JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

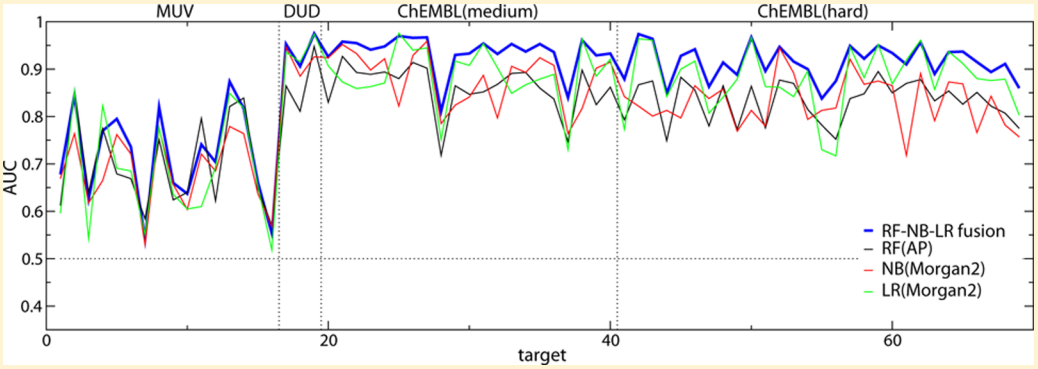
pubs.acs.org/jcim

Heterogeneous Classifier Fusion for Ligand-Based Virtual Screening: Or, How Decision Making by Committee Can Be a Good Thing

Sereina Riniker, Nikolas Fechner, and Gregory A. Landrum*

Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, CH-4056 Basel, Switzerland

Supporting Information



ABSTRACT: The concept of data fusion - the combination of information from different sources describing the same object with the expectation to generate a more accurate representation - has found application in a very broad range of disciplines. In the context of ligand-based virtual screening (VS), data fusion has been applied to combine knowledge from either different active molecules or different fingerprints to improve similarity search performance. Machine-learning (ML) methods based on fusion of multiple homogeneous classifiers, in particular random forests, have also been widely applied in the ML literature. The heterogeneous version of classifier fusion - fusing the predictions from different model types - has been less explored. Here, we investigate heterogeneous classifier fusion for ligand-based VS using three different ML methods, RF, naïve Bayes (NB), and logistic regression (LR), with four 2D fingerprints, atom pairs, topological torsions, RDKit fingerprint, and circular fingerprint. The methods are compared using a previously developed benchmarking platform for 2D fingerprints which is extended to ML methods in this article. The original data sets are filtered for difficulty, and a new set of challenging data sets from ChEMBL is added. Data sets were also generated for a second use case: starting from a small set of related actives instead of diverse actives. The final fused model consistently outperforms the other approaches across the broad variety of targets studied, indicating that heterogeneous classifier fusion is a very promising approach for ligand-based VS. The new data sets together with the adapted source code for ML methods are provided in the Supporting Information.

INTRODUCTION

Data fusion describes the combination of complementary information from multiple sources with the aim to generate a combined representation that is more informative than the individual ones.[1] Originally developed for defense applications in the 1980s, the approach has been applied since then in a wide variety of fields ranging from astronomy and neurocomputing to speech processing.[2]

In similarity-search driven virtual screening (VS), potentially active compounds are retrieved from a pool of compounds based on the similarity to a known active molecule. The basic idea behind this approach is the similarity principle which states that similar molecules have similar properties. However, the description of molecular similarity is not uniquely defined, and a large number of descriptors have been developed for this task.[3−5] Their performances are highly target dependent, and no single best descriptor could be identified so far. Data fusion allows us to take advantage of these differences.[6,7] It can be used to fuse information from different fingerprints of the same active molecule (i.e., similarity fusion) or from different actives with the same fingerprint (i.e., group fusion).[7] Both methods have found to improve the performance of fingerprints in similarity search, and group fusion has become a standard for ligand-based VS.[8−13] In a recent example, the approach was successfully applied to the fusion of both 2D fingerprints and 3D methods.[14] The information is combined based on a fusion rule.[7] Most commonly, unsupervised rules such as MAX, where the maximum input value is the output, or AVE, where the input values are averaged, are used. MAX fusion has found to be the method of choice for similarity data.[7]

Data fusion is an often used approach in the machine-learning (ML) community. The fusion of ML classifiers has many different names: e.g. committee of learners, classifier ensembles, mixture of experts, bagging (see refs 15−23 for reviews).

Received: August 7, 2013
Published: October 30, 2013

The general idea is to generate a strong model by integrating (fusing) a set of weak learners. Usually a homogeneous ensemble of classifiers is trained with subsets of feature and/or sample space, and the predictions are combined using a fusion rule such as voting. The random forest (RF)[24] is an example for homogeneous classifier fusion. An RF is an ensemble of decision trees which see a randomly selected subset of samples and features. Although the individual trees may be overfitted, the ensemble is not and has an increased accuracy.[24] A much less often used form of classifier fusion is the combination of heterogeneous ML models.[25−31] Thereby, models of different ML methods are trained separately on the same set of data, and the predictions are fused to give a final prediction.

ML methods have generally been found to improve the performance of fingerprints in similarity search. Reference 32 gives an overview of ML methods in ligand-based VS. Examples are naïve Bayes (NB),[33−35] binary kernel discrimination (BKD),[12,33,36,37] support-vector machines (SVM),[34,38,39] and neural networks.[40] RFs have been successfully applied to QSAR modeling.[41,42] The performance of the various methods/fingerprint combinations have found to be again highly dependent on the data set, indicating that they may have learned complementary information which could be exploited by classifier fusion. However, apart from the use of RFs, classifier fusion and especially the heterogeneous type has not found wide application in ligand-based VS. One early example is the combination of neural networks and decision trees to predict the inhibition of a P450 enzyme, where the selection of classifiers and combination rule was done using genetic programming.[26,27]

Here, we further investigate and benchmark the performance of heterogeneous classifier fusion in ligand-based VS using three different ML methods, RF, NB, and logistic regression (LR), and four standard 2D fingerprints. The approach is compared to homogeneous classifier fusion and classifiers trained with hybrid fingerprints[43] using an extension of a previously developed benchmarking platform.[44] The MAX group fusion of the 2D fingerprints serves thereby as a baseline. A simple rank-based MAX fusion rule is used for the combination of ML models. In addition to the previously described data sets which contain a diverse set of training actives, data sets for a second use case of VS have been collected which contain a set of related training actives. The additional data sets and the source code of the ML extension of the benchmarking platform are provided as Supporting Information.

METHODS AND MATERIALS

2D Fingerprints. The 2D fingerprints used in this study are described in detail in ref 44. All fingerprints were generated using the open-source cheminformatics toolkit RDKit[45] version 2013.03. Four fingerprints were investigated: atom pairs (AP),[46] topological torsions (TT),[47] RDKit fingerprint with maximum path length = 5 (RDK5),[45] and circular fingerprint with radius = 2 (Morgan2, the RDKit implementation[48] of the familiar ECFP4[49]). In constrast to ref 44, hashed bit-vector versions of AP and TT were used here, where the counts of an atom pair are represented by up to four bits. The performance of the count-version and the hashed-version of AP and TT are compared in Figure S1 in the Supporting Information. The bit vectors of all four fingerprints had the size 1024 bits.

MAX group fusion[6] was used for the 2D fingerprints, and Dice similarity[50] was used as the similarity measure.

Machine-Learning Methods. In machine-learning (ML) methods, each bit of a fingerprint is considered a feature. Three kinds of ML methods were investigated: random forest (RF),[24] naïve Bayes (NB), and logistic regression (LR). The classifiers were calculated using the open-source machine-learning toolkit scikit-learn[51] version 0.13.

Random Forest Classifier. An RF is an ensemble of unpruned decision trees[52] where each tree is built with a random subset of the data, and at each node the most important feature is chosen from a random subset of features.[24] The RF of scikit-learn is controlled by the following parameters: number of trees (NUMT), feature-selection criterion, number of randomly selected features at each node (MAXF), maximum depth (MAXD), minimum samples to split an internal node (MINSP), and minimum samples at a leaf (MINLF). We used the Gini impurity as selection criterion, the square root of the number of bits as MAXF at each node, MINF = 1 and MINSP = 2. The influence of NUMT and MAXD on the performance was found to be very small (data not shown), but the computational effort increases with increasing NUMT. Thus, a relatively small number of trees, i.e. 100, and MAXD = 10 were used throughout this study. To avoid the problem of imbalance in the training set (i.e., many more inactives than actives), a balancing algorithm[53] was applied where for each decision tree the majority class is down-sampled to yield an equal number of instances as the minority class. After training an RF for a given fingerprint and target, the test molecules were ranked based on the predicted probability to be active. As NUMT was chosen relatively small, multiple test molecules could receive the exact same probability. Therefore, molecules with the same predicted probability were ranked further based on the 2D fingerprint similarity.

Naïve Bayes Classifier. The NB classifier is a probabilistic model which is based on the simplifying assumption of conditional independence among the features. Although this assumption is rarely true in real-world applications, the performance of NB has been found in empirical comparisons to be equal to decision trees.[54−56] The Bernoulli NB of scikit-learn is controlled by the following parameters: additive Laplace smoothing parameter α and a flag for learning class prior probabilities. The default parameters, α = 1.0 and learning of prior probabilities, were used in this study. The NB classifier was trained using the complete imbalanced training set for each target and repetition. The test molecules were ranked based on the predicted probability to be active.

Logistic Regression Classifier. LR is a regression-analysis method used for problems with a binary outcome variable as in the present case. Thus, a "hit or miss" function instead of the sum of square residuals (as in linear regression) is minimized. The LR of scikit-learn is controlled by the following parameters: norm for penalizing errors, strength of regularization C, intercept added to the decision function, intercept scaling (INTSC), class weights (CW), and tolerance for the stopping criteria (TOL). The default parameters, "L2" norm, C = 1.0, adding an intercept, INTSC = 1.0, no CW and TOL = 0.001, were used in this study. The LR was trained using the complete imbalanced training set for each target and repetition . The test molecules were ranked based on the predicted probability to be active.

Fingerprint Fusion and Classifier Fusion. Hybrid fingerprints of Morgan2 and AP (each 1024 bits) were generated by simply concatenating the two fingerprints to form a 2048-bit long vector. The first 1024 bits come from the Morgan2 fingerprint and the second 1024 bits from the AP fingerprint. The hybrid fingerprints were used to train an RF or NB classifier. The hybrid fingerprints were also used for traditional similarity-based searching but were not found to be an improvement over the individual fingerprints (data not shown).

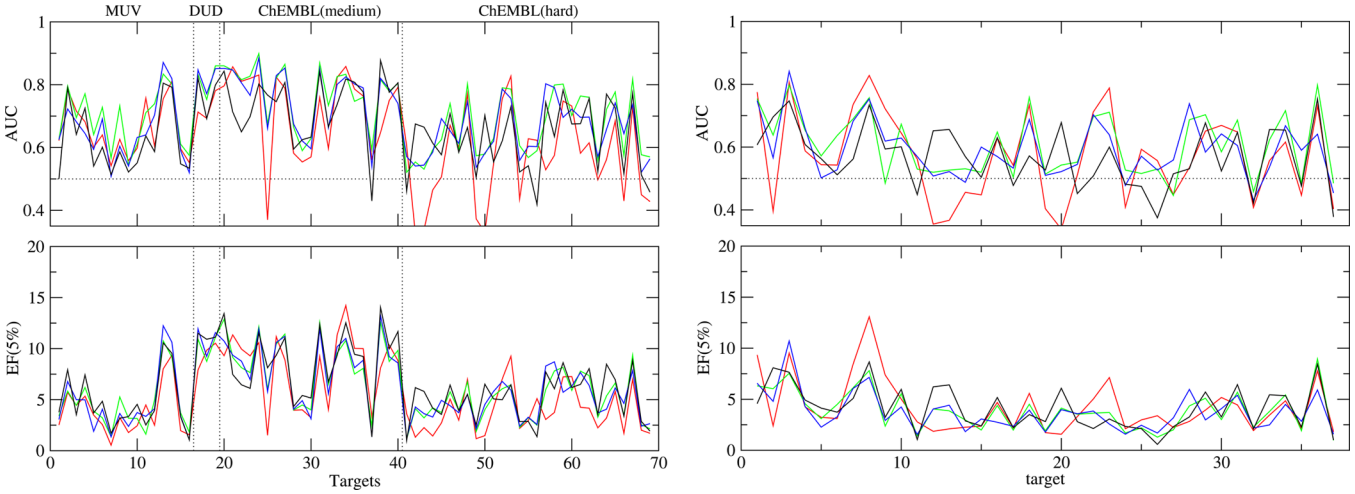


Figure 1. Average performance of four standard fingerprints, Morgan2 (black), AP (red), TT (green), and RDK5 (blue) measured with AUC (top) and EF(5%) (bottom) for data sets I (left) and data sets II (right). The horizontal, dotted line indicates random distribution.

Classifier fusion was performed by applying rank-based MAX fusion to multiple ML models. Each model assigns a rank to a molecule (highest probability = highest rank), and the fusion score of that molecule is the maximum rank assigned. The molecules are then reranked using this fusion score to get a final ranking. If two molecules have the same fusion rank, they are further ranked based on the maximum predicted probability. Classifier fusion was performed for the three ML methods with two 2D fingerprints, AP and Morgan2.

Compound Data Sets. For the comparisons between methods we report here, we have assembled data sets that simulate two different standard VS starting points:

1. A small set of diverse actives from, for example, a high-throughput screen is available.

2. A small set of related actives, i.e. compounds sharing one or two common scaffolds, from a publication or patent is available.

In each case, we have a small number of actives and are interested in finding interesting new actives from a large pool of diverse compounds.

For the first VS use case, data sets from three different public sources were used: the directory of useful decoys (DUD),[57−59] the maximum unbiased validation (MUV) data sets,[60,61] and a selection of targets from ChEMBL[62,63] proposed by Heikamp and Bajorath.[64] These data sets are described in detail in ref 44. For the ChEMBL targets, the 100 most diverse actives were selected for the benchmarking data sets,[44] and query molecules were drawn from this pool in the VS experiments. In order to be able to show some improvement using ML methods, data sets where the average baseline AUC performance using CountMorgan0 (i.e., ECFC0) was above 0.8 were removed. This resulted in 40 data sets. In addition, the 30 ChEMBL targets marked as difficult in the study of Heikamp and Bajorath[64] were added to the collection. One of these targets (target id 100166) had still an average AUC(CountMorgan0) value greater than 0.8 and was not considered. This led to a final number of 69 targets, which are referred to as *data sets I* throughout this study. An overview of the 69 data sets with the target IDs, target description, number of actives, number of decoys, and ratio actives/decoys is given in Table S1 in the Supporting Information. The compound lists of the additional ChEMBL targets with the SMILES and external IDs are given in the Supporting Information.

In order to simulate the second VS use case, a somewhat different approach to constructing data sets is required. Most of the data in ChEMBL have been extracted from the scientific literature. The typical medicinal-chemistry paper includes data on one or two chemical series, together with reference compound(s). These papers provide ideal starting points for our simulation. To build data sets for this use case, we started with the papers from which actives were chosen for the 50 ChEMBL targets in Table S1 in the Supporting Information: 21 from the "medium" set, 29 from the "hard" set. Papers that had less than ten actives were removed, as were targets where less than four papers remained. This resulted in 37 targets with 4 - 37 papers and 10 - 112 actives per paper (Table S2 in the Supporting Information). These data sets are referred to as *data sets II* throughout this study. The compounds lists with the SMILES and external IDs are given in the Supporting Information. The same decoys as for the ChEMBL targets of data sets I were used.

Virtual Screening. The VS experiments in this study were performed using the benchmarking platform described in ref 44. The scoring step of the benchmarking process had to be adjusted for the use of ML methods and the corresponding Python scripts are given in the Supporting Information. The benchmarking platform uses the open-source cheminformatics toolkit RDKit[45] version 2013.03 and the open-source machine-learning toolkit scikit-learn[51] version 0.13.

The VS experiment was divided into three steps. In the scoring step, the test molecules were ranked based on similarity or predicted probability. In the validation step, the performance of the fingerprints and ML methods was evaluated using the area under the ROC curve (AUC) and the enrichment factor at 5% (EF(5%)). AUC considers the whole test set, whereas EF(5%) is an "early-recognition" method. Additional "early-recognition" evaluation methods were not considered because our previous study showed that they are strongly correlated with EF(5%).[44] In the analysis step, the results were collected and ordered for each evaluation method as tables with fingerprints/ML methods (columns) and targets (rows).

For the data sets I, the VS experiment was repeated for each target 50 times with different randomly selected training sets. To ensure reproducibility of the results, the precalculated training sets are provided as part of the benchmarking platform. The training sets consisted of ten actives and 20% of the decoys

dx.doi.org/10.1021/ci400466r | J. Chem. Inf. Model. 2013, 53, 2829−2836

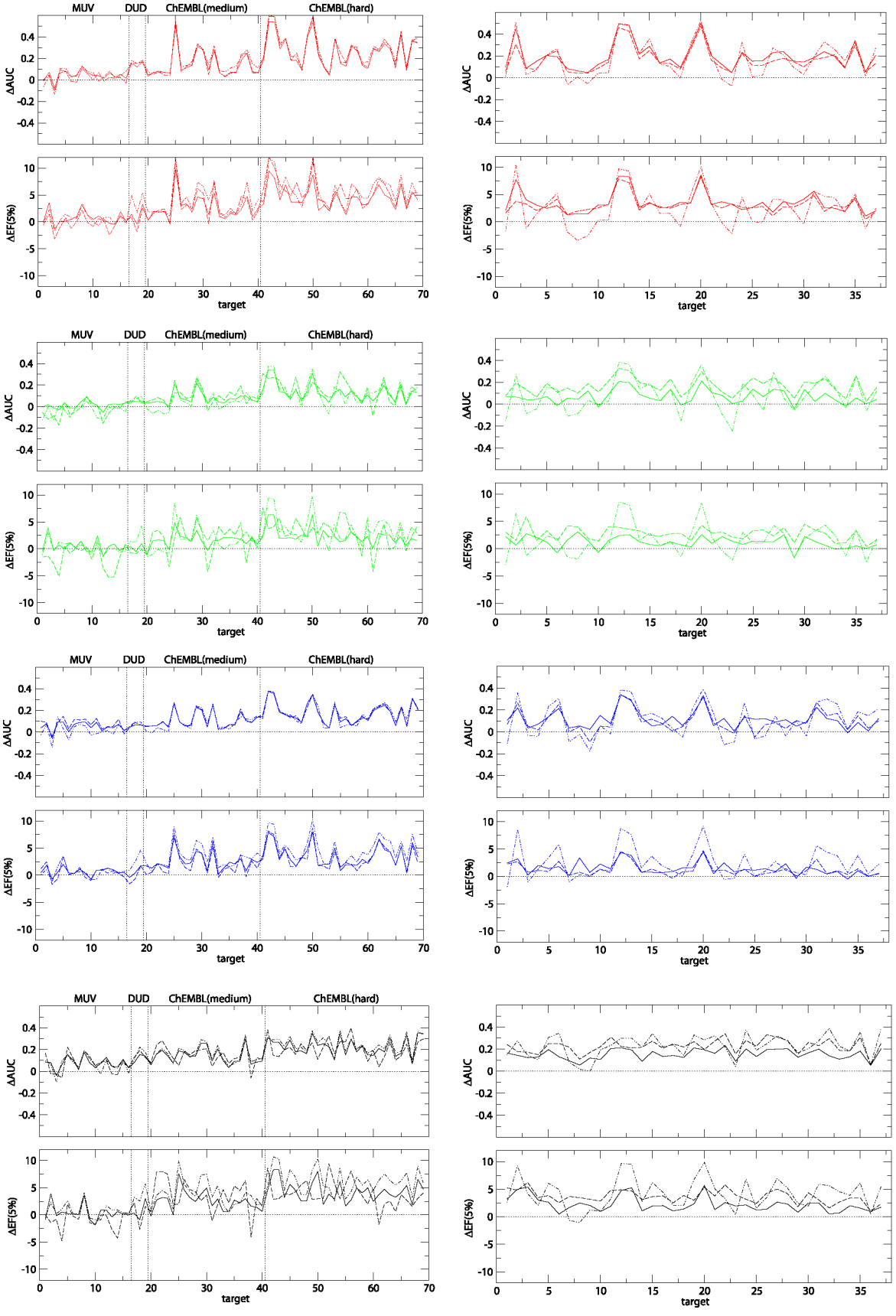


Figure 2. Average performance difference between random forest (RF) (solid lines), naïve Bayes (NB) (dashed lines), or logistic regression (LR) (dashed-dotted lines), and the standard fingerprint measured with ΔAUC and ΔEF(5%) for data sets I (left) and data sets II (right). Four standard fingerprints are compared: AP (red), TT (green), RDK5 (blue), and Morgan2 (black).

dx.doi.org/10.1021/ci400466r | J. Chem. Inf. Model. 2013, 53, 2829−2836

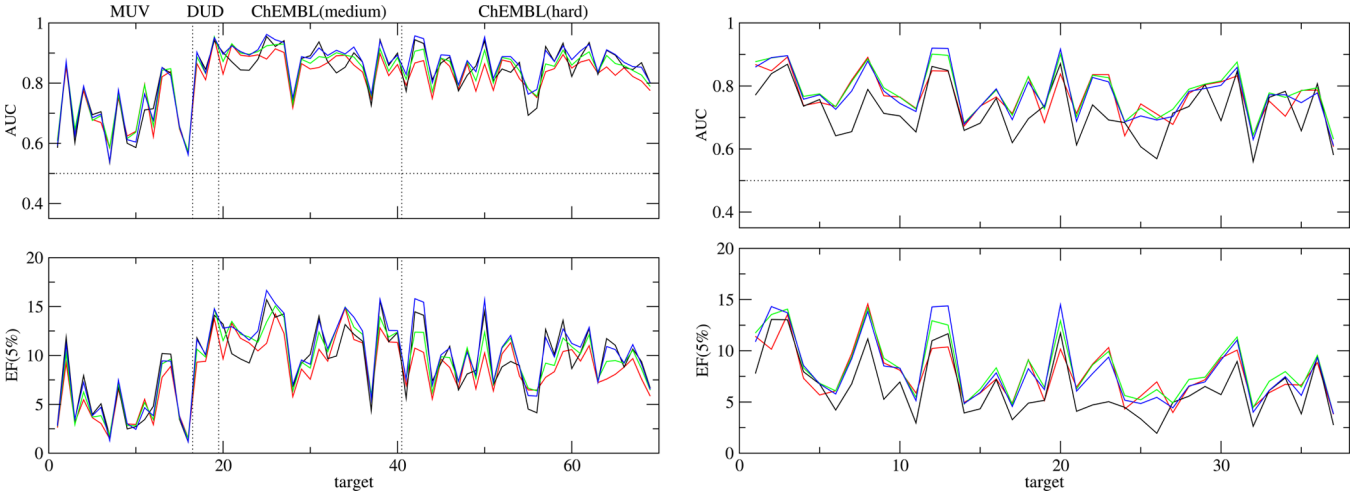


Figure 3. Average performance of RF(AP) (red), RF(Morgan2) (black), RF(AP-Morgan2) (green), and RF(AP)-RF(Morgan2) (blue) measured with AUC (top) and EF(5%) (bottom) for data sets I (left) and data sets II (right).

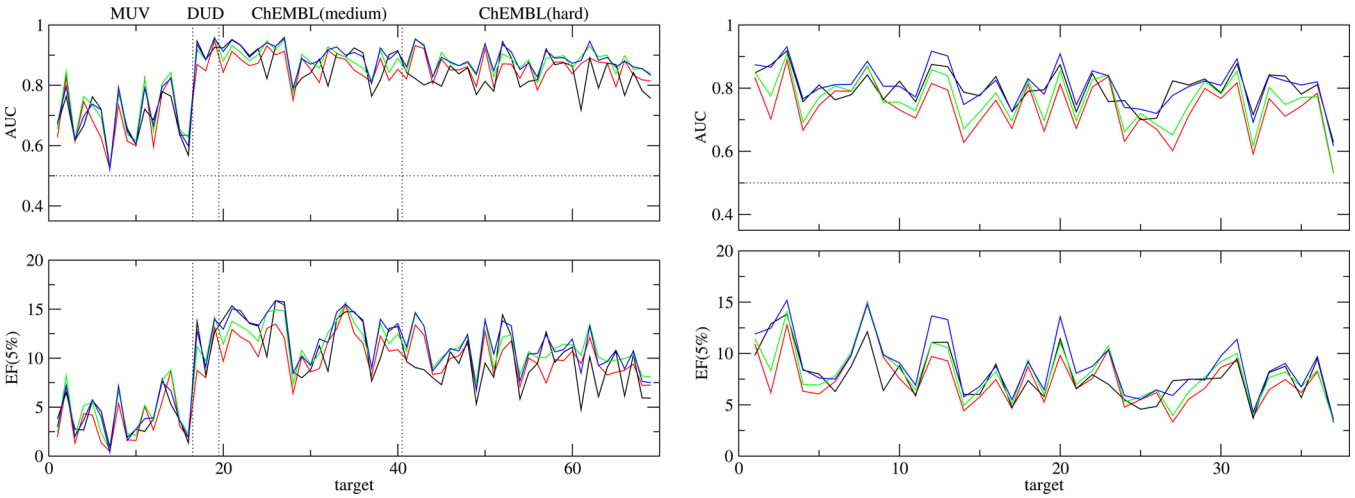


Figure 4. Average performance of NB(AP) (red), NB(Morgan2) (black), NB(AP-Morgan2) (green), and NB(AP)-RF(Morgan2) (blue) measured with AUC (top) and EF(5%) (bottom) for data sets I (left) and data sets II (right).

(random selection). For the data sets II, the VS experiment was performed once for each paper using all actives of the paper and 10% of the decoys for the training. The test set consisted of the 99 actives from the benchmarking data set for the same target and the rest of the decoys. The smaller number of training decoys was chosen to have approximately the same ratio between actives and inactives in the test molecules as in the ChEMBL targets of data sets I.

RESULTS AND DISCUSSION

The numerical values of all figures are given as csv-files in the Supporting Information.

2D Fingerprints. The previous benchmarking study showed that no single best fingerprint could be isolated from the fourteen commonly used 2D fingerprints investigated as the performance is highly target dependent.[44] The four fingerprints selected for this study, i.e. atom pairs (AP), topological torsions (TT), RDKit fingerprint (RDK5), and Morgan fingerprint (Morgan2), were among the top fingerprints previously, and they retrieve rather different actives in the first 5% (Figure S2 in the Supporting Information). The performance of the four fingerprints over the 69 targets of data sets I and the 37 targets of data sets II (Figure 1) showed that the data sets are indeed difficult for 2D fingerprints. The data for all 14 fingerprints used in ref 44 is shown in Figure S3 in the Supporting Information. Note that in constrast to ref 44 a hashed version of AP and TT folded to 1024 bits is used here. For TT, the performance of the two versions is comparable, whereas the hashed bit-vector version of AP showed a somewhat lower performance (Figure S1 in the Supporting Information).

Machine-Learning Methods. The performance of the random forest (RF), naïve Bayes (NB) and logistic regression (LR) classifier was compared to that of the 2D fingerprints (Morgan2, AP, TT, and RDK5) with MAX group fusion. The average difference in performance between the three ML methods and the 2D fingerprints is shown in Figure 2. The absolute performances are shown in Figures S4 and S5 in the Supporting Information. In agreement with previous studies, the use of ML methods generally increased the performance of 2D fingerprints, despite the small number of actives available for training (i.e., ten). The best ML method/fingerprint combination was highly target dependent, with no method consistently outperforming the others. This is also reflected in the results of the pairwise posthoc Friedman tests (described in ref 44) where most methods showed no statistically significant differences to

dx.doi.org/10.1021/ci400466r | J. Chem. Inf. Model. 2013, 53, 2829−2836

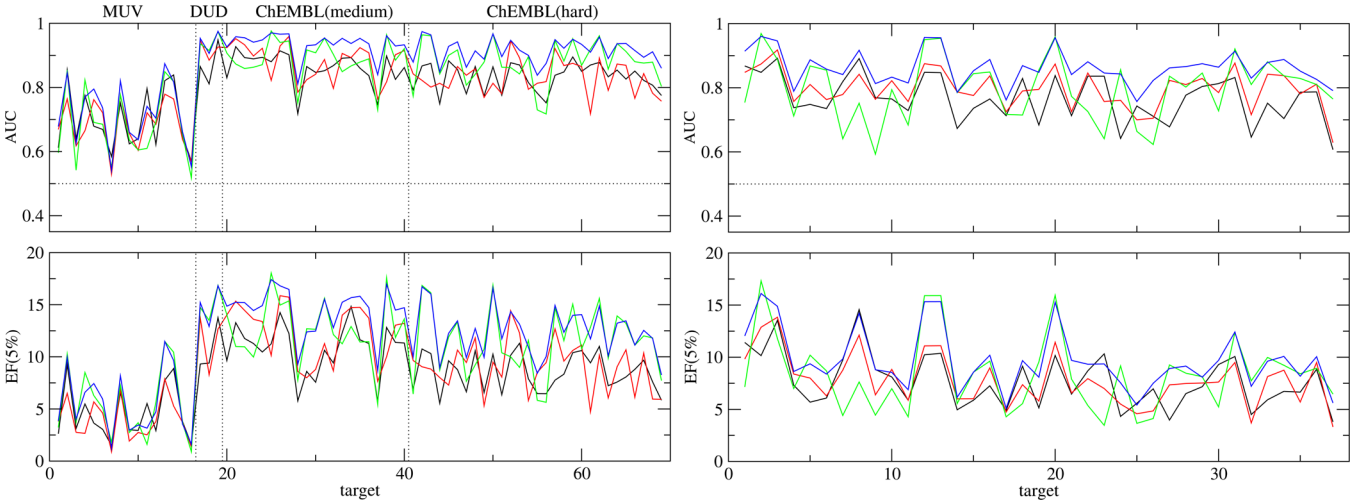


Figure 5. Average performance of RF(AP) (black), NB(Morgan2) (red), LR(Morgan2) (green), and RF(AP)-NB(Morgan2)-LR(Morgan2) (blue) measured with AUC (top) and EF(5%) (bottom) for data sets I (left) and data sets II (right).

Table 1. Results for Data Sets I: Pairwise Post-Hoc Friedman Tests of the Average Rank between the ML Methods Random Forest (RF) Trained with Atom Pairs (AP), Naïve Bayes (NB) Trained with the Morgan Fingerprint (Morgan2, M2) and Logistic Regression (LR) Trained with Morgan2, As Well As the Classifier Fusion Models RF(AP)-RF(M2), NB(AP)-NB(M2), RF(AP)-NB(M2), and RF(AP)-NB(M2)-LR(M2) for the Evaluation Methods AUC (Top) and EF(5%) (Bottom)[a]

AUC	RF(AP)-NB(M2)-LR(M2)	RF(AP)-NB(M2)	NB(AP)-NB(M2)	LR(M2)	RF(AP)-RF(M2)	NB(M2)	RF(AP)
RF(AP)-NB(M2)-LR(M2)		-	-	-	-	-	-
RF(AP)-NB(M2)			X	X	o	-	-
NB(AP)-NB(M2)				X	X	o	-
LR(M2)					X	o	-
RF(AP)-RF(M2)						X	o
NB(M2)							X
RF(AP)							

EF(5%)	RF(AP)-NB(M2)-LR(M2)	LR(M2)	RF(AP)-NB(M2)	NB(AP)-NB(M2)	RF(AP)-RF(M2)	NB(M2)	RF(AP)
RF(AP)-NB(M2)-LR(M2)		-	-	-	-	-	-
LR(M2)			X	X	X	-	-
RF(AP)-NB(M2)				X	o	-	-
NB(AP)-NB(M2)					X	o	-
RF(AP)-RF(M2)						X	o
NB(M2)							X
RF(AP)							

[a]Pairs of ML models with no statistically significant difference are marked with "X", pairs with an adjusted p-value distribution around the confidence level α with "o", and pairs with a statistically significant difference with "-". ML models are ordered according to ascending average rank.

each other (Tables S3 and S4 in the Supporting Information). Nevertheless, some trends can be observed. LR(Morgan2) had consistently the highest average rank for both data sets and evaluation methods. Despite the simplicity of LR, its performance with Morgan2 was clearly better than RF and NB for a number of targets. These targets are all carbonic anhydrases (ChEMBL target ids 12209, 10193, 15, and 12952). For this protein family, the substructure important for activity, i.e. binding to the positively charged zinc-ion in the active site, can be described with a very small number of features. The performance of LR(Morgan2) was not statistically different from the performance of NB(Morgan2), RF(AP), and NB(TT) for both evaluation methods with data sets II. LR(Morgan2) and NB(Morgan2) are also statistically indistinguishable for data sets I using the AUC method. NB(TT) ranked among the lowest methods with data sets I. Generally, the methods trained with TT showed the smallest improvement compared to the other fingerprints. This is partially due to how well TT does with the similarity searching, but it also indicates that this fingerprint is less well suited to train ML models.

Classifier Fusion. The results above show that there is no ML method/fingerprint combination that consistently performs better than the others for all data sets, but that the best method is target dependent. This result is not terribly surprising. As there is currently no approach known to decide *a priori* which method is likely to perform best for a certain target, and we know that the different model types are capturing different information, classifier fusion may present a way around this problem. Ideally, a fused model can be found that is similar to or better than the best individual model across all (or at least most) targets. The large number of possible combinations makes a full enumeration computationally infeasible, so we focus on two fingerprints, Morgan2 and AP, and fusion of up to three models. These two fingerprints generally gave good models and retrieve rather different actives (Figure S2 in the Supporting Information).

dx.doi.org/10.1021/ci400466r | J. Chem. Inf. Model. 2013, 53, 2829−2836

Table 2. Results for Data Sets II: Pairwise Post-Hoc Friedman Tests of the Average rank between the ML Methods Random Forest (RF) Trained with Atom Pairs (AP), Naïve Bayes (NB) Trained with the Morgan Fingerprint (Morgan2, M2) and Logistic Regression (LR) Trained with Morgan2, As Well As the Classifier Fusion Models RF(AP)-RF(M2), NB(AP)-NB(M2), RF(AP)-NB(M2), and RF(AP)-NB(M2)-LR(M2) for the Evaluation Methods AUC (Top) and EF(5%) (Bottom)[a]

AUC	RF(AP)-NB(M2)-LR(M2)	RF(AP)-NB(M2)	NB(AP)-NB(M2)	LR(M2)	NB(M2)	RF(AP)-RF(M2)	RF(AP)
RF(AP)-NB(M2)-LR(M2)		-	-	-	-	-	-
RF(AP)-NB(M2)			X	X	o	-	-
NB(AP)-NB(M2)				X	X	-	-
LR(M2)					X	o	o
NB(M2)						o	o
RF(AP)-RF(M2)							X
RF(AP)							
EF(5%)	RF(AP)-NB(M2)-LR(M2)	RF(AP)-NB(M2)	NB(AP)-NB(M2)	LR(M2)	RF(AP)-RF(M2)	NB(M2)	RF(AP)
RF(AP)-NB(M2)-LR(M2)		o	o	-	-	-	-
RF(AP)-NB(M2)			X	X	-	-	o
NB(AP)-NB(M2)				X	o	o	o
LR(M2)					o	o	o
RF(AP)-RF(M2)						X	X
NB(M2)							X
RF(AP)							

[a]Pairs of ML models with no statistically significant difference are marked with "X", pairs with an adjusted p-value distribution around the confidence level α with "o", and pairs with a statistically significant difference with "-". ML models are ordered according to ascending average rank.

First, we considered homogeneous classifier fusion of two models trained with different fingerprints and using fingerprint fusion with a single ML model. In Figure 3, the fusion of two RFs trained with AP and Morgan2 (i.e., RF(AP)-RF(Morgan2)) is compared to an RF trained with the hybrid fingerprint AP-Morgan2 (i.e., RF(AP-Morgan2)) and the individual models (i.e., RF(AP) and RF(Morgan2)). Both fusion approaches yielded performances that are similar to or better than the best individual method for each target (especially for data sets II). The same comparison is shown for NB in Figure 4. Here, only NB(AP)-NB(Morgan2) yielded similar or better results when compared to the best individual method for each target, whereas the performance of NB(AP-Morgan2) was found to remain similar to NB(AP). Doing the fusion at the classifier level, as opposed to the fingerprint level, not only improves performance more, at least for the RFs, but also provides considerable flexbility. We will explore this in the next section.

Next, we investigated the fusion of heterogeneous classifiers. As the top performer with data sets I for both evaluation methods was LR(Morgan2), and for data sets II LR(Morgan2), NB(Morgan2), and RF(AP), these three methods were combined to give a fused model (Figure 5). The performance of RF(AP)-NB(Morgan2)-LR(Morgan2) was found to be better than or similar to the best individual method and the two-model fusions for each target. The differences are statistically significant for both evaluation methods and data sets (Tables 1 and 2). The classifier fusion approach was able to take advantage of the strengths of each method and fingerprint and thus presents an attractive approach for ligand-based VS.

CONCLUSIONS

The use of machine-learning (ML) methods in ligand-based virtual screening (VS) has generally been found to increase the performance of the 2D fingerprints the models were trained with. The same could be observed for the three ML methods investigated in this study, random forest (RF), naïve Bayes (NB), and logistic regression (LR) trained with four 2D fingerprints: atom pairs (AP), topological torsions (TT), the RDKit fingerprint (RDK5), and the circular fingerprint (Morgan2) using a previously developed benchmarking platform. The data sets of the platform, which consist of small sets of diverse actives, were filtered for difficulty and complimented with other challenging targets from the public database ChEMBL in order to challenge the ML methods. In addition, data sets for a second use case of VS (i.e., searching with a small set of related actives) were collected. Although ML methods generally lead to an increase in performance over fingerprint similarity for both data set collections, the best method/fingerprint combination was found to be highly target-dependent and many methods were statistically indistinguishable. Among the top performers were LR(Morgan2), NB(Morgan2), and RF(AP). The concept of homogeneous and heterogeneous classifier fusion at the ranking step was explored in order to take advantage of the differences between methods and fingerprints. Homogeneous classifier fusion of two models (RF or NB) trained with either AP or Morgan2 was found to perform similarly to or better than the individual models as well as a single model trained with a hybrid fingerprint of AP and Morgan2. The best results were observed, however, for heterogeneous classifier fusion with simple rank-based MAX fusion. The performance of the fused model of RF(AP), NB(Morgan2), and LR(Morgan2) was similar or better than the best individual method for all targets and thus was statistically significantly better than all of the other models. Using this fused model, AUC values between 0.8 and 1.0 could be achieved for the vast majority of targets compared to AUC values close to random for many targets with the fingerprints. Heterogeneous classifier fusion is thus a promising approach for ligand-based VS.

ASSOCIATED CONTENT

Supporting Information

Additional figures and tables mentioned in the text, **supplementary.pdf**; Python scripts of the benchmarking platform[44] modified for the use of machine-learning methods, **python_scripts.zip**; additional compound lists for data sets I, **compounds_I.tar.zip**; compound lists for data sets II, **compounds_II.tar.zip**; training lists with ten query molecules for data sets I, **training_lists_I.tar.zip**; training and test lists for data sets II, **training_test_lists_II.tar.zip**; numerical values of

the results (as csv files), **results.zip**. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: gregory.landrum@novartis.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S. R. thanks the Novartis Institutes for BioMedical Research education office for a Presidential Postdoctoral Fellowship. The authors thank Bernard Rohde for helpful discussions.

REFERENCES

(1) Hall, D. L.; Llinas, J. *Proc. IEEE* **1997**, *85*, 6−23.
(2) Dasarathy, B. V. A. *Inf. Fusion* **2010**, *11*, 299−300.
(3) Sheridan, R. P.; Kearsley, S. K. *Drug Discovery Today* **2002**, *7*, 903−911.
(4) Roth, H.-J. *Curr. Opin. Chem. Biol.* **2005**, *9*, 293−295.
(5) Bender, A. *Expert Opin. Drug Discovery* **2010**, *5*, 1141−1151.
(6) Willett, P. *QSAR Comb. Sci* **2006**, *25*, 1143−1152.
(7) Willett, P. *J. Chem. Inf. Model.* **2013**, *53*, 1−10.
(8) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(9) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.
(10) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23−37.
(11) Ginn, C. M. R.; Willett, P.; Bradshaw, J. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.
(12) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzoui, K.; Jacoby, E.; Schuffenhauer, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.
(13) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840−1848.
(14) Sastry, G. M.; Inakollu, V. S. S.; Sherman, W. *J. Chem. Inf. Model.* **2013**, online.
(15) Kittler, J.; Hatel, J. M.; Duin, R. P.; Matas, J. *IEEE Trans. Patt. Anal. Mach. Intell.* **1998**, *20*, 226−239.
(16) Dietterich, T. In *1st Int. Workshop on Mult. Class. Syst., Lect. Notes in Comput. Sci.*; Kittler, J., Roli, F., Eds.; Springer Verlag, 2000; pp 1−15.
(17) Ruta, D.; Gabrys, B. *Comput. Inf. Syst.* **2000**, *7*, 1−10.
(18) Kuncheva, L. I.; Bezdek, J. C.; Duin, R. P. W. *Pattern Recognit.* **2001**, *34*, 299−314.
(19) Gunes, V.; Ménard, M.; Loonis, P. *Int. J. Patt. Recognit. Artif. Intel.* **2003**, *17*, 1303−1324.
(20) Rokach, L. *Artif. Intell. Rev.* **2010**, *33*, 1−39.
(21) Yang, P.; Yang, Y. H.; Zhou, B. B.; Zomaya, A. Y. *Curr. Bioinf.* **2010**, *5*, 296−308.
(22) Polikar, R. In *Ensemble Machine Learning: Methods and Applications*; Zhang, C., Ma, Y., Eds.; Springer Verlag: 2012.
(23) Zhou, Z.-H. *Ensemble methods. Foundations and algorithms*; Machine Learning & Pattern Recognition Series; CRC Press, Taylor & Francis Group: Boca Raton, FL, 2012.
(24) Breiman, L. *Machine Learning* **2001**, *45*, 5−32.
(25) Bahler, D.; Navarro, L. *Methods for combining heterogeneous sets of classifiers*; Proceedings of the 17th National Conference on Artificial Intelligence, Workshop on New Research Problems for Machine Learning; AAAI Press: 2000.
(26) Buxton, B. F.; Langdon, W. B.; Barrett, S. J. *Meas. Control* **2001**, *34*, 229−234.
(27) Langdon, W. B.; Barrett, S. J.; Buxton, B. F. *Genetic Programming*; Springer: Berlin, Heidelberg, 2002; pp 60−70.
(28) Tsoumakas, G.; Angelis, L.; Vlahavas, I. *Intell. Data Analysis* **2005**, *9*, 511−525.
(29) Bian, S.; Wang, W. *Investigation on Diversity in Homogeneous and Heterogeneous Ensembles*; IEEE: 2006; pp 3078−3085.
(30) Chandra, A.; Yao, X. *Neurocomputing* **2006**, *69*, 686−700.
(31) de Oliveira, D. F.; Canuto, A. M. P.; de Souto, M. C. P. *Proceed. Int. Joint Conf. Neural Networks*; IEEE: 2009.
(32) Plewczynski, D.; Spieser, S. A. H.; Koch, U. *Comb. Chem. High Throughput Screening* **2009**, *12*, 358−369.
(33) Bender, A.; Mussa, H. Y.; Glen, R. C. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.
(34) Liu, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1823−1828.
(35) Lounkine, E.; Nigsch, F.; Jenkins, J. L.; Glick, M. *J. Chem. Inf. Model.* **2011**, *51*, 3158−3168.
(36) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.
(37) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzoui, K.; Jacoby, E.; Schuffenhauer, A. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.
(38) Geppert, H.; Horváth, T.; Gärtner, T.; Wrobel, S.; Bajorath, J. *J. Chem. Inf. Model.* **2008**, *48*, 742−746.
(39) Chang, C.-Y.; Hsu, M.-T.; Esposito, E. X.; Tseng, Y. J. *J. Chem. Inf. Model.* **2013**, *53*, 958−971.
(40) Simmons, K.; Kinney, J.; Owens, A.; Kleier, D. A.; Bloch, K.; Argentar, D.; Walsh, A.; Vaidyanathan, G. *J. Chem. Inf. Model.* **2008**, *48*, 2196−2206.
(41) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(42) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. *J. Chem. Inf. Model.* **2012**, *52*, 792−803.
(43) Nisius, B.; Bajorath, J. *ChemMedChem* **2009**, *4*, 1859−1863.
(44) Riniker, S.; Landrum, G. *J. Cheminf.* **2013**, *5*, 26.
(45) RDKit: Cheminformatics and Machine Learning Software. 2013. http://www.rdkit.org (accessed November 5, 2013).
(46) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(47) Nilakantan, R.; Baumann, N.; Dixon, J. S.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(48) Landrum, G.; Lewis, R.; Palmer, A.; Stiefl, N.; Vulpetti, A. *J. Cheminf.* **2011**, *3* (Suppl 1), O3.
(49) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(50) Dice, L. R. *Ecology* **1945**, *26*, 297−302.
(51) Pedregosa, F.; et al. *J. Machine Learning Res.* **2011**, *12*, 2825−2830.
(52) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, P. J. *Classification and regression trees*; Wadsworth International Group: Belmont, CA, 1984.
(53) Chen, C.; Liaw, A.; Breiman, L. *Using random forest to learn imbalanced data*; 2004.
(54) Kononenko, I. In *Current Trends in Knowledge Acquisition*; Wielinga, B., Ed.; IOS Press: 1990.
(55) Langley, P.; Iba, W.; Thomas, K. *An analysis of Bayesian classifiers*; Proceedings of the 10th National Conference of Artificial Intelligence; AAAI Press: 1992; pp 223−228.
(56) Pazzani, M. J. In *Learning from Data: Artificial Intelligence and Statistics V*; Fisher, D., Lenz, H. J., Eds.; Springer Verlag: 1996.
(57) Irwin, J. J. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193−199.
(58) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. *J. Cheminf.* **2009**, *1*, 14−37.
(59) DUD LIB VS 1.0. http://dud.docking.org (accessed November 5, 2013).
(60) Rohrer, S. G.; Baumann, K. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.
(61) MUV. http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html (accessed November 5, 2013).
(62) 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. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(63) ChEMBL: European Bioinformatics Institute (EBI). 2012. http://www.ebi.ac.uk/chembl/ (accessed November 5, 2013).
(64) Heikamp, K.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 1831−1839.