JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Using Information from Historical High-Throughput Screens to Predict Active Compounds

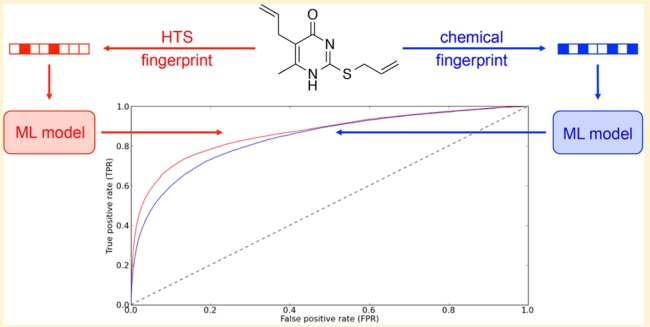
Sereina Riniker,[†,§] Yuan Wang,[‡] Jeremy L. Jenkins,[‡] and Gregory A. Landrum*[,†]

[†]Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4056 Basel, Switzerland

[‡]Novartis Institutes for BioMedical Research Inc., 220 Massachusetts Avenue, Cambridge, Massachussets 02139, United States

Supporting Information

ABSTRACT: Modern high-throughput screening (HTS) is a well-established approach for hit finding in drug discovery that is routinely employed in the pharmaceutical industry to screen more than a million compounds within a few weeks. However, as the industry shifts to more disease-relevant but more complex phenotypic screens, the focus has moved to piloting smaller but smarter chemically/biologically diverse subsets followed by an expansion around hit compounds. One standard method for doing this is to train a machine-learning (ML) model with the chemical fingerprints of the tested subset of molecules and then select the next compounds based on the predictions of this model. An alternative approach would be to take advantage of the wealth of bioactivity information contained in older (full-deck) screens using so-called HTS fingerprints, where each element of the fingerprint corresponds to the outcome of a particular assay, as input to machine-learning algorithms. We constructed HTS fingerprints using two collections of data: 93 in-house assays and 95 publicly available assays from PubChem. For each source, an additional set of 51 and 46 assays, respectively, was collected for testing. Three different ML methods, random forest (RF), logistic regression (LR), and naïve Bayes (NB), were investigated for both the HTS fingerprint and a chemical fingerprint, Morgan2. RF was found to be best suited for learning from HTS fingerprints yielding area under the receiver operating characteristic curve (AUC) values >0.8 for 78% of the internal assays and enrichment factors at 5% (EF(5%)) >10 for 55% of the assays. The RF(HTS-fp) generally outperformed the LR trained with Morgan2, which was the best ML method for the chemical fingerprint, for the majority of assays. In addition, HTS fingerprints were found to retrieve more diverse chemotypes. Combining the two models through heterogeneous classifier fusion led to a similar or better performance than the best individual model for all assays. Further validation using a pair of in-house assays and data from a confirmatory screen—including a prospective set of around 2000 compounds selected based on our approach—confirmed the good performance. Thus, the combination of machine-learning with HTS fingerprints and chemical fingerprints utilizes information from both domains and presents a very promising approach for hit expansion, leading to more hits. The source code used with the public data is provided.



INTRODUCTION

High-throughput screening is a well-established technique in the pharmaceutical industry for lead discovery.[1] State-of-the-art HTS enables the testing of 1−5 million compounds within a few weeks.[2] Much effort has been put into the development of sophisticated compound management, miniaturization and automation,[2,3] but despite these advances other factors such as assay throughput, reagent costs and limited availability of reagents (cells or protein) can present bottlenecks which limit the size of the compound library that can be screened. Different approaches have been proposed for the selection of focused sublibraries which aimed to maximize the chemical and/or biological diversity.[4−8] Diversity was thereby measured either by the number of Bemis−Murcko[9] scaffolds present,[4] by molecular-fingerprint similarity (using e.g. ECFP4[10]),[5,6] by HTS-fingerprint similarity,[7] or by the number of modulated targets/genes.[8]

HTS fingerprints are based on the wealth of information present in historical HTS campaigns using both cellular and biochemical formats, i.e. they describe the bioactivity profile of a compound.[7] Each element or bit in the HTS fingerprint corresponds to the activity outcome in a historical assay. HTS fingerprints can be used like molecular fingerprints for various cheminformatics applications such as subset design, virtual screening and clustering,[7,11] as well as target prediction.[12] As HTS fingerprints describe the biological rather than the chemical similarity between compounds, they were found to have an increased capability to discover novel active chemotypes.[7] Similar approaches have been developed using data on growth inhibition of 60 cancer cell lines,[13,14] a panel of eight proteins,[15] the panels of the CEREP BioPrint database,[16,17] or

Received: March 27, 2014
Published: June 16, 2014

   dx.doi.org/10.1021/ci500190p | J. Chem. Inf. Model. 2014, 54, 1880−1891

the assays of the National Toxicology Program (NTP)[18,19] and were subjected to similarity search,[13,14] ligand binding prediction,[15] SAR analysis,[17] and QSAR for predicting adverse health effects.[18,19]

A limitation of such approaches is that compounds need to have been tested previously in order to have an HTS fingerprint. This poses a problem as new compounds are constantly added to screening collections. Wassermann et al.[20] proposed the combination of chemical similarity with HTS fingerprints to circumvent this problem. The approach is based on the assumption that structurally similar molecules have similar HTS fingerprints, i.e. the unknown HTS fingerprint of a new compound can be substituted by the HTS fingerprints of its neighbors.[20]

HTS fingerprint-based diversity selection[7] was proven more efficient regarding hit rate and chemical diversity of active compounds compared to the molecular fingerprint-based one.[5] Diverse-gene selection in turn, which maximizes the number of targets by the selected compounds based on dose−response data, was found to outperform the HTS fingerprint-based selection.[8]

With the screening results of a sublibrary at hand, different cheminformatics approaches can be utilized to expand the number of compounds and chemotypes around the primary hits such as nearest-neighbor search[21] or machine-learning (ML) models.[22] Glick et al.[22] tested naïve Bayes (NB) and support vector machines (SVM) using the ECFP4 fingerprint and found that both perform well, even on noisy data sets, with NB being computationally more efficient.

Here, we investigate the combination of HTS fingerprints and machine learning for application to hit expansion and compare it to learning with a standard molecular fingerprint, ECFP4, as well as simple similarity search. Three different ML methods, i.e. NB, random forest (RF), and logistic regression (LR), are tested for both fingerprints. The best models are further combined using heterogeneous classifier fusion. The methods are compared over a large set of 51 in-house and 46 public HTS screens which cover a broad range of targets and target classes. Two sets of HTS fingerprints are generated based on 93 in-house and 95 public assays, respectively. For a pair of in-house assays, the performance is further validated using data from a confirmatory screen including a prospective set of around 2000 molecules that were selected for validation based on our approach. The implementation is based on two open-source Python libraries, RDKit[23] and scikit-learn,[24] and the source code used for the public assays is provided.

METHODS AND MATERIALS

General Procedure. The general workflow for building and validating HTS fingerprints is shown schematically in Figure 1. HTS fingerprints were constructed from a set of historical assays. A machine-learning (ML) model was trained using the HTS fingerprints for a set of training molecules, e.g. active and inactive compounds from a new assay, and predictions were generated for the molecules from a compound library (i.e., the test molecules). The test molecules were ranked based on the predicted probability of being active. Ties were broken using the maximum chemical similarity to the training actives (MAX group fusion[25]). From the ranked list, the area under the ROC curve (AUC) and the enrichment factor (EF) at different percentages were calculated. The evaluation methods are described in more detail in ref 26. Other "early-recognition" methods were not considered because our previous study showed that they are strongly correlated with EF.[26] Given the much larger test sets here, we calculated EF at percentages smaller than the 5% used in earlier studies. With 1 million molecules to choose from, the cutoffs at 0.1−1% correspond to 1000−10000 molecules, which is in the range of what is reasonable for a typical confirmatory screen.

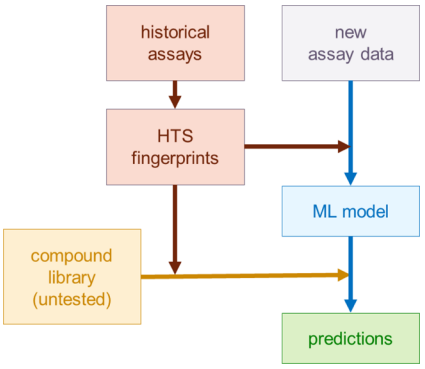


Figure 1. Schematic representation of the workflow for constructing HTS fingerprints from a set of historical assays, training a machine-learning (ML) model with the molecules from a new assay and generating predictions for a set of untested compounds.

The approach was compared to different ML methods trained with chemical fingerprints as well as similarity search using both the chemical fingerprints and the HTS-fingerprints.

The procedure was implemented using a series of Python scripts based on the open-source libraries RDKit[23] (version 2013.09) and scikit-learn[24] (version 0.13). The scripts used for the public data set are available in the Supporting Information.

Chemical Fingerprints. The chemical similarity was calculated using Morgan2 fingerprints (RDKit[23] implementation of the well-known ECFP4 fingerprint[10]). The bit-vector version of Morgan2 was folded to 1024 bits. Similarities were calculated using the Tanimoto coefficient and MAX group fusion[25] was applied for the similarity search. In addition, the unfolded (dictionary) version of Morgan2 fingerprints was used for training of ML models.

HTS Fingerprints. Each bit/element in an HTS fingerprint corresponds to the activity outcome or *Z*-score of the given molecule in one of the historical assays. *Z*-scores were calculated from the measured activity values x using,

$$Z\text{-score}(x_i) = \frac{x_i - \mu(x)}{\sigma(x)} \quad (1)$$

where $\mu(x)$ is the mean of the measured activity values and $\sigma(x)$ is the standard deviation. The function *zscore* from the module stats.mstats of the Python library scipy[27] (version 0.9.0) was used for the calculation. A float and a binary version of HTS fingerprints were constructed. For the float version, an element was set to the corresponding *Z*-score value (positive or negative). For the binary version, a bit was set to one if the *Z*-score was >2.0 or <−2.0 depending on the design of the assay;[28] else, the bit was zero.

Often the matrix of historical assays and compounds is not complete, i.e. not all compounds were tested in all assays. In the almost 1.7 million compounds in the internal data set, the average compound was not tested in 30.2 of the 93 assays (32.4%). The missing bits/*Z*-scores were assumed to be equal to the mean, i.e. zero. The assumption that unknown data is inactive may, of course, introduce false negatives. To investigate the impact of this, we tried building ML models based on

chemical fingerprints for each historical assay and used these to predict the missing bits. This approach can, however, lead to both false positives and false negatives, and we found that HTS fingerprints incorporating such model-generated bits were more problematic (data not shown). Therefore, we decided to use the inactivity assumption for the missing bits.

The similarity between two float-HTS fingerprints was calculated using a generalized Tanimoto coefficient

$$\text{Tanimoto}_{\text{gen}} = \frac{\langle A, B\rangle}{|A^2| + |B^2| - \langle A, B\rangle} \quad (2)$$

where $\langle A, B\rangle$ is the scalar product of the floating-point vectors A and B, and $|X^2|$ is the sum of the squared elements of X, $\sum_i X_i^2$, with $X = A, B$. The AVE group fusion[25] was applied for the similarity search. The MAX group fusion was also tried with the HTS fingerprints, but the resulting performance was much lower (data not shown).

Machine-Learning Methods. Three different ML methods were tested in this study: random forest (RF),[29] naïve Bayes (NB), and logistic regression (LR). A detailed description of the methods is given in ref 30.

A downsampling balancing algorithm[31] was applied for the training of the scikit-learn RF models. The RF parameters used were 100 trees, maximum depth = 10, minimum samples to split = 4, and minimum samples for a leaf = 2. The folded bit-vector Morgan2 fingerprints generated by RDKit were converted to numpy[32] arrays prior to training and testing.

Default parameters were used for the scikit-learn LR classifier. The folded Morgan2 fingerprints from RDKit were converted to numpy arrays prior to training and testing. LR was also trained with unfolded Morgan2 fingerprints from RDKit which were converted to a scipy sparse matrix using the DictVectorizer functionality from scikit-learn prior to training and testing.

For NB, the Bernoulli model of scikit-learn was used with the default parameters. As Bernoulli-NB is designed for binary features, it was only trained with the folded bit-vector fingerprints and the binary version of the HTS fingerprints.

The Laplacian-modified naïve Bayes classifier in PipelinePilot[33] 8.5 (PP) was trained with the unfolded ECFP4.[10] The default parameters were used.

For the rank-based classifier fusion of RF(HTS-fp) and LR(Morgan2), the molecules were first ranked by each classifier separately (with the highest rank being the best), the maximum of the two ranks was kept and the molecules were ordered again. This is the same procedure for fusing model predictions used in ref 30.

Data Sets. *Historical Assays.* The HTS fingerprints were constructed using two sets of historical assays: (i) a collection of 93 in-house in vitro assays (biochemical and cell-based) with ~1.7 million unique compounds, and (ii) a collection of 96 assays from PubChem[34] with 427 880 unique compounds. The in-house collection consists of a broad set of kinases, proteases, ion channels, GPCRs and other target classes, where all modes of action (inhibitor, agonist, antagonist, modulator) are represented.

The public collection consists of the 96 HTS screens from the NIH molecular libraries program (MLP) with more than 338 000 tested compounds. The assays were collected starting with those with the highest number of tested compounds and then taking only assays submitted by NCGC, the Scripps Research Institute Molecular Screening Center, and the Burnham Center for Chemical Genomics. This was done to limit the number of different procedures. The column which had been used to define the PubChem activity outcome was taken to calculate the Z-scores for the HTS fingerprints. The PubChem assay ID (AID), target name, gene id, target class, source, mode of action, number of tested compounds, and number of actives are given in Table S1 in the Supporting Information.

Test Assays. For the in-house assays, we have two kinds of test sets. The first set (test set A) consists of two biochemical screens for the same target using the same assay method. These two assays are not present in the historical set. The small screen (~200 K compounds) had been run before the large one (~1 million compounds). This pair represents a typical use case for our approach. The small screen was used to train the ML model and predictions were generated for the compounds in the large screen. Compounds with Z-score <−3.0 were considered actives. For some of the compounds, the results from a confirmatory screen were available and could be used for further comparison. The second set (test set B) consists of 50 internal in vitro assays (biochemical and cell-based) that were not present in the historical set, with 750 000 to 1.2 million compounds. The targets of the assays were again a broad set of kinases, proteases, ion channels, GPCRs, and other target classes, but different from the targets of the historical assays. For each test assay, 200 K compounds were randomly selected for training, the rest were used for testing. This was repeated ten times for each assay to collect statistics. Depending on the design of the assay, compounds with a Z-score >3.0 or <−3.0 were considered actives.[28] We used a more strict definition here than for the construction of the binary HTS fingerprints in order to reduce the number of false positives, i.e. the noise, for training of the model. Using a cutoff of 2.0 led to a decrease in performance (data not shown).

The public set consists of 46 PubChem[34] assays from the NIH MLP with 300 000−338 000 tested compounds. Each assay is for a different target and the targets are different from those of the historical assays. Each assay in the public set was divided randomly into a 50 000 compound training set and a test set composed of the remaining compounds. This was repeated ten times for each assay to collect statistics. The smaller training set size of 50 000 compared to the 200 000 for the in-house assays was chosen due to the smaller size of the PubChem assays. The activity outcome from PubChem was used to determine the actives and inactives for training and testing. The PubChem assay ID (AID), target name, gene id, target class, source, mode of action, number of tested compounds, and number of actives are given in Table S2 in the Supporting Information.

RESULTS AND DISCUSSION

In-House Test Set A. Different parameters for the HTS-fingerprints approach were investigated using the pair of assays from in-house test set A, and the results were compared to those of machine-learning (ML) models trained with the molecular fingerprint, Morgan2, as well as similarity search with Morgan2 and the HTS fingerprints. Three ML methods were tested: random forest (RF), naïve Bayes (NB), and logistic regression (LR).

Using the maximum folded-Morgan2 similarity to rank the test molecules, an enrichment factor (EF) at 0.1% of 29.78 and an AUC of 0.775 was achieved (Table 1). When the Morgan2 fingerprints were used to train ML models, the AUC and

Table 1. Results Using In-House Test Set A for Morgan2 and Float-HTS Fingerprint Similarity Search and ML Models Trained with Morgan2 (Folded and Unfolded) or HTS Fingerprints (Float or Binary Version), as Well as the PipelinePilot NB Trained with Unfolded ECFP4[a]

method	AUC	EF(0.1%)	EF(0.5%)	EF(1%)	EF(5%)
Morgan2(folded) similarity	0.775	29.78	21.13	16.41	7.17
RF(Morgan2, folded)	0.794	37.32	20.75	15.36	6.90
LR(Morgan2, folded)	0.799	37.87	23.24	17.20	7.46
NB(Morgan2, folded)	0.782	36.96	18.54	13.36	6.46
PP-NB(ECFP4, unfolded)	0.838	22.53	25.16	20.41	9.10
LR(Morgan2, unfolded)	0.860	54.83	37.43	28.24	10.75
HTS-fp(float) similarity	0.825	45.00	28.88	21.32	8.88
RF(HTS-fp, float)	0.864	51.60	35.87	28.41	11.25
LR(HTS-fp, float)	0.817	54.83	37.38	27.45	10.37
RF(HTS-fp, binary)	0.826	50.52	33.29	24.83	9.42
LR(HTS-fp, binary)	0.795	52.53	34.30	24.84	9.23
NB(HTS-fp, binary)	0.801	30.07	17.31	13.56	7.57

[a]The performance was measured using the area under the ROC curve (AUC) and enrichment factors (EF) at 0.1, 0.5, 1, and 5%.

EF(0.1%) values increased, whereas the EFs at 0.5, 1 and 5% remained similar (Table 1). RF and LR performed similarly,

Table 2. Results for RF and LR Models Trained with Float-HTS Fingerprints Where Molecules That Were Active in More than Ten Assays Were Excluded from Training and Testing, and Ranking Based on the Median of the *Z*-Scores, Using In-House Test Set A[a]

method	AUC	EF(0.1%)	EF(0.5%)	EF(1%)	EF(5%)
RF(HTS-fp)	0.855	46.28	33.99	26.45	10.87
LR(HTS-fp)	0.819	52.81	35.71	26.47	10.29
Median(*Z*-score)	0.769	20.60	13.23	11.70	7.33

[a]The performance was measured using AUC and EF values at 0.1, 0.5, 1, and 5%.

whereas NB yielded a lower performance than the other two methods.

The PipelinePilot-NB(ECFP4) model is the standard approach internally for hit expansion.[22] For the current assay pair, it performed better than the scikit-learn ML models (except in EF(0.1%)) (Table 1). However, PipelinePilot uses an unfolded dictionary-based version of ECFP4 whereas scikit-learn models typically take bit vectors as input, i.e. in our case folded Morgan2 fingerprints. While the folding has no influence on similarity search performance,[26] the performance of the LR classifier increased substantially and exceeded that of PP-NB(ECFP4) when LR was trained using the unfolded version of Morgan2 (Table 1). For the other two scikit-learn ML methods, the unfolded version of Morgan2 could either not be

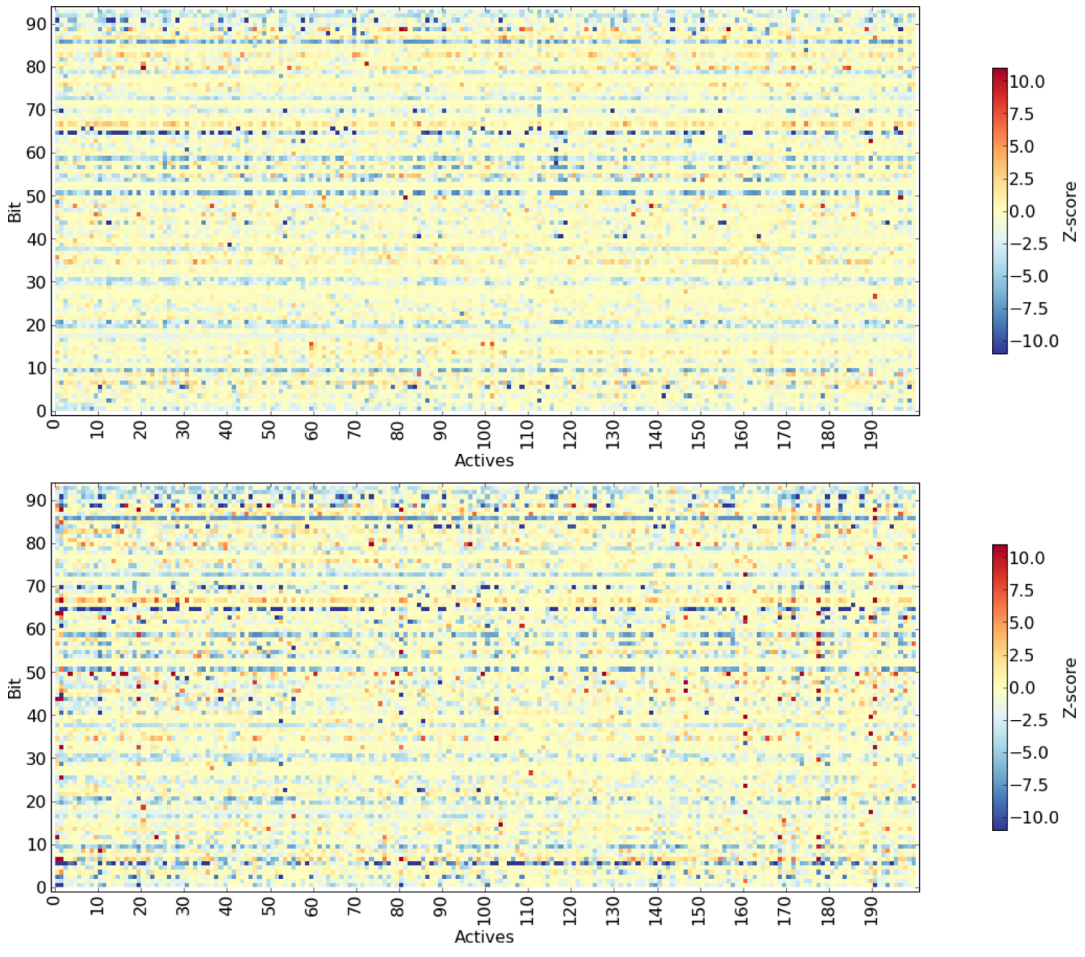


Figure 2. Heatmap of the float HTS fingerprints of the top 200 actives found with RF(HTS-fp) (top) and LR(HTS-fp) (bottom) using in-house test set A. The values >11 were set to 11 and values <−11 to −11 for visualization purposes.

dx.doi.org/10.1021/ci500190p | J. Chem. Inf. Model. 2014, 54, 1880−1891

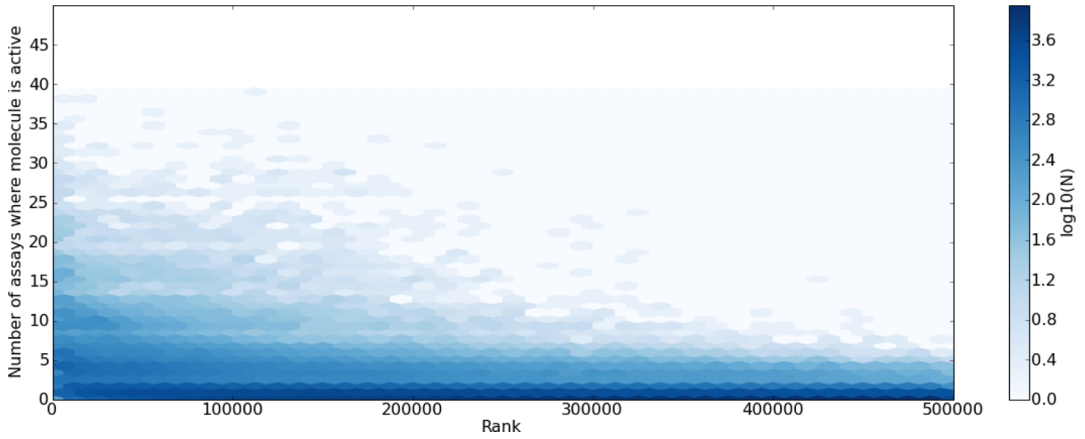


Figure 3. 2D histogram (with hexagonal cells) of the number of historical assays where a compound was tested active as a function of the rank of the compound as predicted by RF(HTS-fps) using in-house test set A. Only the top 500 000 test molecules are shown. The color is representing the number of molecules on a log10 scale.

Table 3. Number of True Actives (N_{act}), Number of Validated Compounds (N_{val}), Percentage of True Actives N_{act} in N_{val}, and the Total Number of Molecules (N_{mol}) Found at Cutoffs of 0.1, 0.5, 1, and 5% of the Ranked Lists by LR(HTS-fp), RF(HTS-fp), and LR(Morgan2) Either Trained with All Primary Actives or with Only Validated Actives (val) from the Small Screen in In-House Test Set A[a]

method	cutoff [%]	N_{act}	N_{val}	N_{act}/N_{val} [%]	N_{mol}
LR(HTS-fp)	0.1	344	779	44.2	1049
RF(HTS-fp)	0.1	368	744	49.5	1049
LR(Morgan2)	0.1	268	782	34.3	1049
RF(HTS-fp, val)	0.1	355	537	66.1	1049
LR(Morgan2, val)	0.1	271	337	80.4	1049
LR(HTS-fp)	0.5	984	2741	35.9	5249
RF(HTS-fp)	0.5	1019	2676	38.1	5249
LR(Morgan2)	0.5	876	2779	31.5	5249
RF(HTS-fp, val)	0.5	934	1702	54.9	5249
LR(Morgan2, val)	0.5	640	980	65.3	5249
LR(HTS-fp)	1.0	1338	4125	32.4	10499
RF(HTS-fp)	1.0	1439	4312	33.4	10499
LR(Morgan2)	1.0	1315	4228	31.1	10499
RF(HTS-fp, val)	1.0	1184	2401	49.3	10499
LR(Morgan2, val)	1.0	899	1536	58.5	10499
LR(HTS-fp)	5.0	2265	8160	27.8	52496
RF(HTS-fp)	5.0	2543	8881	28.6	52496
LR(Morgan2)	5.0	2523	8291	30.4	52496
RF(HTS-fp, val)	5.0	2135	5468	39.1	52496
LR(Morgan2, val)	5.0	1709	3909	43.7	52496

[a]The float version of the HTS fingerprints and the unfolded version of Morgan2 were used.

Table 4. Number of True Actives (N_{act}), Number of Validated Compounds (N_{val}), Percentage of True Actives N_{act} in N_{val}, and the Total Number of Molecules (N_{mol}) Found at Cutoffs of 0.1, 0.5, 1, and 5% of the Ranked Lists by the Fusion of RF(HTS-fp) and LR(Morgan2) Either Trained with All Primary Actives or with Only Validated Actives (val) from the Small Screen in In-House Test Set A[a]

method	cutoff [%]	N_{act}	N_{val}	N_{act}/N_{val} [%]	N_{mol}
Fusion	0.1	347	788	44.0	1049
Fusion	0.5	1086	3055	35.6	5249
Fusion	1.0	1565	4775	32.8	10499
Fusion	5.0	2905	9961	29.2	52496
Fusion(val)	0.1	359	475	75.6	1049
Fusion(val)	0.5	1045	1627	64.2	5249
Fusion(val)	1.0	1418	2474	57.3	10499
Fusion(val)	5.0	2476	5600	44.2	52496

[a]The float version of the HTS fingerprints and the unfolded version of Morgan2 were used.

Table 5. Results for RF(HTS-fp), LR(Morgan2), and the Fusion of RF(HTS-fp) and LR(Morgan2) Trained with the Primary Actives from the Small Screen of In-House Test Set A[a]

method	AUC	EF(0.1%)	EF(0.5%)	EF(1%)	EF(5%)
LR(Morgan2, unfolded)	0.860	54.83	37.43	28.24	10.75
RF(HTS-fp, float)	0.864	51.60	35.87	28.41	11.25
Fusion	0.903	55.19	41.42	31.81	12.75

[a]The float version of the HTS fingerprints and the unfolded version of Morgan2 were used. The performance was measured using AUC and EF values at 0.1, 0.5, 1, and 5%.

used due to technical limitations (RF) or led to a decrease in performance (NB) (data not shown). To put the EF(0.1%) value of 54.83 with LR(Morgan2, unfolded) in context, 764 of the top 1049 molecules (72.8%) were found to be active in the primary assay.

Ranking the test molecules based on the average float-HTS fingerprint similarity led to a higher performance than Morgan2 similarity search (Table 1). Using the HTS fingerprints to train a RF or LR model, the performance increased even further (Table 1). The RF(HTS-fp, float) yielded the highest AUC, EF(1%), and EF(5%) values, whereas LR(HTS-fp, float) gave higher EF(0.1%) and EF(0.5%) values. Thus, with RF(HTS-fp, float) 719 of the top 1049 compounds were primary hits and with LR(HTS-fp, float) 764 out of 1049.

Instead of using the *Z*-scores directly as elements in the HTS fingerprint (i.e., the float version), we also tested a binary version where a bit is set to one if *Z*-score >2.0 or <−2.0 depending on the design of the assay. The performance decreased slightly for both the RF and LR (Table 1), so we used only the float-HTS fingerprints in the following. The comparatively poor performance of NB(HTS-fp, binary) relative to RF(HTS-fp, binary) and LR(HTS-fp, binary) is

Table 6. Number of Proposed True Actives (N^p_{act}), Number of Proposed Compounds (N^p_{val}), Percentage of N^p_{act} in N^p_{val}, and the Total Number of Previously Untested Molecules (N^p_{mol}) Found at Cutoffs of 0.1, 0.5, 1, and 5% of the Ranked Lists by RF(HTS-fp, val), LR(Morgan2, val), and the Fusion of Both Trained with Only Validated Actives (val) from the Small Screen in In-House Test Set A[a]

method	cutoff [%]	N^p_{act}	N^p_{val}	N^p_{act}/N^p_{val} [%]	N^p_{mol}
RF(HTS-fp, val)	0.1	194	225	86.2	512
LR(Morgan2, val)	0.1	143	257	55.6	712
Fusion(val)	0.1	211	302	69.9	574
RF(HTS-fp, val)	0.5	636	764	83.2	3547
LR(Morgan2, val)	0.5	271	555	48.8	4269
Fusion(val)	0.5	604	889	67.9	3622
RF(HTS-fp, val)	1.0	865	1099	78.7	8098
LR(Morgan2, val)	1.0	332	674	49.3	8963
Fusion(val)	1.0	876	1287	68.1	8025
RF(HTS-fp, val)	5.0	1010	1400	72.1	47028
LR(Morgan2, val)	5.0	484	978	49.5	48587
Fusion(val)	1.0	1191	1895	62.8	46896

[a]The float version of the HTS fingerprints and the unfolded version of Morgan2 were used.

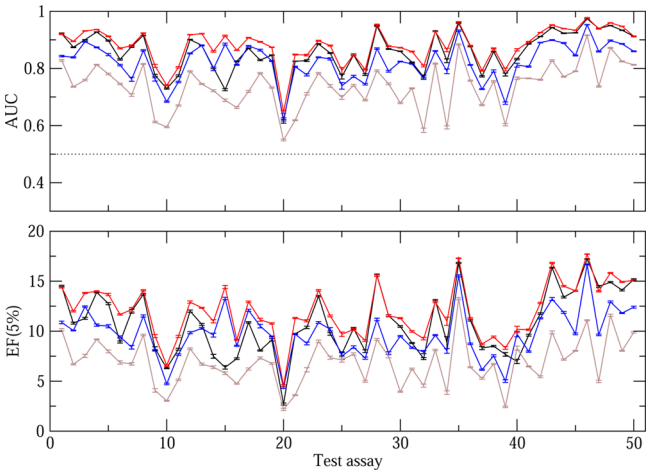


Figure 4. Comparison of average AUC (top) and EF(5%) values (bottom) from RF(HTS-fp, float) (black), LR(Morgan2) (blue), and the fusion of RF(HTS-fp) and LR(Morgan2) (red) and ranking by the Median(*Z*-scores) (brown) over the 50 assays in the in-house test set B. The dotted black line in the top panel indicates random selection. Error bars represent the standard deviation.

likely explained by the assumed independence of features in the NB ansatz. Although LR and NB both learn feature weights, in LR all weights are set together in order to maximize a likelihood function. In NB on the other hand, the weight of each feature is set independently based on its probabilities for each class (for a discussion on the difference between NB and LR see ref 35). In this case, taking the relationships between the bits into account is clearly helpful.

The exceptionally high performance of RF(HTS-fp) and LR(HTS-fp) naturally leads to the question of artifacts or biases. Possible causes could be closely related targets in the historical assays which would lead to a few bits being present in the majority of actives, or the effect of frequent hitters. To investigate these possibilities, we generated heatmaps of the HTS fingerprints for the top 200 actives ranked by RF(HTS-fp) and LR(HTS-fp) (Figure 2). It is clearly visible that some

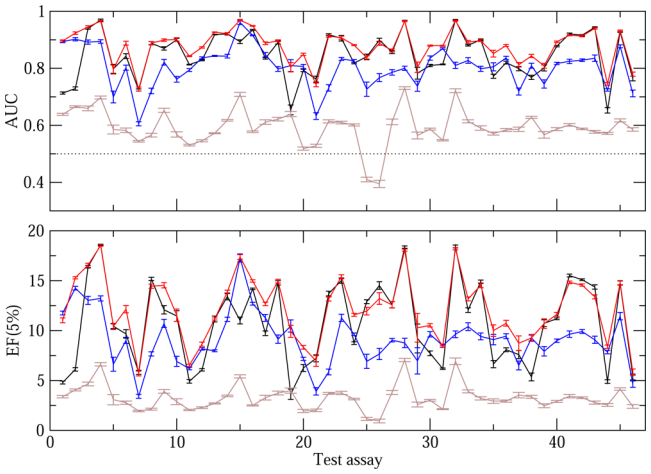


Figure 5. Comparison of average AUC (top) and EF(5%) values (bottom) from RF(HTS-fp, float) (black), LR(Morgan2) (blue), and the fusion of RF(HTS-fp) and LR(Morgan2) (red) and ranking by the Median(*Z*-scores) (brown) over the 46 assays in the PubChem test set. The dotted black line in the top panel indicates random selection. Error bars represent the standard deviation.

bits are set in most of the actives. The *Z*-scores of assay No. 86 for example are <−3.0 in 144 of the 200 actives found by RF(HTS-fp) and in 165 of the 200 actives found by LR(HTS-fp). Other assays appearing in more than 50% of the 200 actives in both models are assay nos. 59 (oxidoreductase), 51 (protein-tyrosine kinase), 65 (*N*-methyltransferase), and 92 (phosphatase). None of these targets are closely related to the glycosyltransferase studied in the current assay, so we can exclude a target-driven bias as a cause for the good performance. However, the HTS fingerprints shown in Figure 2 have rather many darker dots (red or blue), i.e. the corresponding molecules tested active in many assays. Such molecules could be frequent hitters whose "activity" originates from interference with the assay, a reactive group, or toxicity. To investigate the possible bias by frequent hitters, we removed compounds that were active in more than ten of the 93 historical assays from both training and testing. These were ∼1000 training molecules and ∼20 000 test molecules. The resulting performance of RF(HTS-fp) and LR(HTS-fp) was only slightly lower than that of the unfiltered models (Table 2), indicating that frequent hitters were not causing a significant bias. Nevertheless, there was a tendency for HTS fingerprints with many high *Z*-score values (negative or positive depending on the design of the assay) to be ranked in the beginning of the list (Figure 3). To quantify this effect, we ranked the test molecules using the median of their HTS-fingerprint *Z*-scores. For historical assays where actives have *Z*-scores <−3.0, the inverted values (i.e., $Z\text{-score}'_i = -Z\text{-score}_i$) were used for the calculation of the median. The resulting performance was lower than that of both RF(HTS-fp) and Morgan2 similarity search, but it was far from random (Table 2). It represents therefore a kind of baseline performance, to which other approaches should be compared.

As shown in Table 1, the RF(HTS-fp), LR(HTS-fp), and LR(Morgan2) models were able to retrieve a large number of primary hits. However, many of these primary hits may be false positives that turn out to be inactive in a confirmatory screen. We would like to see if our approach can enrich the number of true actives. For both assays in the in-house test set A, we had data from subsequent confirmatory screens available which

dx.doi.org/10.1021/ci500190p | J. Chem. Inf. Model. 2014, 54, 1880−1891

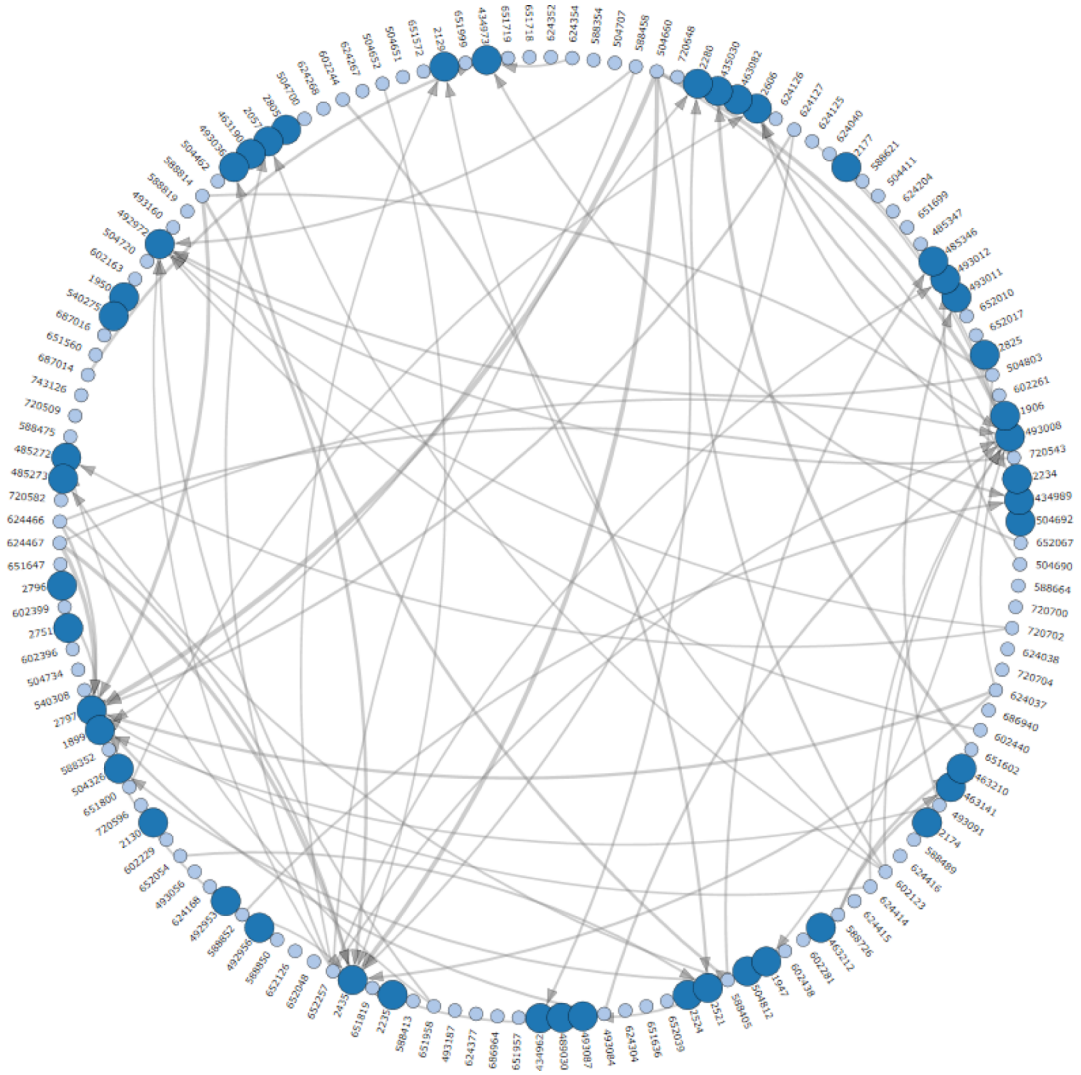


Figure 6. Network of the overlap of actives between the PubChem historical assays (small, light blue circles) and the test assays (big, dark blue circles) for all test assays. The threshold for the directed arrows is 20% overlap of actives and the thickness indicates the amount of overlap (in percent).

allowed us to address this question. Of the ∼1 million compounds in the large screen, 16 265 were tested in a confirmatory screen, 4748 of those were validated as hits (i.e., 29.2%). The numbers of true actives and validated compounds found in the first 0.1, 0.5, 1, and 5% of the compounds ranked by RF(HTS-fp), LR(HTS-fp), and LR(Morgan2) are given in Table 3. Of the molecules within the top *X*% that were not tested in the confirmatory screen, nearly 100% were primary inactives. This is to be expected since primary inactives are rarely selected for confirmatory screens. At 1 and 5% there are 1 or 2 compounds with a *Z*-score >3.0 or <−3.0, depending on the design of the assay, that were not tested. The enrichment of true actives is highest for the smallest cutoff, 0.1%. With RF(HTS-fp) both the precision (N_{act}/N_{val}) and the absolute number of true actives were higher compared to LR(HTS-fp), so the rest of the analysis focuses on RF(HTS-fp). 744 of 1049 compounds found by RF(HTS-fp) were tested in the confirmatory screen, and 368 of these (i.e., 49.5%) were true actives. At a cutoff of 5%, the percentage decreased to 28.6%. With LR(Morgan2) the precision was lower: only 268 of 782 tested molecules were true actives. Thus, although LR-(Morgan2) and RF(HTS-fp) retrieved a similar number of primary hits, the ones found by RF(HTS-fp) were more likely to be confirmed in the follow-up assay.

The percentage of true actives in the validated molecules recovered could be increased by training the RF using only the 420 validated hits from the small screen in the in-house test set A (Table 3). Using RF(HTS-fp, val), a similar number of true actives was found as before with many fewer false positives (i.e., primary actives that were inactive in the confirmatory screen), thus increasing the precision to 66.1%. The same was observed for LR(Morgan2) where the precision increased even more to 80.4%. In absolute numbers RF(HTS-fp) still retrieved more true actives. In summary, this analysis indicates that information about the validated hits can be learned by the ML model that distinguishes them from the other primary actives that were inactive in the confirmatory screen. Therefore, if more reliable dose−response data is available for an HTS screen, this information should be used for training of the ML model and subsequent hit expansion as it will reduce the number of false positives in the top part of the ranked list.

As both RF(HTS-fp) and LR(Morgan2) were found to perform well and they learned from different descriptions of molecules (chemical versus biological), it is desirable to take

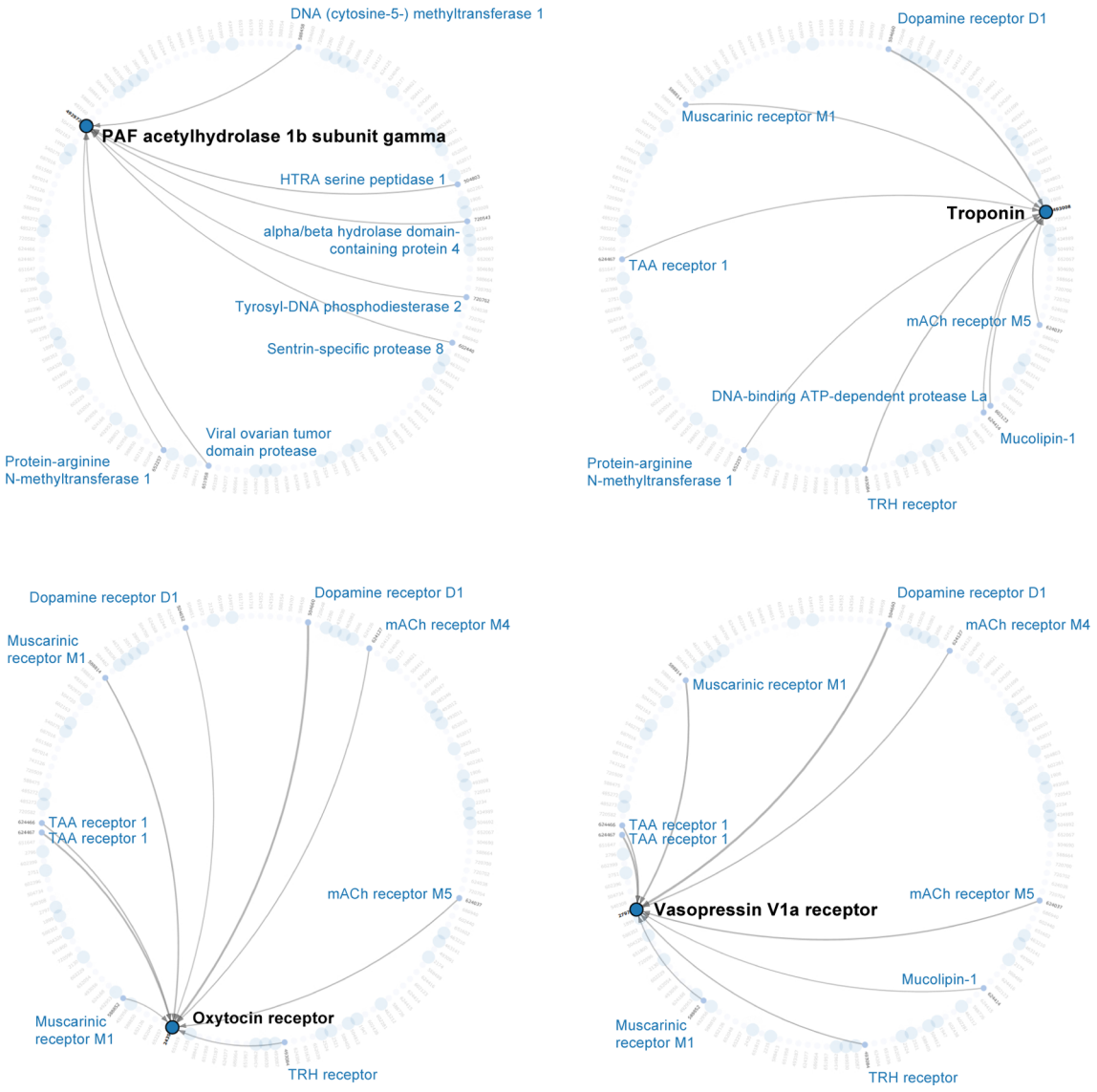


Figure 7. Network of the overlap of actives between the PubChem historical assays (small, light blue circles) and the test assays (big, dark blue circles) for AID 492972, 493008, 2435, and 2797. The threshold for the directed arrows is 20% overlap of actives and the thickness indicates the amount of overlap (in percent).

advantage of both methods. One possibility to do this is to combine them using heterogeneous classifier fusion[30] where the ranked lists of the two models are fused based on the maximum (or minimum) ranks. The results for the fusion of RF(HTS-fp) and LR(Morgan2) are given in Tables 4 and 5. Except for cutoff 0.1% where the performance of the fusion model was between that of the individual models, the absolute number of true actives increased compared to RF(HTS-fp) and LR(Morgan2). The AUC and EF values of the fusion model trained with the primary actives were higher than those of the individual models with a remarkably high AUC of 0.903 and an EF(0.1%) value of 55.19.

537 of the 1049 compounds in the first 0.1% of the ranked list of the RF(HTS-fp, val) model were tested in an confirmatory screen. What about the other 512 compounds? They were inactive in the primary screen and therefore not selected for validation, but some of them may be false negatives. As a prospective test of our approach, we had the opportunity to suggest 2000 additional compounds in a second confirmatory screen. These compounds were selected based on an older version of the fusion approach implemented in PipelinePilot using binary HTS fingerprints (a detailed method description is given in the Supporting Information). The compounds already tested in a confirmatory screen were removed from the ranked list and the top 2188 compounds of the remaining list were selected for validation. Of the 2188 compounds 2015 were available for testing, and 1230 of these tested active (i.e., 61.0%). Table 6 lists how many of the 2015 compounds were in the ranked list of the current model and how many of these were false negatives. For RF(HTS-fp) at 0.1%, these were 225 compounds of which 194 (i.e., precision = 86.2%) were found active in the confirmatory screen. This means that in total 72.1% of the 762 validated molecules in the first 0.1% were true actives. For LR(Morgan2), the precision was generally lower but some of the molecules were different from the ones found by RF(HTS-fp), which led to 211 true actives being retrieved by the fusion model.

In-House Test Set B. The conclusions drawn from the in-house test set A were tested in a next step on a set of 50 assays covering a broad range of targets. The targets were selected

dx.doi.org/10.1021/ci500190p | J. Chem. Inf. Model. 2014, 54, 1880−1891

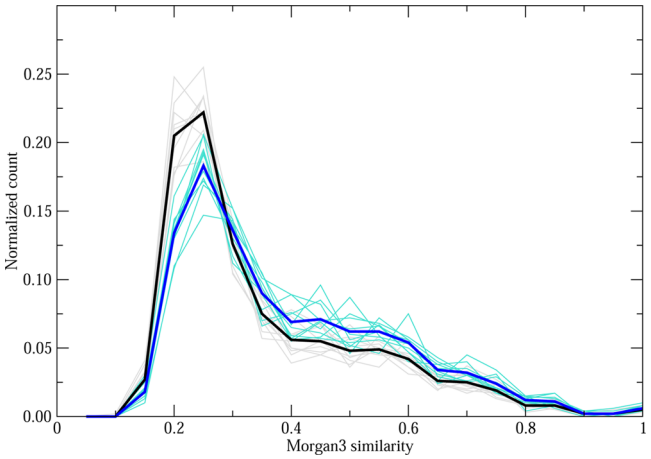


Figure 8. Comparison of the normalized distribution of Morgan3 similarities between the training actives and the actives in the top 5% test molecules of RF(HTS-fp) (black) and LR(Morgan2) (blue) for AID 492972. The ten repetitions are shown for each model.

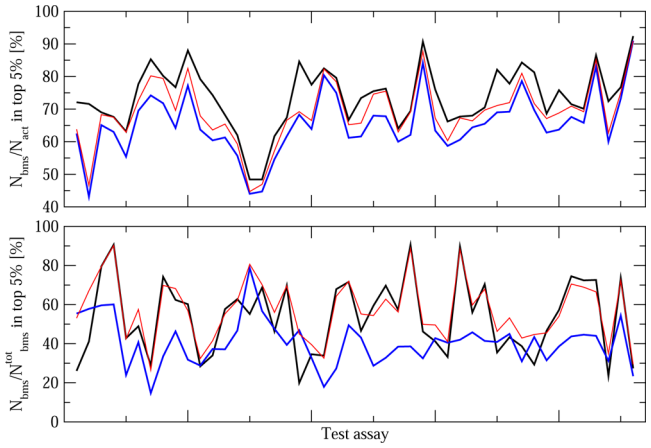


Figure 9. Average ratio of the number of Bemis–Murcko scaffolds (BMS)[9] and the number of actives N_{bms}/N_{act} in the top 5% test molecules (top) and average ratio of N_{bms} in the top 5% and the total number BMS in the test molecules N_{bms}/N^{tot}_{bms} (bottom) of RF(HTS-fp) (black), LR(Morgan2) (blue), and the fusion of both (red).

such that there was no overlap with targets of the historical assays used to construct the HTS fingerprints. For each test assay, the compounds were randomly separated into a training set of 200 000 molecules and a test set. This process was repeated ten times and the results were averaged to remove biases from the random selection. The performance of RF(HTS-fp) using the float version of the HTS fingerprints was compared to that of LR(Morgan2) (Figure 4), as well as that of Morgan2 and HTS-fingerprint similarity search (Figure S1 in the Supporting Information). The numerical values for AUC, EF(0.1%), EF(0.5%), EF(1%), and EF(5%) are given in Tables S3–S6 in the Supporting Information. 80% of the AUC values for RF(HTS-fp) were above 0.8, 94% above 0.75, and only one assay had an AUC value below 0.7. For LR(Morgan2), 68% of the AUC values were above 0.8 and 88% above 0.75. Regarding the enrichment of actives in the beginning of the ranked list, 56% of the assays had an EF(5%) value above 10 with RF(HTS-fp) and 42% with LR(Morgan2). For five assays, EF(5%) was even above 15 with RF(HTS-fp). The RF(HTS-fp) generally outperformed LR(Morgan2), except for four assays in AUC (nos. 15, 17, 18, 20), and 12 assays in EF(5%). The performance of ranking by Median(*Z*-scores) was below that of RF(HTS-fp) and LR(Morgan2) but clearly not random as the AUC values of 10 assays (i.e., 20%) were above 0.8.

The results for the heterogeneous classifier fusion of RF(HTS-fp) and LR(Morgan2) are shown as red lines in Figure 4. The fusion model performed similar to or better than the best individual model for all assays, independent of whether RF(HTS-fp) or LR(Morgan2) was better. In addition, classifier fusion presents an alternative—although conceptually similar—approach to bioturbo similarity search[20] for solving the problem of missing HTS fingerprints for new compounds. As LR(Morgan2) (or any other ML model trained with a molecular fingerprint) can generate predictions for any compounds, such molecules can be incorporated through rank-based fusion, where a molecule simply gets the rank from LR(Morgan2) if it is not present in the ranked list of RF(HTS-fp).

Public Test Set. In addition to the in-house data sets, we collected a set of total 131 assays from PubChem[34] to repeat the same approach. The 95 assays with the most compounds were taken as "historical" assays to construct the HTS fingerprints. For each of the remaining 46 assays, the compounds were randomly separated into a training set of 50 000 molecules and a test set. As before, this process was repeated ten times and the results were averaged. The performances of RF(HTS-fp) and LR(Morgan2) are shown in Figure 5. The comparison to HTS-fingerprint similarity search and Morgan2 similarity search is shown in Figure S2 in the Supporting Information. The numerical values for AUC, EF(1%), and EF(5%) are given in Tables S7–S9 in the Supporting Information. As the public assays were smaller than the internal ones, we omitted the EFs at 0.1 and 0.5%. Again, RF(HTS-fp) yielded generally higher AUC and EF(5%) values compared to LR(Morgan2). The AUC value was above 0.8 in 72% (above 0.75 in 89%) of the assays for RF(HTS-fp), and in 63% (above 0.75 in 76%) for LR(Morgan2). Using the fusion of RF(HTS-fp) and LR(Morgan2), 89% of the assays had an AUC value above 0.8 (93% above 0.75). The simple ranking by the Median(*Z*-scores) yielded a much lower but not random performance with the exception of two assays. For assay No. 19 (AID 2524, intestinal alkaline phosphatase), the performance of RF(HTS-fp) was similar to that of Median(*Z*-scores), indicating that for this assay the model was not able to learn something. The maximum value for EF(5%) is 20. EF(5%) was above 10 in 56% of the assays for RF(HTS-fp), in 28% for LR(Morgan2), and in 74% for the fusion model. For three assays, the EF(5%) was even above 18 for RF(HTS-fp). Again, for assay No. 19 the values for RF(HTS-fp) and median(*Z*-scores) were similar. LR(Morgan2) performed better than RF(HTS-fp) in 14 assays (in at least two of the three evaluation methods AUC, EF1% and EF5%), i.e. assay nos. 1, 2, 12, 15, 17, 19, 20, 24, 30, 31, 35, 36, 38, and 44 (AID 492953, 492956, 463212, 2805, 485273, 2524, 2825, 463082, 2751, 2796, 540275, 2130, 2177, and 1906). The targets of these assays are very diverse, indicating that no particular target class or mode of action was doing better with LR(Morgan2) than RF(HTS-fp).

The assays where RF(HTS-fp) yielded an EF(5%) value >16 were assay nos. 3, 4, 28, and 32 (AID 492972, 493008, 2435, and 2797). Here, the targets are PAF acetylhydrolase 1b (subunit γ), troponin C type 1, oxytocin receptor, and vasopressin V1a receptor. When the overlap of active molecules between the historical assays and the test assays was analyzed (Figures 6 and 7); these four assays were found to share a

minimum of 20% actives with the highest number of historical assays. The oxytocin receptor and vasopressin V1a receptor are both GPCRs, and the targets of the historical assays that share more than 20% actives are also GPCRs (except mucolipin-1 in the case of vasopressin V1a receptor). The other two targets share actives with a more diverse set of targets. There is always a possibility that the overlap in actives between these assays is due to compounds that interfere with the assay readout. We investigated this here by analyzing the assay technologies and could not find any suspicious regularities. The four best-performing test assays mentioned above are all fluorescence based: two of them (AID 492972 and 493008) are biochemical, and the other two (AID 2435 and 2797) are cellular. There is a large number of fluorescence-based biochemical and cellular assays in the historical collection. Some of these do share many actives with the four test assays, but many do not. For example, assay AID 492953 used the same assay technology as AID 492972 (and is for a closely related target) but does not share many actives with any of the historical assays and has a low EF(5%) using RF(HTS-fp) of 4.8. We conclude that interference with the assay readout is not a sufficient explanation for the overlap. In summary, this analysis shows that if information about the active compounds in an assay is present in the historical assays—with similar target and technology or not—the corresponding model is able to retrieve the majority of actives.

Combining RF(HTS-fp) and LR(Morgan2) using heterogeneous classifier fusion led—as for the in-house assays—to a similar or even better performance than the best individual models for all assays, including those where LR(Morgan2) outperformed RF(HTS-fp).

The HTS fingerprints capture a bioactivity profile of a compound and are thus likely to retrieve more diverse chemotypes, i.e. to have a higher scaffold-hopping potential.[36,37] This was confirmed previously for HTS-fingerprint similarity search by Petrone et al.[7] Here, we investigated it in the context of machine learning. The diversity of scaffolds was assessed by calculating the distribution of Morgan3 similarities between the training actives and the actives in the first 5% test molecules, as well as by calculating the number of Bemis−Murcko scaffolds (BMS)[9] in the first 5% of the test molecules. The Morgan3 similarity distributions of RF(HTS-fp) and LR(Morgan2) for the assay AID 492972 are shown in Figure 8. The thin, light-colored lines are the distributions of the individual repetitions, whereas the thick, dark-colored lines are the averages. Using HTS fingerprints, more diverse actives (i.e., similarity < 0.5) were retrieved compared to Morgan2. This can also be seen in Figure 9 where the ratio between the number of BMS and the number of actives in the top 5% of test molecules (N_{bms}/N_{act}) as well as the ratio between the number of BMS in the top 5% and the total number of BMS in the test molecules (N_{bms}/N^{tot}_{bms}) are shown. The numerical values are given in Table S11 in the Supporting Information. For all test assays, N_{bms}/N_{act} was higher for RF(HTS-fp) compared to LR(Morgan2), even for assays where LR(Morgan2) retrieved more actives in the top 5%. For the N_{bms}/N^{tot}_{bms} which is highly correlated with the EF(5%) (Figure S3 in the Supporting Information), LR(Morgan2) performed better than RF(HTS-fp) for the assays where LR(Morgan2) retrieved more actives. These findings confirm and extend previous studies:[7,20] HTS fingerprints lead to more scaffold hopping when used for similarity search or for machine learning.

CONCLUSIONS

HTS fingerprints constructed from historical HTS data describe the bioactivity profile of a compound. The use of these fingerprints for the training of an ML model and subsequent application of the model in hit expansion was investigated using both in-house and public HTS assay collections. In hit expansion, a large library is searched for potential actives based on the data from a (focused) HTS screen. 93 internal and 95 public assays, respectively, were used as historical assays to construct HTS fingerprints. Three different ML methods were tested: NB, RF, and LR. A pair of two in-house screens for the same target was used to investigate the choice of ML method and the choice of HTS fingerprints (float or binary). RF(HTS-fp), LR(HTS-fp), and LR(Morgan2) all performed clearly well with ∼70% of the molecules in the top 0.1% being primary hits. For LR(Morgan2), it was found to be important to use unfolded fingerprints.

ML models trained with the float version of the HTS fingerprints, where each element corresponds to the *Z*-score of the activity value in a historical assay, were found to perform better than models trained with the binary version, where a bit was set to one when the molecules was tested active in the corresponding assay.

Of the three best performing models, RF(HTS-fp) retrieved the highest absolute number of validated hits (i.e., true positives) in the top part of the ranked list with ∼50% of the 744 molecules in the top 0.1% that were tested in a confirmatory screen being true positives. The enrichment of true positives could be increased further to ∼60% when the ML models were trained with only the validated hits as actives. The more accurate data from a confirmatory screen should, when available, be used for training. Heterogeneous classifier fusion allowed the combination of the chemical and biological information contained in LR(Morgan2) and RF(HTS-fp) models. The performance of the fusion of the two was similar or better than the individual models. The fusion approach was also tested in a prospective manner. Around 2000 top ranked molecules which were inactive in the primary screen were proposed for validation in a confirmatory assay, and ∼60% of them were found to be active.

The findings were further tested on a set of 50 internal and a set of 46 public assays covering a broad range of targets. RF(HTS-fp) performed better than LR(Morgan2) for most assays, and the fusion of both models was found similar or better than the best individual model for all assays. In addition, the rank-based fusion allows us to overcome a limitation of the HTS-fingerprint approach and incorporate new molecules which have not been tested in previous screens and therefore do not have an HTS fingerprint.

The scaffold-hopping potential of the HTS fingerprints was assessed and compared to that of Morgan2 using the PubChem assays by calculating similarity distributions between the training actives and the actives in the top 5% of the test molecules as well as the number of Bemis−Murcko scaffolds. Using RF(HTS-fp), more diverse actives were retrieved compared to LR(Morgan2).

In summary, combining experimental information (HTS fingerprints) and chemical information (Morgan2 fingerprints) by fusing heterogeneous machine-learning classifiers is a very promising approach for hit expansion based on HTS data. The method enables us to extract knowledge from orthogonal descriptions using focused pilot screens of limited scale. In the

current study, the size of the training set was predetermined. As a future step, we will investigate the effect of the training-set size on the RF(HTS-fp) and LR(Morgan2) performance.

■ ASSOCIATED CONTENT

Supporting Information

PDF containing the additional figures and tables mentioned in the text and ZIP containing the Python scripts used for the PubChem data set. 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.

Present Address

§(S.R.) Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland.

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 Nikolas Fechner for his help with scikit-learn and insightful discussions on machine-learning and Allen Cornett, Maxim Popov, and Ansgar Schuffenhauser for their help with the in-house HTS assays.

■ REFERENCES

(1) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzend, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* **2011**, *10*, 188−195.

(2) Mayr, L. M.; Bojanic, D. Novel trends in high-throuput screening. *Curr. Opin. Pharmacol.* **2009**, *9*, 580−588.

(3) Battersby, B. J.; Trau, M. Novel miniaturized systems in high-throughput screening. *Trends Biotechnol.* **2002**, *20*, 167−173.

(4) Crisman, T. J.; Jenkins, J. L.; Parker, C. N.; Hill, A. G.; Bender, A.; Feng, Z.; Nettles, J. H.; Davies, J. W.; Glick, M. Plate cherry picking: a novel semi-sequential screening paradigm for cheaper, faster, information-rich compound selection. *J. Biomol. Screen.* **2007**, *12*, 320−327.

(5) Sukuru, S. C. K.; Jenkins, J. L.; Beckwith, R. E. J.; Scheiber, J.; Bender, A.; Mikhailov, D.; Davies, J. W.; Glick, M. Plate-based diversity selection based on empirical HTS data to enhance the number of hits and their chemical diversity. *J. Biomol. Screen.* **2009**, *14*, 690−699.

(6) Bakken, G. A.; Bell, A. S.; Boehm, M.; Everett, J. R.; Gonzales, R.; Hepworth, D.; Klug-McLeod, J. L.; Lanfear, J.; Loesel, J.; Mathias, J.; Wood, T. P. Shaping a screening file for maximal lead discovery efficiency and effectiveness: elimination of molecular redundancy. *J. Chem. Inf. Model.* **2012**, *52*, 2937−2949.

(7) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chem. Biol.* **2012**, *7*, 1399−1409.

(8) Petrone, P. M.; Wassermann, A. M.; Lounkine, E.; Kutchukian, P.; Simms, B.; Jenkins, J. L.; Selzer, P.; Glick, M. Biodiversity of small molecules - a new perspective in screening set selection. *Drug Discovery Today* **2013**, *18*, 674−680.

(9) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(10) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(11) Dančik, V.; Carrel, H.; Bodycombe, N. E.; Seiler, K. P.; Fomina-Yadlin, D.; Kubicek, S. T.; Hartwell, K.; Shamji, A. F.; Wagner, B. K.; Clemons, P. A. Connecting small molecules with similar assay performance profiles leads to new biological hypotheses. *J. Biomol. Screen.* **2014**, *19*, 771−781.

(12) Wassermann, A. M.; Lounkine, E.; Urban, L.; Whitebread, S.; Chen, S.; Hughes, K.; Guo, H.; Kutlina, E.; Fekete, A.; Klumpp, M.; Glick, M. A screening pattern recognition method finds new and divergent targets for drugs and natural products. *ACS Chem. Biol.* **2014**, DOI: 10.1021/cb5001839.

(13) Paul, K. D.; Shoemaker, R. H.; Hodes, L.; Monks, A.; Scudiero, D. A.; Rubinstein, L.; Plowman, J.; Boyd, M. R. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst.* **1989**, *81*, 1088−1092.

(14) Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. Identifying compound-target associations by combining bioactivity profile similarity search and public databases mining. *J. Chem. Inf. Model.* **2011**, *51*, 2440−2448.

(15) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.

(16) Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon, J. C. Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug. Discovery Devel.* **2003**, *6*, 470−480.

(17) Fliri, A.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261−266.

(18) Zhu, H.; Rusyn, I.; Richard, A.; Tropsha, A. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ. Health Perspect.* **2008**, *116*, 506−513.

(19) Zhu, H.; Rusyn, I.; Richard, A.; Tropsha, A. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ. Health Perspect.* **2011**, *116*, 506−513.

(20) Wassermann, A. M.; Lounkine, E.; Glick, M. Bioturbo similarity searching: combining chemical and biological similarity to discover structurally diverse bioactive molecules. *J. Chem. Inf. Model.* **2013**, *53*, 692−703.

(21) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-directed nearest-neighbor searching. *J. Med. Chem.* **2005**, *48*, 240−248.

(22) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. *J. Chem. Inf. Model.* **2006**, *46*, 193−200.

(23) *RDKit: Cheminformatics and Machine Learning Software*; 2013; http://www.rdkit.org.

(24) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(25) Willett, P. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR Comb. Sci.* **2006**, *25*, 1143−1152.

(26) Riniker, S.; Landrum, G. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminf.* **2013**, *5*, 26.

(27) *SciPy: Open-source software for mathematics, science, and engineering*, version 0.9.0; 2011; http://www.scipy.org.

(28) Zhang, X. D. Illustration of SSMD, z score, SSMD*,z* score, and *t* statistic for hit selection in RNAi high-throughput screens. *J. Biomol. Screen.* **2011**, *16*, 775−785.

(29) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5−32.

(30) Riniker, S.; Fechner, N.; Landrum, G. Heterogeneous Classifier Fusion for Ligand-Based Virtual Screening: Or, How Decision Making

by Committee Can Be a Good Thing. *J. Chem. Inf. Model.* **2013**, *53*, 2829−2836.
(31) Chen, C.; Liaw, A.; Breiman, L. *Using random forest to learn imbalanced data*; 2004; http://statistics.berkeley.edy/sites/default/files/tech-reports/666.pdf.
(32) *NumPy: Fundamental package for scientific computing with Python*, version 1.7.1; 2013; http://www.numpy.org.
(33) *PipelinePilot*, version 8.5; Accelrys Software Inc.: San Diego, CA.
(34) *PubChem: National Center for Biotechnology Information (NCBI)*. http://pubchem.ncbi.nlm.nih.gov (accessed March, 26, 2014).
(35) Ng, A.; Jordan, M. I. On discriminative vs. generative classifiers: a comparision of logistic regression and naïve Bayes. *Adv. Neur. Inf. Process. Syst.* **2002**, *2*, 841−848.
(36) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? *J. Med. Chem.* **2010**, *53*, 5707−5715.
(37) Gardiner, E. J.; Holliday, J. D.; O'Dowd, C.; Willett, P. Effectiveness of 2D fingerprints for scaffold hopping. *Future Med. Chem.* **2011**, *3*, 405−411.