **9** in the enrichment curves of the three different screening methods are indicated with an asterisk.

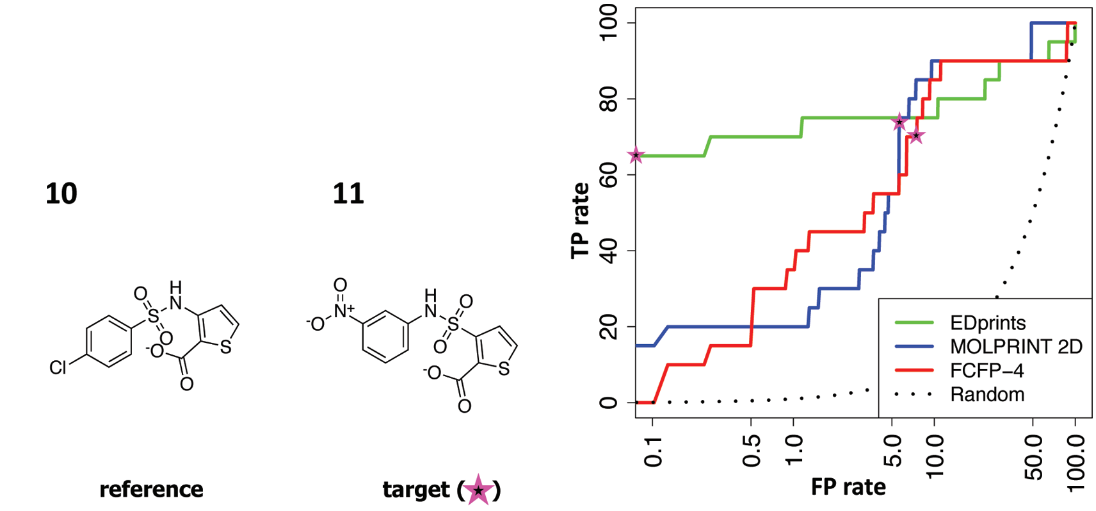


Figure 11. VS enrichment plots of known AmpC ligands (FP) and corresponding decoys (TP) from the DUD data set[13] using EDprints (green), MOLPRINT 2D[9,10] (blue), and FCFP-4[11] (red) based on AmpC reference compound **10**. Rankings of specific search target compound **11** in the enrichment curves of the three different screening methods are indicated with an asterisk.

ligands (e.g., P38 ligands **6** and **7** in Figure 9), compounds containing multiple ionizable groups are less suitable reference ligands for EDprints screening (compare p38 reference ligands **12** and **13** to retrieve target ligand **14** in Figure 12).

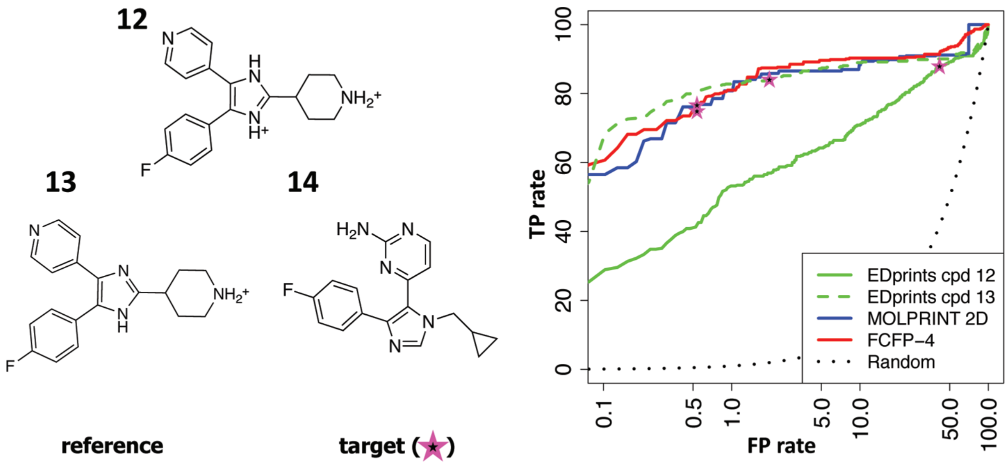


Figure 12. VS enrichment plots of known p38 ligands (FP) and corresponding decoys (TP) from the DUD data set[13] using EDprints (green and purple), MOLPRINT 2D[9,10] (blue), and FCFP-4[11] (red) based on P38 reference cpds **12** and **13**. Rankings of specific search target cpd **14** in the enrichment curves of the three different screening methods are indicated with an asterisk.

The performance of MOLPRINT 2D and FCFP-4 is generally not that much affected by differences in protonation state (data not shown). The MMFF94 partial charge model is considered as a good alternative to more computationally expensive semiempirical charge models[28,29] but is independent of compound conformation. Therefore, it will be interesting to investigate the possibility of including ligand conformation-dependent atomic partial charge[41] and chemical shift models[42] in EDprints in the future.

Faster VS with EDprints. While the retrospective VS accuracy of EDprints is similar to the accuracy of the established MOLPRINT 2D and FCFP-4 methods, EDprints offers a much faster computation time for comparing fingerprints. The benchmarks for the comparison of two compound-representing descriptors were performed on a single core of an Intel Core 2 Duo E6420 processor with 2.13 GHz, and the comparison time for two EDprints was determined to be 4.1 μs. This is about 233 times faster than MOLPRINT 2D (954.4 μs per comparison) and 56 times faster than FCFP-4 (231.6 μs per comparison). The high screening rate of EDprints can mainly be ascribed to the use of bitwise operations on the binary fingerprints. Each fingerprint is stored as multiple 64-bit substrings, which current cpus can process as a whole. On the other hand, the creation rate of the fingerprint currently implemented in EDprints creation rate is relatively low for EDprints (118 953 μs per molecule) compared to MOLPRINT 2D (346 μs per molecule) and FCFP-4 (214 μs per molecule). The overhead through the use of the third party programs, instead of direct integration, together with the calculation time of the chemical shifts and the partial charges accounts for 96% of the creation time. This process, however, has to be performed only once for a molecule, while the screening rate will influence the computing time for each screening. Consequently, EDprints is highly suitable for the repetitive VS of very large chemical databases[3] and the construction all-against-all similarity matrices.[4]

CONCLUSIONS

While most molecular fingerprints explicitly encode molecular topology, we have developed a new method, EDprints, which uses numerical values reflecting the molecular electron density distribution for molecular similarity searches. Our study shows that EDprints is able to achieve similar screening accuracies compared to other two-dimensional (2D) ligand-based similarity methods. Specific examples highlight interesting differences between the new EDprints fingerprint method and the atom-centered circular fingerprint methods of MOLPRINT 2D and FCFP-4. On one hand, EDprints similarity searches can be biased by the molecular protonation state, especially the presence of multiple ionizable groups. On the other hand, EDprints models are more robust toward subtle rearrangements of chemical groups and are more suitable for screening against reference molecules with fused ring systems, rather than MOLPRINT 2D and FCFP-4. EDprints is the fastest method under investigation for comparing fingerprints with an average of 4.1 μs per molecule comparison. This makes EDprints highly suitable for all-against-all similarity searches and for repetitive virtual screening against large chemical databases of millions of compounds.

ACKNOWLEDGMENT

We would like to thank Porta Nova Software GmbH for the use of their BatchNMRPrediction software, which was used to calculated all ^{13}C and ^{1}H NMR shifts.

Supporting Information Available: Distributions of ^{13}C and ^{1}H shifts and partial charges of all molecules in the DUD data set, virtual screening results of EDprints, FCFP-4, and MOLPRINT for individual targets in the DUD data set, and the workflow to optimize EDprints parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Rester, U. From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 559–568.
(2) Irwin, J.; Shoichet, B. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(3) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.
(4) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.

(5) 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.

(6) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2009**, *53*, 539–558.

(7) Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.

(8) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49* (1), 108–119.

(9) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments and information-based feature selection and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.

(10) 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. Comput. Sci.* **2004**, *44*, 1708–1718.

(11) *Pipeline Pilot*, version 6.1.5; Accelrys: San Diego, CA, 2010.

(12) *Molecular Operating Environment (MOE)*, version 2009.10; Chemical Computing Group: Montreal, Canada, 2010.

(13) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.

(14) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular mocking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49* (6), 1455–1474.

(15) Kirchmair, J.; Distinto, S.; Markt, P.; Schustera, D.; Spitzer, G.; Liedl, K.; Wolber, G. How to optimize shape-based virtual screening: choosing the right query and including chemical information. *J. Chem. Inf. Model.* **2009**, *49* (3), 678–692.

(16) 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* (2), 209–231.

(17) Ebalunode, J.; Zheng, W. Unconventional 2D shape similarity method affords comparable enrichment as a 3D shape method in virtual screening experiments. *J. Chem. Inf. Model.* **2009**, *49* (6), 1313–1320.

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

(19) 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. Chem. Inf. Model.* **2006**, *49* (23), 6802–6810.

(20) Yeap, S. K.; Walley, R. J.; Snarey, M.; van Hoorn, W. P.; Mason, J. S. Designing Compound Subsets: Comparison of Random and Rational Approaches Using Statistical Simulation. *J. Chem. Inf. Model.* **2007**, *47*, 2149–2158.

(21) Kogej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multifingerprint Based Similarity Searches for Targeted Class Compound Selection. *J. Chem. Inf. Model.* **2006**, *46*, 1201–1213.

(22) Bender, A.; Mussa, H. Y.; Glen, R. C. Screening for Dihydrofolate Reductase Inhibitors Using MOLPRINT 2D, a Fast Fragment-Based Method Employing the Naive Bayesian Classifier: Limitations of the Descriptor and the Importance of Balanced Chemistry in Training and Test Sets. *J. Biomol. Screen.* **2005**, *10*, 658–666.

(23) *BatchNMRPredictor*, version 1.1a; Porta Nova Software GmbH: Zürich, Switzerland, 2010.

(24) Pretsch, E.; Furst, A.; Badertscher, M.; Buergin, R.; Munk, M. E. C13Shift: a computer program for the prediction of carbon-13 NMR spectra based on an open set of additivity rules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 291–295.

(25) Schaller, R. B.; Munk, M. E.; Pretsch, E. Spectra Estimation for Computer-Aided Structure Determination. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 239–243.

(26) Vainio, M.; Johnson, M. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.* **2007**, *47* (6), 2462–2474.

(27) Halgren, T. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

(28) Mittal, R. R.; Harris, L.; McKinnon, R. A.; Sorich, M. J. Partial Charge Calculation Method Affects CoMFA QSAR Prediction Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 704–709.

(29) Hou, T.; Zhu, L.; Chen, L.; Xu, X. Mapping the Binding Site of a Large Set of Quinazoline Type EGF-R Inhibitors Using Molecular Field Analyses and Molecular Docking Studies. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 273–287.

(30) Naylor, E.; Arredouani, A.; Vasudevan, S. R.; Lewis, A. M.; Parkesh, R.; Mizote, A.; Rosen, D.; Thomas, J. M.; Izumi, M.; Ganesan, A.; Galione, A.; Churchill, G. C. Identification of a chemical probe for NAADP by virtual screening. *Nat. Chem. Biol.* **2009**, *5*, 220–226.

(31) Gasteiger, J.; Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, *19*, 3181–3184.

(32) Willet, P.; Barnard, J. M.; Downs, G. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996.

(33) Wartik, S.; Fox, E.; Chen, Q.; Heath, L. S. Hashing Algorithms 293−362. *Information retrieval: data structures and algorithms*; Prentice-Hall: Englewood Cliffs, NJ, 1992.

(34) Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **1972**, *28*, 11–20.

(35) McConnaughey, B. H. The determination and analysis of plankton communities. *Penelitian laut di Indonesia (Marine research in Indonesia)* **1964**, 1–40, special number.

(36) Team, R. D. C. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(37) Jain, A.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 133–139.

(38) Ferrara, P.; Gohlke, H.; Price, D.; Klebe, G.; Brooks, C. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047.

(39) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(40) Krüger, D.; Evers, A. Comparison of Structure- and Ligand-Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. *ChemMedChem* **2010**, *5*, 148–158.

(41) Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7821–7839.

(42) Bulo, R. E.; Jacob, C. R.; Visscher, L. NMR Solvent Shifts of Acetonitrile from Frozen Density Embedding Calculations. *J. Phys. Chem. A* **2008**, *112*, 2640–2647.

CI1002608