JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Ligand-Based Target Prediction with Signature Fingerprints

Jonathan Alvarsson,*[,†] Martin Eklund,[†,‡] Ola Engkvist,[¶] Ola Spjuth,[†,§] Lars Carlsson,[∥] Jarl E. S. Wikberg,[†] and Tobias Noeske*[,∥]

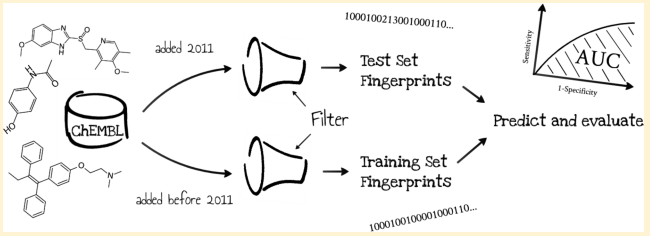
[†]Department of Pharmaceutical Biosciences and [§]Science for Life Laboratory, Uppsala University, SE-751 05 Uppsala, Sweden

[‡]Department of Surgery, University of California at San Francisco (UCSF), San Francisco, California 94115, United States

[¶]Discovery Sciences, Chemistry Innovation Centre, and [∥]Computational ADME and Safety, DSM, AstraZeneca R&D, SE-431 83 Mölndal, Sweden

Supporting Information

ABSTRACT: When evaluating a potential drug candidate it is desirable to predict target interactions in silico prior to synthesis in order to assess, e.g., secondary pharmacology. This can be done by looking at known target binding profiles of similar compounds using chemical similarity searching. The purpose of this study was to construct and evaluate the performance of chemical fingerprints based on the molecular signature descriptor for performing target binding predictions. For the comparison we used the area under the receiver operating characteristics curve (AUC) complemented with net reclassification improvement (NRI). We created two open source signature fingerprints, a bit and a count version, and evaluated their performance compared to a set of established fingerprints with regards to predictions of binding targets using Tanimoto-based similarity searching on publicly available data sets extracted from ChEMBL. The results showed that the count version of the signature fingerprint performed on par with well-established fingerprints such as ECFP. The count version outperformed the bit version slightly; however, the count version is more complex and takes more computing time and memory to run so its usage should probably be evaluated on a case-by-case basis. The NRI based tests complemented the AUC based ones and showed signs of higher power.



INTRODUCTION

Safety issues are one of the major reasons for attrition in drug discovery and development, accounting for approximately 20% of all discontinuations in late-stage development.[1] A compound can exert its toxic effect via excess pharmacology at the primary target, a chemically reactive functionality, or activity at an unintended target, referred to as *secondary* or *off-target pharmacology*. Off-target pharmacology has gained an increased interest within the past decade with the realization that compounds usually interact not only with one, but more often with multiple targets[2,3] and hence prediction of target binding profiles is an important task.

Prediction of targets that a given chemical compound bind to can be made by querying existing databases with the aim of finding structurally similar compounds and examine targets to which they are known to bind,[4−6] assuming the so-called similarity principle, i.e. that similar compounds will behave similarly.[7,8]

Similarity searching with two-dimensional (2D) fingerprints is, due to its speed, probably the most common method for querying databases for structures similar to a given query molecule.[9] Molecular 2D fingerprints are predominantly bit vector representations of molecular fragments, where the bit represents whether a property is found in the molecule or not.

The aim of the present study was to generate fingerprints based on the signature descriptors[10−12] and to evaluate their performance in target binding predictions. Introduced about a decade ago, signature descriptors are topological descriptors canonically describing the connectivity of each atom in a molecule with its neighboring atoms in a tree-like fashion. Structurally similar compounds can be found by comparing the molecular fingerprints using a distance measure, such as the Tanimoto coefficient[13] which is among the most commonly used distance measures.[4] In doing so, a ranking list of molecules ordered by similarity is generated where the highest ranking molecule is predicted as the most likely candidate to bind.

One common approach for evaluating ranking lists is the enrichment factor (EF).[14] This method looks at the top few percent of the list to see whether there are more substances which bind to the studied target (hereby denoted as positive substances) in those top percent than would be expected from a random outcome. This gives a number corresponding to the amount of enrichment that has occurred in those top few percent. The use of EF as a basis for performance comparison between studies suffers from at least these problems:[14]

- It depends on the number of positive and negative substances.

Received: June 19, 2014
Published: September 17, 2014

- It has a free parameter (what level of percentage to look at) which can vary between studies and complicate comparison.

Another approach is to study the receiver operating characteristic (ROC) curve and the area under that curve (AUC). In a ROC curve, the rate of true positives (sensitivity) is plotted on the vertical axis and the rate of false positives (1 − specificity) on the horizontal. AUC can be interpreted as the probability of being right when trying to differentiate between a positive and a negative outcome.[15] The AUC value thus has the beneficial property of being easy to interpret.

However, the statistical power of AUC is not always satisfactory when testing for the difference in predictive performance between two different models.[16] The net reclassification improvement (NRI) statistic was therefore introduced as a complement to classic comparison of AUC values.[17] NRI is especially suitable when a specific cutoff is used to separate positive and negative predictions but can also be used without such a cutoff.

In this paper we introduce two open-source fingerprints based on molecular signature descriptors and compare them with existing 2D fingerprints using a publicly available data set with over 200 000 entries covering 161 targets, where we predict the known interactions and evaluate the performance by AUC and NRI using temporal validation.

METHODS

Data. ChEMBL[18] is an open database that has binding data for a large number of drug-like compounds. We downloaded ChEMBL 15 and filtered out substances screened in a binding assay (i.e., with K_i, IC_{50}, or K_d obtained) against a human protein and with between 5 and 50 non-hydrogen atoms for the compound. Activities of up to 10 000 nM were considered to signify binding for IC_{50} and EC_{50}, whereas activities of up to 5000 nM were used for K_i and K_d (thus using the conversion factor of 2 suggested by Kalliokoski et al.[19]). Where multiple measurements were reported, the ChEMBL specific confidence score was used to select only the interaction measure judged as the most trustworthy.

Study Design. Training data were used for training the models and independent test data were used for evaluating the performance of the models fitted to the training data. The separation between training and test data was done using temporal validation in order to evaluate how well the behavior of future substances could be predicted based on what was known about contemporary substances. Since each target was evaluated separately, the data were split into multiple training and test sets, one for each target. The training sets contained chemical structures, with associated interaction measurements, added before 2011, and the test sets contained structures from 2011. This time-based split simulated the situation where the model would be used: a model would be trained on all available data before performing wet-lab experiments, allowing for prioritization and early warnings. The model would be run for a while and then be retrained on all, at that time, available data.

Both the training and the test sets were filtered (Figure 1) so that they contained at least 50 substances for each target (to have enough data points as basis for the evaluation). The training set contained 284 186 observations (measured interaction values) for 99 790 unique substances against 161 targets. There was an overlap in the sense that multiple

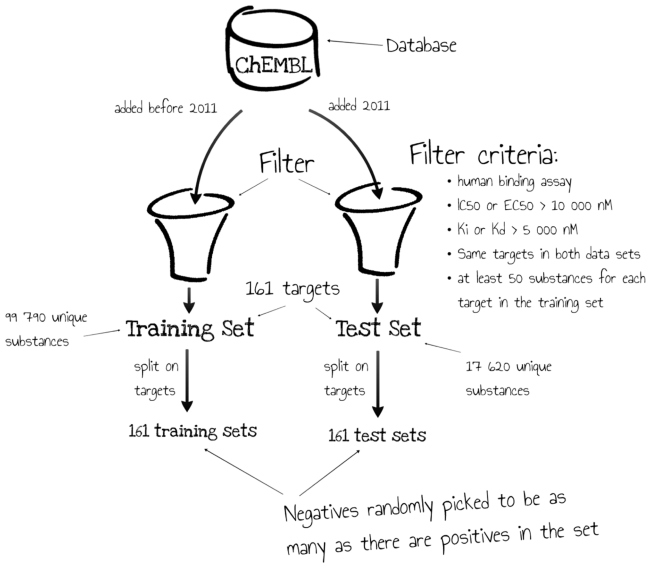


Figure 1. ChEMBL 15 database was downloaded and split into training and test data set based on time of addition to the database. The data sets were filtered so that there were at least 50 substances for each target and the same targets were used in both sets. Only binding assays were considered. The sets were split up according to which target they were known to bind to, and an equal number of negatives which had been added to ChEMBL during the same time span were added randomly to make up training and test sets.

substances in the training set were listed as binding to multiple targets (Figure 2). The number of substances known to bind to the target in each training set varied between 57 and 4 279. All in all, the test sets contained 17 620 different substances.

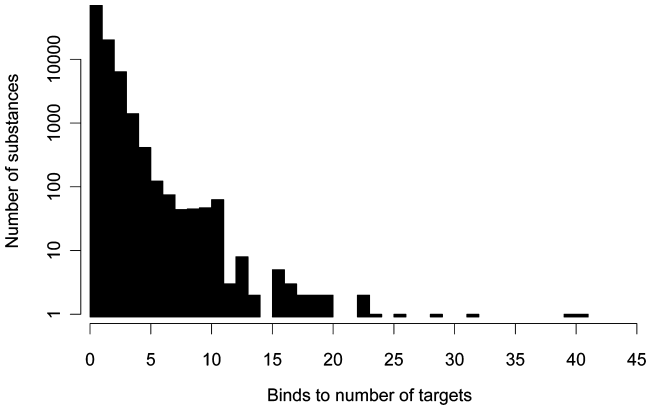


Figure 2. Number of substances binding to number of targets. Bar plot showing the number of substances in the training data that bind to how many targets among the targets. The two substances binding to the most targets bind to 40 and 41 out of the 161 targets.

The test sets for each target were made up of all substances in the whole test set known to bind to the target and an equal number of substances randomly chosen from the entries from the same time span, not listed as binding. Since the main cost for including data is data set size (affecting execution time and memory usage), we used the proportion 1:1 between positives and negatives.

In the medical field this study design is known as a *case control study*. In a case control study there is one group having a property (in our case: binding to the target; i.e. the positive

dx.doi.org/10.1021/ci500361u | J. Chem. Inf. Model. 2014, 54, 2647−2653

substances) and another group, known as control, which does not have that property; i.e. the negative substances. Another term from the medical field is *prevalence*, which refers to the proportion of a population having a certain condition, such as for example a disease. The prevalence needs to be known in order to estimate statistical properties such as positive predictive value (PPV or recall) and negative predictive value (NPV). In our case, knowing the prevalence corresponds to knowing the proportion of tested substances being positive. Our data come from a public repository containing only positive results, so statistics about the repository cannot be used to estimate the prevalence of binding substances because the amount of substances that do not bind to anything is unknown and not counted.

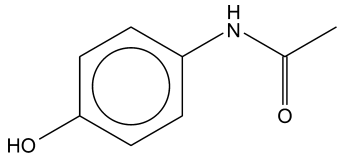
The assumption that substances that were not listed as binding to the examined target were negative could potentially be wrong for some substances, but the number of unreported binding substances can be expected to be low and thus the effect should not be important. Also, as long as they are randomly included they will not affect the predicted ranking. Likely a training set of confirmed negatives would produce lower AUC values because randomly selected substances from ChEMBL will differ more than substances tested and confirmed as negative,[20] however this effect should not change the individual order between the fingerprints. Although we are not able to estimate PPV and NPV (and therefore not any statistics based on PPV or the NPV, e.g., the F1-score), we can estimate AUC and NRI. As long as our positives are selected at random from the entire set of possible positives, a predictor will produce probabilities that only differ by a constant factor.[21] Under the assumption of random selection, it is possible to estimate sensitivity and specificity, which is all we need in order to compute the AUC.

Chemical Fingerprints. This study covered six different fingerprints (Table 1) based on three different principles.

Table 1. Six Different Fingerprints with Two Different Ways of Generalizing the Tanimoto Coefficient Count Fingerprint

designation	short description	source
BitSignFPHeight0to5	bit signature fingerprint	made publicly available in CDK
CountTc1	count signature fingerprint with Tanimoto generalization 1	made publicly available in CDK
CountTc2	count signature fingerprint with Tanimoto generalization 2	made publicly available in CDK
CDKExtended	hashed path fingerprint	publicly available in CDK
FOYFI	hashed path fingerprint	AstraZeneca in-house
ECFP	extended connectivity fingerprint (ECFP_6)	PipeLine Pilot
ECFI	extended connectivity fingerprint	AstraZeneca in-house

Signature Fingerprints. In constructing our new fingerprints we used the signature molecular descriptor which has the attractive property that all topological indices based on counts of walks, paths, and distances can be derived from it.[10] It has been successfully used in computer-aided molecular design[22] and QSAR modeling.[23−25] A molecular signature consists of an array of atom signatures. An atom signature is a canonical string representation of the atom's environment (Figure 3). A variable



Height	Signature for the nitrogen atom
0	[N]
1	[N]([C][C])
2	[N]([C](p[C]p[C])[C]([C]=[O]))

Figure 3. Example of atom signatures, for the nitrogen in acetaminophen (paracetamol). Signatures of height 0 to 2 for the nitrogen atom in acetaminophen (paracetamol) molecule depicted above. At height 0 the signature only specifies the atom type and as the height increases more information about the atom environment is included.

determines the height of the signature, which corresponds to the distance (number of atoms) from the atom that is described. We used heights 0 to 5 and created a bit fingerprint based on the signatures.

We also created a count version of the fingerprint (sometimes referred to as a holographic fingerprint) where each element was not a bit but an integer corresponding to how many times the signature was present in the given molecular structure.

In creating a fingerprint from the signature descriptor already implemented in The Chemistry Development Kit (CDK),[26,27] which is an open-source Java library for cheminformatics, we used the default hashcode method in the programming language Java to hash each atom signature string into an integer. Integers in Java use 32 bits, which has been suggested to be enough to minimize collisions during hashing of molecular fingerprints.[28] A naive dense representation would store not only the set fingerprint bits but also the unset bits amounting to a total of $2^{32} \approx 4$ billion bits or half a gigabyte of memory. The number of set bits were significantly smaller and thus sparse representation of the data was important in order to be able to work with a great number of fingerprints in memory.

The bit fingerprint consisted of an ordered list of these integers with duplicates removed. This made up a bit fingerprint where the ordered list represented the set bits and all other bits were unset. The count fingerprint was made up of two lists, one containing the indices and the other the corresponding counts.

Because the signature fingerprint consists of an array of atom signatures it is somewhat similar to the circular fingerprints, i.e., ECFP and ECFI (see below) but the atom signature is canonical. Also the hashing to a 32 bit integer without folding down to an indexing space of 10 bits is similar to the tested ECFP.

Hashed Path Fingerprints. Starting from an atom in a molecule it is possible to trace different paths through the molecule. By storing the visited atoms in a list and hashing this down to an integer, a fingerprint can be constructed. CDK supplies fingerprints of this type. We tested the fingerprint from CDK named *Extended*, which is a fingerprint based on all paths in the molecule truncated at length 8 and extended by a few ring features. It is hashed down to 1024 bits, (i.e., the indexing space is made up of 10 bits, which can be compared to the 32 bits which made up the indexing space in our signature

dx.doi.org/10.1021/ci500361u | J. Chem. Inf. Model. 2014, 54, 2647−2653

fingerprint). A similar fingerprint which was tested is FOYFI, which is an AstraZeneca in-house fingerprint.[29]

Extended-Connectivity Fingerprints. Extended-connectivity fingerprint (ECFP)[30] is a class of topological fingerprints, which is developed specifically for structure−activity modeling. The fingerprints in this class are circular fingerprints. ECFP has shown good performance at predicting activities for ligands[31] and is arguably at the top performance level among 2D fingerprints.[32] We used ECFP with a height of 6 (ECFP_6) generated by *Pipeline Pilot*.[33]

We also tested the *ECFI* fingerprint, which is an AstraZeneca in-house fingerprint based on the extended-connectivity method which has seen previous use.[34]

Prediction by Similarity Searching. A Tanimoto-based similarity search was implemented using a *k*-nearest neighbor (kNN) approach to search for similar chemical structures and on that basis predict target interactions.

For each test substance the *k* nearest neighboring substances (with regards to Tanimoto coefficient) were used to create an estimation of the probability that the test substance binds to the examined target by a voting procedure according to

$$p_{\text{binding}} = \frac{m}{k} \tag{1}$$

where m was the number of neighbors that was known to bind to the given target and k was the total number of studied neighbors. In case of ties, k was extended until no more ties were found. The k variable was determined through a 10-fold cross validation varying k from 1 to the number of training substances and choosing the median of the 10 individual k that maximized the Mann−Whitney U, calculated with the org.apache.commons.math[35] Java library (maximizing the Mann−Whitney U is equivalent to maximizing the AUC).

Generalization of the Tanimoto coefficient to count fingerprints can be done in many ways and we have tried two previously tested approaches.[36] The first method (hereby referred to as CountTc1), for fingerprints x and y was defined as

$$\text{CountTc1}(x, y) = \frac{\sum_i^N x_i y_i}{\sum_i^N (x_i^2 + y_i^2 - x_i y_i)} \tag{2}$$

and the second method (hereby referred to as CountTc2) was defined as

$$\text{CountTc2}(x, y) = \frac{\sum_i^N \min(x_i, y_i)}{\sum_i^N \max(x_i, y_i)} \tag{3}$$

where x_i is the count in bin i for the first fingerprint and y_i is the count for bin i in the second fingerprint.

Assessment Metrics. We used the AUC measurement as our main metric for evaluating the performance of the fingerprints. With AUC we could make method comparisons in a straightforward, well-defined and quantifiable way. For all predicted binding probabilities we generated ROC curves in R[37] using the pROC[38] package.

The AUC values were complemented by NRI values which were generated using the R package Hmisc.[39] NRI compares the ordering of two ranking lists scoring positives and negatives on whether they have moved up or down in the ranking when comparing the first list with the second list, according to

$$\text{NRI} = \frac{\sum_i^p v(i)}{p} - \frac{\sum_j^n v(j)}{n} \tag{4}$$

where $v(\cdot)$ is a movement indicator defined as

$$v(\cdot) = \begin{cases} 1, & \text{for upward movement} \\ 0, & \text{for no movement} \\ -1, & \text{for downward movement} \end{cases} \tag{5}$$

Let N and P be the sets of negative and positive observations, respectively. Now, $n = |N|$ is the number of negatives (i.e., the cardinality of N) and $p = |P|$ the number of positives (i.e., the cardinality of P). The variables i and j index the elements in P and N, respectively.

NRI was developed to address the problem of how to compare and quantify the improvement of risk prediction models. NRI has become very popular in the medical field.[40] Like all prediction model performance statistics, NRI has its strengths and drawbacks. Among its strengths is that NRI has much higher statistical power to detect differences between prediction models than e.g. ROC AUC, and among its drawbacks are that NRI does not take prevalence and costs of misclassification into account.[40] AUC is still the recommended standard measurement in the field of this study and we here complement the use of AUC by reporting on NRI.

RESULTS

CDK was extended to support count fingerprints and sparse representations of fingerprints. The signature fingerprints were implemented using these CDK extensions, which have been integrated into the CDK code base and made available from the CDK code repository at https://github.com/cdk.

We compared the bit-signature fingerprint Tanimoto approach with the CountTc1, CountTc2, CDKExtended, FOYFI, ECFP, and ECFI by calculating AUC and NRI. The AUC values are shown in violin plots[41] for each method in Figure 4. We also made repeated Wilcoxon tests with

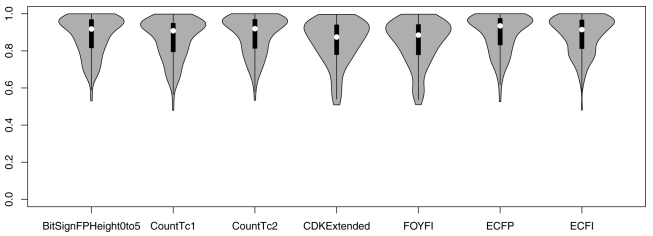


Figure 4. Area under ROC curves. Violin plots for the area under the ROC curve for the various fingerprints. A violin plot is a combination of a box plot and a rotated kernel density plot on each side of the box plot.[41]

Bonferroni correction for multiple testing for the values as seen in Table 2. The results of the NRI analysis are shown as violin plots showing the performance of the fingerprints when compared to the bit signature fingerprint in Figure 5. Table 3 contains the results of repeated Wilcoxon tests with Bonferroni correction for the NRI comparisons.

The results are further summarized and discussed below.

DISCUSSION

Performance of the Signature Fingerprints. From Figure 4 together with Tables 2 and 3 it is clear that ECFP,

dx.doi.org/10.1021/ci500361u | J. Chem. Inf. Model. 2014, 54, 2647−2653

Table 2. Wilcoxon Signed Rank Tests for Areas under ROC Curves[a]

	AUC median	95% CI	*p*-value
BitSignFPHeight0to5	0.892	(0.871, 0.909)	
CountTc1 (diff.)	−0.014	(−0.020, −0.008)	<0.0001
CountTc2 (diff.)	0.003	(0.000. 0.006)	0.0392
CDKExtended (diff.)	−0.028	(−0.036, −0.021)	<0.0001
FOYFI (diff.)	−0.028	(−0.037, −0.020)	<0.0001
ECFP (diff.)	0.008	(0.004, 0.012)	<0.0001
ECFI (diff.)	0.003	(0.000, 0.007)	0.0285

[a]The first row of the table shows the median with confidence interval for the signature fingerprint, and the following rows show the differences in AUC for the other fingerprints compared to the bit signature fingerprint. The null hypothesis for the test with the reported *p*-values is that there are no differences between the tested fingerprints and the bit signature fingerprint. A series of seven Wilcoxon signed rank tests were performed. With Bonferroni corrections, the corresponding 0.05 *p*-value threshold becomes 0.007 14. The confidence interval is a non-parametric confidence interval and was calculated by the wilcox.test function in R.[37]

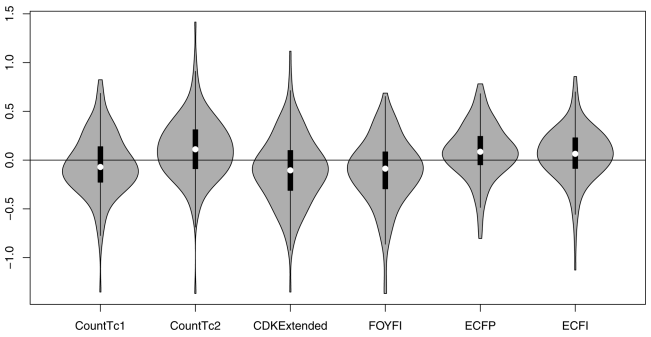


Figure 5. NRI values for fingerprints compared with the bit signature fingerprint. Violin plots for the NRI values against the bit signature fingerprint. The line at NRI 0 corresponds to no difference between the tested fingerprint and the bit signature fingerprint.

Table 3. Wilcoxon Signed Rank Tests for NRI Differences to Bit Signature Fingerprint[a]

	NRI median	95% CI	*p*-value
CountTc1	−0.054	(−0.099, −0.004)	0.0355
CountTc2	0.111	(0.065, 0.157)	<0.0001
CDKExtended	−0.100	(−0.153, −0.048)	0.0003
FOYFI	−0.103	(−0.153, −0.053)	<0.0001
ECFP	0.089	(0.048, 0.131)	<0.0001
ECFI	0.065	(0.027, 0.103)	0.0013

[a]NRI values were calculated as each fingerprint compared with the bit signature fingerprint. The null hypothesis for the test with the reported *p*-value is that there are no differences between the tested fingerprints and the bit signature fingerprint. A series of six Wilcoxon signed rank tests were performed. With Bonferroni corrections, the corresponding 0.05 *p*-value threshold becomes 0.0083. The confidence interval is a nonparametric confidence interval and was calculated by the wilcox.test function in R.[37]

CountTc2, and ECFI perform better than the other fingerprints. Looking only at Table 2 and taking the Bonferroni corrections into account it is unclear whether ECFI and ECFP indeed behaves better but the extra power of the NRI based tests in Table 3 indicate that in fact they belong to that group. Based on the 95% confidence interval in Table 2 it is tempting to say that the ECFP fingerprint behaves better than the CountTc2 which seems to be the best performing signature fingerprint, but those values are uncorrected for multiple testing and the NRI based values in Table 3 do not confirm that so in fact we are not able to distinguish which of the fingerprints behaves the best.

The higher power of NRI is reflected in the much lower p-values for CountTc2 and ECFI in Table 3 where they both are under the Bonferroni corrected p-value threshold. The count fingerprint which is used in CountTc2 takes extra effort (execution time and memory) compared to the bit signature fingerprint and whether it is worth it is probably something that needs to be decided on a case-by-case basis.

The results were obtained using kNN as the learning method based on Tanimoto distances. The absolute differences between the tested fingerprints are very small and since the optimal choice of learner is data set specific, we recognize that the results may change if a different learning method is used. However, these results show that the signature fingerprints as open-source alternatives, can perform at a similar level as the, to our knowledge, best other fingerprints available.

Performance Metrics. AUC has, in the field of virtual chemical screening, been criticized for not handling the "early recognition" problem.[42,43] In other words, the AUC does not specifically score the performance of a method in the top few percent but rather is an average measurement for the entire ranking list. This is a valid problem for a virtual screen where the goal is to avoid doing a lab screen of all the compounds, but rather only of the best scoring fraction from the virtual screen. However, this is based on an assumed cost structure where it is known how much false positives and false negatives cost. Since this is rarely the case, and because AUC has other positive properties,[14] we do not consider this an argument strong enough for disqualifying AUC from the kind of studies performed for this article.

AUC is an average score for the entire operating span but when actually using this method to classify substances a decision threshold has to be chosen. In this situation NRI would be a highly relevant metric for comparing the performance of two classification approaches. The AUC score has been shown to, in most cases, correspond well to the performance of a classifier based on such a threshold.[14] The choice of this threshold depends highly on project specific preferences so we include all predictions in the Supporting Information so different choices of this threshold can be tested by other researchers.

A recently published article by Riniker and Landrum[44] introduces a fingerprint benchmarking platform which calculates not only AUC but also a few measurements aimed at early recognition. They find that these early recognition measurements deliver different results than AUC and make the recommendation that at least one of these early recognition measurements should be used as complement to AUC. Although we believe this result can be interesting in some cases, we think it is important to remember that these measurements are prevalence dependent and that the amount of weight to be put on early recognition is based on an assumption of a cost structure that is not always fully known. Certainly not so in our case.

Given a prevalence estimation, it would be possible to select a decision threshold such that a certain precision or PPV is achieved when using the method as a predictor, e.g. it might be considered desirable to get on average 40% true positives when

using the method as an interactive decision support. Since we have no prevalence estimation this is impossible.

CONCLUSION

We have introduced two open-source signature fingerprints (bit and count versions) and shown that they perform on par with the, to our knowledge, best performing fingerprints available. The count signature fingerprint, the ECFP fingerprint and the ECFI fingerprint show statistically significant better results than the bit signature fingerprint. The NRI based tests complements the AUC based ones and show indications of higher power than the AUC based tests. AUC values are at around 0.9, meaning that in 90% of the cases a positive and a negative example drawn from our test set would be correctly distinguished.

Target interactions predicted with methods introduced in this study can be employed to predict off-target pharmacology and, in a next step, e.g., adverse drug reactions (ADR) since it is well-known that ADRs are often associated with particular targets.[45,46]

ASSOCIATED CONTENT

Supporting Information

zip file with training and test sets as well as all predictions and information describing this data further. This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: jonathan.alvarsson@farmbio.uu.se (J.A.).
*E-mail: tobias.noeske@astrazeneca.com (T.N.).

Author Contributions

J.A. and O.E. extracted the data set. J.A. implemented the signature fingerprint and conducted the analysis. M.E. helped with planning the statistical tests and the study design. L.C. provided valuable feedback regarding the signatures and contributed to the analysis. T.N. and O.E. provided expertise in chemical fingerprints. O.S. contributed with result interpretation. J.E.S.W. coordinated the Uppsala lab. J.A., O.S., M.E., and J.E.S.W. drafted the manuscript. All authors read, provided valuable feedback to, and approved the final manuscript.

Notes

The authors declare the following competing financial interest(s): J.E.S.W., O.S., and M.E. hold shares in Genetta Soft AB, a Swedish incorporated company.

ACKNOWLEDGMENTS

The authors want to thank the CDK development team who have made this study possible. This study was financially supported by AstraZeneca, the Swedish Research Council (VR-2011-6129), and the Swedish strategic research program eSSENCE. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project p2010047.

REFERENCES

(1) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug. Discovery* **2004**, *3*, 711−5.
(2) Mestres, J.; Martn-Couce, L.; Gregori-Puigjané, E.; Cases, M.; Boyer, S. Ligand-based approach to in silico pharmacology: nuclear receptor profiling. *J. Chem. Inf. Model.* **2006**, *46*, 2725−36.
(3) Hopkins, A. L. Network pharmacology. *Nat. Biotechnol.* **2007**, *25*, 1110−1.
(4) Stumpfe, D.; Bajorath, J. Similarity searching. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 260−282.
(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Model.* **1998**, *38*, 983−996.
(6) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.
(7) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(8) Boström, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion? *J. Med. Chem.* **2006**, *49*, 6716−6725.
(9) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(10) Faulon, J.-L.; Visco, D. P.; Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707−20.
(11) Faulon, J.-L.; Churchwell, C. J.; Visco, D. P. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721−734.
(12) Faulon, J.-L.; Collins, M. J.; Carr, R. D. The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 427−436.
(13) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, *132*, 1115−1118.
(14) Nicholls, A. What do we know and when do we know it? *J. Comput. Aided Mol. Des.* **2008**, *22*, 239−255.
(15) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29−36.
(16) Ware, J. H. The limitations of risk factors as prognostic tools. *N. Engl. J. Med.* **2006**, *355*, 2615−2617.
(17) Pencina, M. J.; D'Agostino, R. B.; Vasan, R. S. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **2008**, *27*, 157−172 discussion 207−212.
(18) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *44*, 1−8.
(19) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data - A Statistical Analysis. *PloS one* **2013**, *8*, e61007.
(20) Heikamp, K.; Bajorath, J. Comparison of confirmed inactive and randomly selected compounds as negative training examples in support vector machine-based virtual screening. *J. Chem. Inf. Model.* **2013**, *53*, 1595−1601.
(21) Elkan, C.; Noto, K. Learning classifiers from only positive and unlabeled data. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining−KDD 08*, Las Vegas, Aug 24−27, 2008; p 213.
(22) Weis, D. C.; Visco, D. P. Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. *Comput. Chem. Eng.* **2010**, *34*, 1018−1029.
(23) Carlsson, L.; Helgee, E. A.; Boyer, S. Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. *J. Chem. Inf. Model.* **2009**, *49*, 2551−8.
(24) Spjuth, O.; Eklund, M.; Ahlberg Helgee, E.; Boyer, S.; Carlsson, L. Integrated decision support for assessing chemical liabilities. *J. Chem. Inf. Model.* **2011**, *51*, 1840−7.
(25) Norinder, U.; Ek, M. E. QSAR investigation of NaV1.7 active compounds using the SVM/Signature approach and the Bioclipse Modeling platform. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 261−263.
(26) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-

source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.
(27) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111−2120.
(28) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771−784.
(29) Steffen, A.; Kogej, T.; Tyrchan, C.; Engkvist, O. Comparison of molecular fingerprint methods on the basis of biological profile data. *J. Chem. Inf. Model.* **2009**, *49*, 338−347.
(30) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(31) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **2005**, *10*, 682−6.
(32) Heikamp, K.; Bajorath, J. Large-scale similarity search profiling of ChEMBL compound data sets. *J. Chem. Inf. Model.* **2011**, *51*, 1831−1839.
(33) Pipeline Pilot. http://accelrys.com/products/pipeline-pilot/ (accessed September 29, 2014).
(34) Karunaratne, T.; Bostrom, H.; Norinder, U. Pre-Processing Structured Data for Standard Machine Learning Algorithms by Supervised Graph Propositionalization−A Case Study with Medicinal Chemistry Datasets. *2010 Ninth International Conference on Machine Learning and Applications*, Bethesda, MD, Dec 12−14, 2010; pp 828−833.
(35) The Apache Commons Mathematics Library. http://commons.apache.org/proper/commons-math/ (accessed September 29, 2014).
(36) Roger, S. *Benchmarking and Validation of JChem. ECFP and FCFP Fingerprints*; NextMove Software Ltd: Cambridge UK., 2011; http://www.chemaxon.com/wp-content/uploads/2011/05/NextMovePoster3.pdf (accessed September 29, 2014).
(37) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008; ISBN 3-900051-07-0.
(38) Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Muller, M. pROC: an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinformatics* **2011**, *12*, 77.
(39) Harrell, F.; et al. *Hmisc: Harrell Miscellaneous*; 2012; R package version 3.9-1.
(40) Kerr, K. F.; Wang, Z.; Janes, H.; McClelland, R. L.; Psaty, B. M.; Pepe, M. S. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology* **2014**, *25*, 114−121.
(41) Hintze, J. L.; Nelson, R. D. Violin Plots: A Box Plot Density Trace Synergism. *Am. Stat.* **1998**, *52*, 181−184.
(42) Truchon, J.-F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.
(43) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput. Aided Mol. Des.* **2008**, *22*, 141−146.
(44) Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* **2013**, *5*, 26.
(45) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361−367.
(46) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; et al. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175−181.