Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

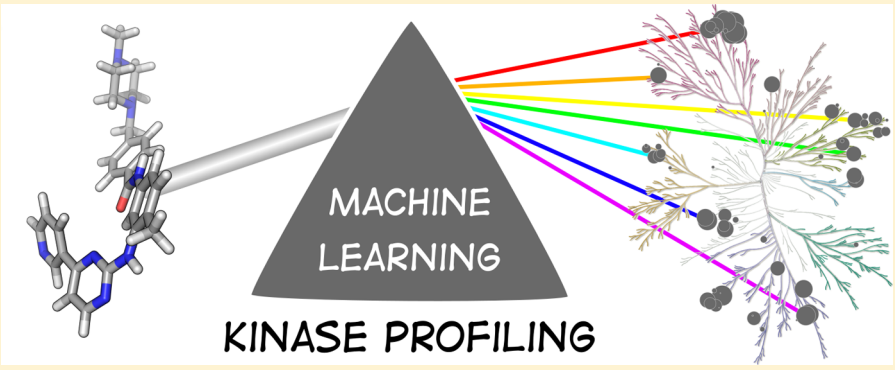
Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay

Benjamin Merget,[†] Samo Turk,[†] Sameh Eid,[†] Friedrich Rippmann,[‡] and Simone Fulle*[,†]

[†]BioMed X Innovation Center, Im Neuenheimer Feld 515, 69120 Heidelberg, Germany

[‡]Global Computational Chemistry, Merck KGaA, Frankfurter Strasse 250, 64293 Darmstadt, Germany

Supporting Information



ABSTRACT: Kinome-wide screening would have the advantage of providing structure−activity relationships against hundreds of targets simultaneously. Here, we report the generation of ligand-based activity prediction models for over 280 kinases by employing Machine Learning methods on an extensive data set of proprietary bioactivity data combined with open data. High quality (AUC > 0.7) was achieved for ∼200 kinases by (1) combining open with proprietary data, (2) choosing Random Forest over alternative tested Machine Learning methods, and (3) balancing the training data sets. Tests on left-out and external data indicate a high value for virtual screening projects. Importantly, the derived models are evenly distributed across the kinome tree, allowing reliable profiling prediction for all kinase branches. The prediction quality was further improved by employing experimental bioactivity fingerprints of a small kinase subset. Overall, the generated models can support various hit identification tasks, including virtual screening, compound repurposing, and the detection of potential off-targets.

INTRODUCTION

Protein kinases have been the focus of drug discovery efforts for many years due to their central roles in signaling pathways involved in the formation and progression of human cancer, inflammation, and Alzheimer's disease.[1,2] Until February 2016, 30 small molecules targeting kinases were approved by the FDA with many potential compounds still in clinical trials.[3,4] The majority of kinase inhibitors bind to the highly conserved ATP-binding pocket, leading to low selectivity, which can easily translate into unwanted side effects.[5] Thus, having an understanding about the binding profile of kinase inhibitors is a prerequisite for drug discovery efforts. Sequence-based phylogenetic relationships of kinases do not always allow extrapolation to the bioactivity space,[6] underlining the need for kinome-wide profiling data for cross-reactivity estimation. Although experimental profiling of compounds against a large fraction of the kinome is experimentally feasible, it is too expensive to be done on a regular basis for hundreds of compounds even for big pharma companies. Extensive panels of compounds tested against many different kinases exist, containing both positive and, importantly, also negative results.[5,7−10] Together with other open resources, like ChEMBL,[11] a wealth of bioactivity measurements is freely accessible. These sources provide valuable training data for computational activity prediction models or *virtual assays*, which are of great importance for hit identification, compound repurposing and off-target detection.

Several attempts to predict binding profiles or free-energy differences exist (including classical QSAR models and free energy calculation tools) and can be feasible for drug design projects (e.g., with respect to accuracy and project time lines).[12−14] The similarity ensemble approach (SEA) is based on the chemical similarity of query ligands to known inhibitors and resulted in successful predictions of unanticipated cross-reactivity.[15,16] Aside from predictors based on chemical similarity, Machine Learning (ML) algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), and Neural Networks comprise a popular toolbox for ligand-centric activity and selectivity prediction.[7,14,17−22] For instance, Yabuuchi and colleagues successfully employed SVM to detect experimentally confirmed inhibitors of GPCR and kinase targets.[20] Furthermore, NB classification and regression yielded kinase activity models with high predictive power.[14,22] Neural Networks employed by Manallack and colleagues achieved 79% correct classifications on an external test set of 120 kinase inhibitors.[17]

Received: October 31, 2016
Published: November 18, 2016

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485

Deep Learning networks showed an average AUC of 0.83 on ChEMBL data, outperforming competing methods, such as SVM, K-Nearest Neighbor (KNN), NB, and also SEA.[19] Beside these ML algorithms, Random Forest (RF) is a popular method to solve classification and regression problems using an ensemble of decision trees and is able to yield very good performance in QSAR modeling even without careful feature selection and extensive parameter tuning.[21,23,24]

Extending a traditional target-centric screen to multiple targets (e.g., kinome-wide) has the advantage of providing multidimensional structure−activity relationships against hundreds of targets simultaneously.[25] Several studies address this issue using a proteochemometric (PCM) approach and, thus, combine chemical information with biological data about the target.[26] Although PCM holds promises, this survey will primarily focus on assessing the predictive power of activity prediction models derived solely from compound fingerprints (FP) and experimental activities. Accordingly, we will address the following questions: (1) which ML algorithm is best suited for generating high-quality (HQ) activity prediction models for a large kinase panel, (2) how does the composition of the used data set and the inherent chemical diversity influence the predictive quality in internal cross-validations and external testing, (3) what is the most suited strategy to balance the data in preprocessing and thus address the class imbalance problem of ML,[27] and (4) can ML be employed for accurate selectivity and off-target prediction? In doing so, different ML methodologies and data balancing schemes are evaluated to yield kinase-specific models with high accuracy. A combination of encoding the chemical (*Morgan fingerprints*) and biological space (*bioactivity fingerprints*) resulted in the best activity prediction and allowed accurate inference of compound selectivity.

MATERIAL AND METHODS

Data Sets. Three data sets were employed for the generation of activity prediction models, hereinafter termed *Proprietary*, *Open*, and *Combined*. (1) The data set *Proprietary* is composed from an in-house profiling panel from Merck KGaA with 4,712 compounds, 220 kinases, and 1,035,549 data points in the form of pIC_{50} values (~100% coverage).[28] (2) The second data set (*Open*) contains the Tang set[10] (which is a collection of a the kinase profiling data sets of Metz,[5] Davis,[8] and Anastassiadis[9]), PKIS,[29−31] and a curated ChEMBL kinase inhibitor panel.[32] The *Open* set provides a rich source of kinase inhibitor data accessible to the public domain and can be obtained from https://github.com/Team-SKI/Publications. After initial filtering, the Tang data set comprises 1,356 compounds, 188 kinases, and a total of 120,194 data points (~50% coverage). The KIBA scores introduced by Tang were converted to the negative $\log_{10}$ of the molar concentration to make the values comparable to the remaining data sets. The *Open* set was then extended by the *Published Kinase Inhibitor Set* (PKIS) of GlaxoSmithKline, containing 366 compounds, 195 kinases, and a total of 71,369 data points in the form of pIC_{50} values (100% coverage).[29,30] In the last step, the *Open* set was extended by a curated in-house databank of kinase inhibitor data from ChEMBL 21.[11] This sparse panel added 38,988 compounds, 314 kinases, and 64,157 measurements. (3) The third data set is a *Combined* master table of the data sets *Proprietary* and *Open*. Overlaps in compounds and kinases from the different sources are shown in Figure 1. Only one representative of duplicate compounds was kept according to the following priority: *Proprietary*, Tang, PKIS, in-house ChEMBL. Furthermore, we only considered data sets with at least 50 bioactivity values per kinase (Table 1). Although it is known that bioactivity values derived from different experimental designs do not always agree and the ChEMBL database can be error prone, it is still feasible to combine these data sets for large-scale ML predictions.[33] The chemical space of the used data sets was subsequently analyzed by means of a Principal Component Analysis (PCA) using the statistical framework R and associated plug-ins.[34−37] Two different types of FPs were calculated for all compounds using the RDKit[38] (version 2015.09.2) implemented connectivity- and feature-based Morgan FPs[39] (ECFP-like and FCFP-like, respectively) with an array length of 4,096 bits and a radius of 4 each. The two FPs were concatenated for each compound. In contrast to classical binary FPs, all calculated FPs were count-based, where each feature is an integer number indicating how often a substructure appears in a molecule. The pIC_{50} cutoff to compose the active (positive) and inactive (negative) classes for each kinase was set to 6.3, which corresponds to a concentration of 500 nM. This cutoff was also used for the generation of binary bioactivity fingerprints, where a compound is described by the experimental bioactivity against a selected subset of kinases. For evaluation, an activity threshold of pIC_{50} = 6 (corresponding to 1 μM) was also tested.

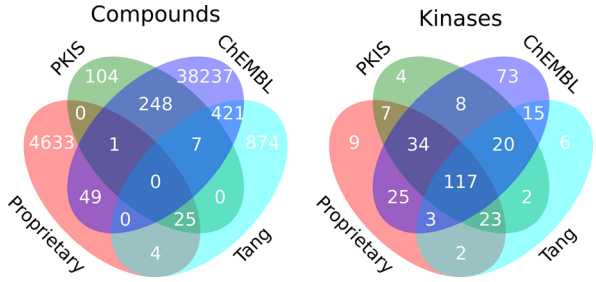


Figure 1. Venn diagram of compounds (left) and kinases (right) from different sources. Whereas only few kinases are unique for a single data source, many compounds only appear in one source.

Table 1. Sizes of Final Data Sets Used for Creating Activity Prediction Models

	compounds	kinases	data points
Proprietary	4,712	220	1,035,549
Open	39,970	263	247,739
Combined	44,603	291	1,280,016

The RF classifier of the Python library Scikit-learn (version 0.17.1) was mainly used for the generation of activity classification models.[40] The number of estimators (decision trees) was set to 2500 and the maximum number of features to the $\log_2$ of the total number of features. To assess the performance of the RF models over other classification techniques, Scikit-learn's NB classifiers (Gaussian NB and Bernoulli NB[41] using default parameters) and a Tensorflow Deep Neural Network (DNN) classifier were used (version 0.7.1 using skflow version 0.1.0).[42] For DNN calculations, a Deep Learning network with two hidden layers of 2048 neurons each, rectified linear units,[43] and stochastic gradient descent for weight optimization were used. The input layer was preprocessed using normalization and application of a tanh function. Furthermore, neuron Dropout at a threshold of 0.5 was introduced to avoid overfitting.[44] DNN parameters were selected by starting from values reported in the literature[19] and optimizing on ABL1 as an exemplary kinase. A simple K-Nearest Neighbor classification serves as a baseline model. All ML calculations were performed in Python, making further use of the packages NumPy (version 1.10.4) and Pandas (version 0.17.1). An exemplary jupyter notebook is provided at https://github.com/Team-SKI/Publications.

Balancing Methods. In kinase profiling data, active compounds are heavily outnumbered by inactive compounds. This is referred to as the class imbalance problem, which can lead to poor performance of ML algorithms.[27] Besides *random* under- and oversampling of the majority/minority class, undersampling was performed based on centroid clustering, Nearest Neighbor (NearMiss[45]) search and *PCA-Centroids*, while oversampling of the minority class was performed by the SMOTE-algorithm (Supporting Information).[46]

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485

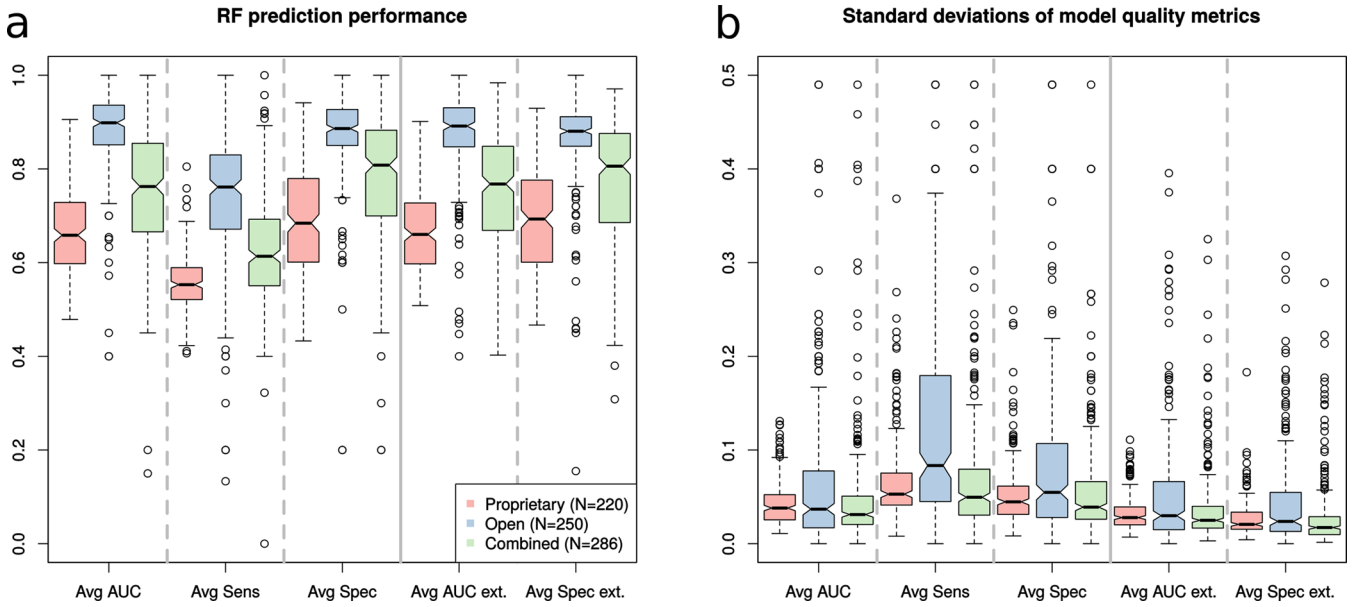


Figure 2. Boxplots of (a) average Random Forest model quality metrics and (b) standard deviations of 5-fold cross-validations (CV) on various data sets. Black lines depict the median, and boxes illustrate the interquartile range (IQR) of the distribution. Whiskers extent to 1.5·*IQR* from the median. Although models based on *Open* data show higher average AUC, sensitivity, and specificity, the standard deviations of the 5-fold CV are significantly higher compared to the results obtained based on the *Proprietary* and *Combined* data sets. Hence, models derived from the latter two data sets are more robust.

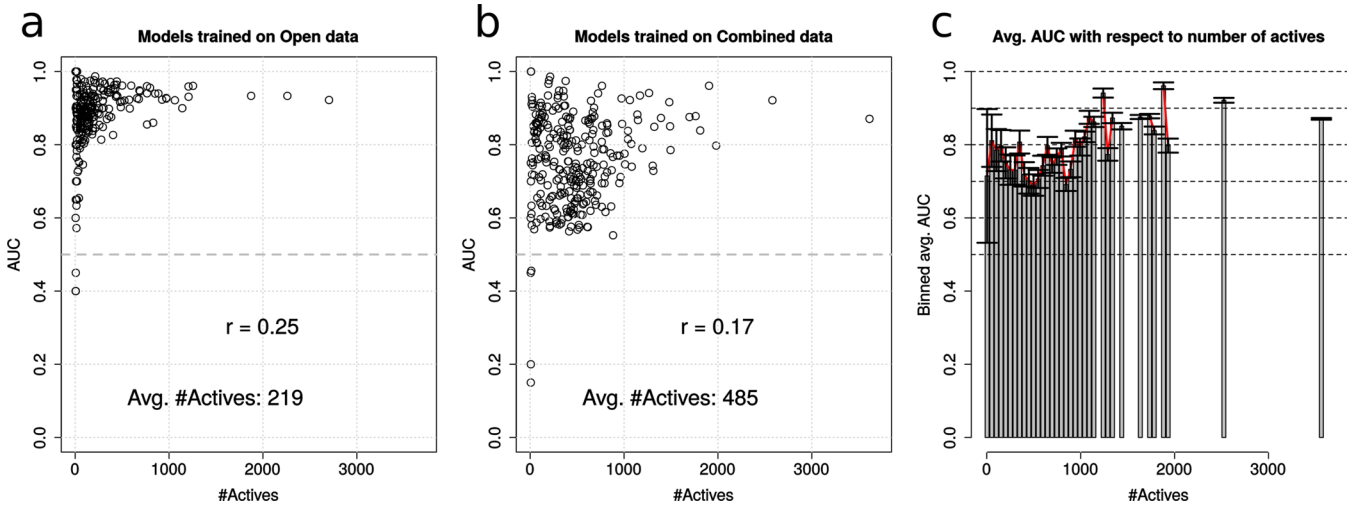


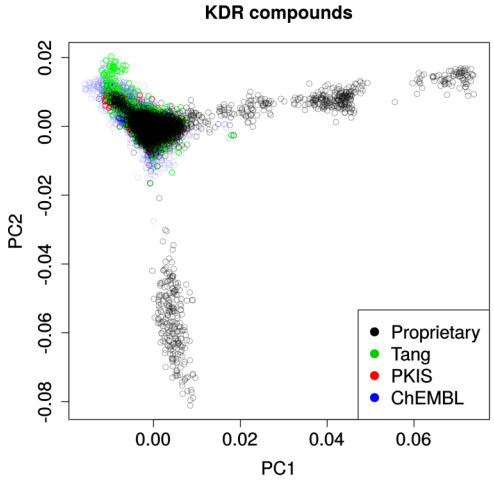
Figure 3. AUC plotted against the number of active compounds per kinase for the (a) *Open* and (b) *Combined* data sets. In both cases, no correlation between the achieved AUC and number of actives exists. (c) Binned average AUC against number of active compounds. Bins have a range of 50 actives. Models with a high number of actives generally result in prediction models with high AUC values.

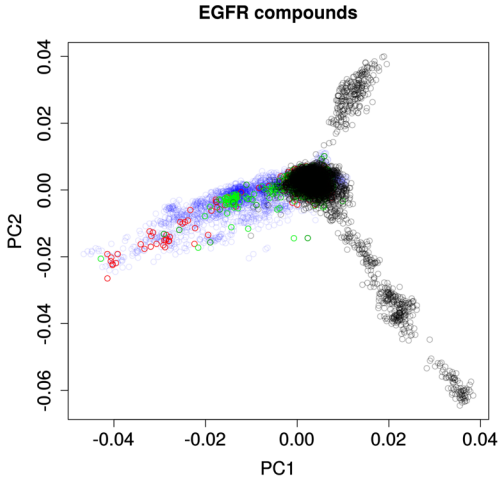
Cross-Validation and Model Evaluation. The balanced data sets were split into five equal parts for a 5-fold cross-validation (CV). Thus, in each fold, a model was trained on 80% and tested on the remaining 20% of the data. Additionally, 10 *external* (left-out) data sets were composed for every single fold, each containing the active compounds from the respective test set combined with randomly drawn inactives from the compounds dismissed in the preceding undersampling step. For model quality assessment, the average Areas Under the ROC-Curve (AUC), the average sensitivities (recall of positive class) and the average specificities (recall of negative class) of the 5-fold CVs were evaluated. Furthermore, the average AUCs and specificities of the *external* test sets were calculated. There is no need to calculate the sensitivity for these data sets because the positive compounds are identical to those of the CV test sets. Paired Mann−Whitney U tests were performed to assess statistical significance.

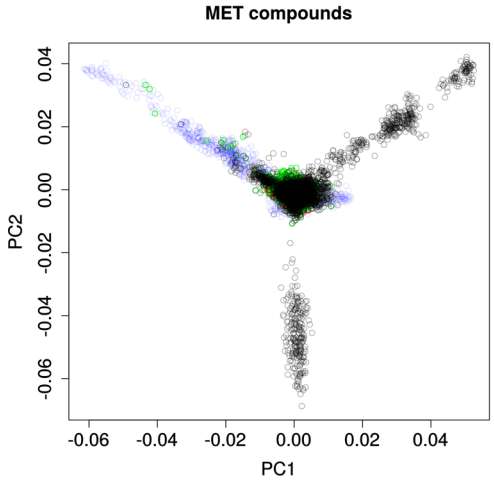
RESULTS

This section will be structured as follows: first, the effect of different data sources (*Proprietary* and/or *Open*) on the predictive power of RF models will be assessed. An analysis of the chemical space of the data sets will serve as an explanation for the dependency of the model applicability on the chosen data source. Moreover, the distribution of high-quality (HQ) models across the kinome phylogeny will be analyzed. Second, alternative ML classifiers will be trained and the predictive power evaluated. In an analogous fashion, various data balancing techniques will be used and evaluated in comparison to the results obtained by random undersampling. Third, the training data will be extended by bioactivity fingerprints and the corresponding RF classifiers evaluated. In

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485







AUC of cross-data set validation

	Open trained	Proprietary trained
KDR	0.56	**0.66**
EGFR	0.65	**0.83**
MET	0.58	**0.80**

Figure 4. First two principal components of PCA of compounds from the *Combined* data set with measurements against kinases KDR, EGFR, and MET, respectively. Connectivity-based Morgan FPs were used as input. Besides the majority cluster, the *Proprietary* data (from Merck KGaA) occupies several further distinct clusters in the chemical space.

the last section, the applicability of the various presented methods for predicting selectivity scores will be analyzed.

Kinase Activity Classification Models. *Activity Prediction with Random Forest Classifier Yields High-Quality Models.* The Random Forest (RF) methodology was used on the Morgan FPs to create a dedicated activity prediction model for each kinase after balancing the data sets using random undersampling. Random undersampling is the easiest, most intuitive method for data balancing but nevertheless yields very good performance.[27,47] The RF models yield reasonable prediction results on the *Proprietary* data set with an average AUC of 0.67 ± 0.09, a sensitivity of 0.56 ± 0.06, and a specificity of 0.69 ± 0.11 (Figure 2a). Surprisingly, models trained on *Open* data exhibit a much higher average predictive power with an average AUC of 0.88 ± 0.09. However, the much higher standard deviations of the 5-fold CVs indicate that these results are largely affected by the random splits of the underlying training and test data (Figure 2b). Therefore, the models trained with *Open* data are much less robust and can be expected to generally have less prediction power. The *Combined* data set shows the best results regarding average model quality metrics (AUC of 0.76 ± 0.12, sensitivity of 0.63 ± 0.12, and specificity of 0.78 ± 0.13) and simultaneously very high robustness (i.e., low standard deviations in Figure 2b). It is notable that only 0.7% of all models show a drop in AUC > 0.05 on left-out data, making the models trained on *Combined* data well applicable on new data sets. Overall, 118 models were generated with an average AUC > 0.8 based on the *Combined* data set. As a baseline, the results were compared to a simple Nearest Neighbor (KNN with K = 1) classification on raw, unbalanced data. The resulting models have significantly lower predictive power with an average AUC of 0.66 ± 0.10 ($p \ll 0.001$). Employing a threshold of $pIC_{50} = 6$ did not change the model performance (0.76 ± 0.11), although the average number of actives increased from 485 to 651.

Effect of Training Data on Model Performance. Next, we investigated how the content of the training set, such as the number of actives and the diversity of covered chemical space, affects the prediction power. Interestingly, the obtained AUCs do not correlate with the number of actives (Figures 3a,b). Please note that most models of the *Open* set were trained on a low number of actives (33% have <50 actives), indicating that these models might be overfitted. By including the *Proprietary* kinase panel, the average number of actives increased from 219 to 485 compounds. Although the average AUC values exhibit a slight drop, the models are more robust, as evidenced by the much lower standard deviation in the CV (cf. Figure 2b). As expected, the higher number of active compounds (and associated increased size of the balanced training set) has a positive effect on the model quality. To further test whether the

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485

models derived from the *Open* data are overfitted, the *Proprietary* data set was used for external validation. The performance of the *Open* models drops significantly (average AUC of 0.56 ± 0.06, Supporting Information, Figure S1). On the other hand, models based on *Proprietary* data show reasonable prediction results when externally tested on the *Open* data (average AUC of 0.65 ± 0.13; $p \ll 0.001$ compared to *Open* models; Supporting Information, Figure S1). Although there is no direct correlation between the number of active compounds and the AUC, models with a large number of actives (>1000) usually result into prediction models with very good AUC values above 0.8 (Figure 3c). The large error bar of the first bin in Figure 3c indicates that, regardless of a possibly high AUC, reliable predictions can hardly be derived from models with very small training sets.

Analysis of Chemical Space of Data Sets Explains Importance of Proprietary Data Set. To explain the positive effect of the *Proprietary* data set on model robustness, a PCA was performed on the connectivity-based part of the Morgan FPs of compounds measured against the kinases KDR, EGFR, and MET (Figure 4), respectively. These kinases show drastic differences in AUC when trained on *Proprietary* and evaluated on *Open* data and vice versa ($p \ll 0.001$); furthermore, these three kinases provide very large data sets with overall 10,072, 9,307, and 7,636 measurements, respectively. Distinct clusters are observable in the chemical space of the PC1−PC2 planes. Whereas the majority of the compounds in the *Proprietary*, Tang, and PKIS data accumulate to one big cluster, the *Proprietary* panel additionally occupies several further distinct clusters. Thus, the *Proprietary* data set has a larger chemical diversity in the training and test sets. Resulting models accordingly have a higher chance of being applicable in diverse drug discovery efforts. This indicates that aside from model quality metrics, such as AUC, also the chemical space of the training sets and real test sets should always be assessed. Interestingly, the compounds from the public domain (Tang, PKIS, and ChEMBL) tested against EGFR also show a notable diversity from the main cluster, which might explain the higher AUC of the EGFR model based on *Open* data compared to the KDR and MET models (0.65 vs 0.56 and 0.58, respectively). Analyzing the principal component space to higher dimensions up to PC20 further underlines the chemical diversity in the *Proprietary* panel (data not shown). Worth mentioning is that on average 97% of the compounds are DFG-in binders according to a classification scheme described by Zhao et al.[48] Considering only "DFG-in" binders did not change the prediction performance (both AUC: 0.76 ± 0.12), while only 92 models with AUC ≥ 0.7 could be obtained for "DFG-out" binders. This indicates that DFG-out binders can safely be included when training DFG-in models and that not enough data is in general available to derive dedicated DFG-out models.

High-Quality Models Are Evenly Distributed Across the Kinome. Although a strong traditional research bias shifts kinase drug discovery toward already validated drug targets, such as the tyrosine kinases (TK group),[49,50] driver mutations in kinases in all kinase groups are present in a variety of cancer types.[49] Many of these yet untargeted kinases show high predicted druggability scores and might, thus, provide opportunities for novel, competition-free drug discovery projects.[32] Encouragingly, HQ models, i.e., models with an AUC of ≥0.8, are evenly distributed across the entire kinome (Figure 5).

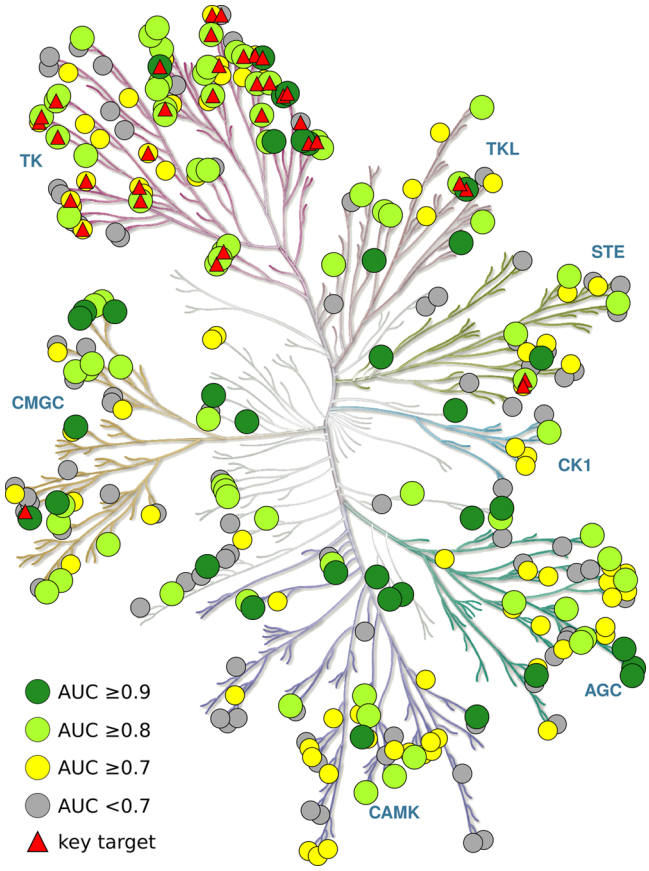


Figure 5. Kinome map of the performance of activity prediction models by Random Forest. Kinases are colored based on their AUC value. High-quality models (AUC ≥ 0.8) are scattered well across the kinome tree and cover almost all kinase families. FDA approved kinase inhibitor targets are depicted as red triangles. Figure was created with KinMap (http://kinhub.org/kinmap).

Table 2 summarizes the 35 best models (AUC ≥ 0.9). Beside AUC, sensitivity and specificity are generally also very high in these models, making them excellent tools for the identification of truly active and inactive compounds. As indicated by the high AUC and specificity on external data sets, the models are well applicable on external data sets for VS for new inhibitors of these kinases. Exceptions might be the models obtained for kinases DCAMKL1, DYRK4, MNK1, NEK9, and TGFbR2, as they are derived from a very small number of active compounds (Table 2).

Random Forest Generally Outperforms Alternative Machine Learning Methods. Model generation was repeated using two NB classifiers (Gaussian and Bernoulli NB), a K-Nearest Neighbor (KNN; with default settings, K = 5) and a Deep Neural Network (DNN) predictor, again employing randomly undersampled data sets. Each method was tested on the *Combined* data set (using the same data preparation as before). Comparison of AUC and specificity, extracted from the 5-fold CV and external data sets, indicates that the RF models are generally superior to the alternative ML methods in our analyses (Figure 6). Only DNN yields similarly good results with an average AUC of 0.76 ± 0.12. Interestingly, the DNN and KNN classifiers achieve higher sensitivity than the RF models, which, however, comes with a decline in specificity, particularly in the case of KNN.

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485

Table 2. Quality Measures of Activity Prediction Models[a] with an AUC ≥ 0.9 of the 5-Fold CV

group	kinase	av AUC[b]	av sens[b]	av spec[b]	av AUC ext[b]	av spec ext[b]	no. actives
AGC	PKCa	0.93 ± 0.01	0.85 ± 0.03	0.86 ± 0.01	0.92 ± 0.01	0.84 ± 0.02	293
	PKCb	0.97 ± 0.02	0.87 ± 0.03	0.94 ± 0.06	0.97 ± 0.02	0.95 ± 0.00	195
	PKCg	0.91 ± 0.03	0.79 ± 0.06	0.93 ± 0.05	0.91 ± 0.02	0.92 ± 0.01	98
	ROCK2	0.93 ± 0.01	0.85 ± 0.02	0.90 ± 0.03	0.93 ± 0.01	0.90 ± 0.01	1169
CAMK	CHK1	0.92 ± 0.01	0.80 ± 0.02	0.96 ± 0.01	0.91 ± 0.01	0.94 ± 0.01	1482
	DCAMKL1	1.00 ± 0.00	0.40 ± 0.49	0.80 ± 0.40	0.88 ± 0.13	0.92 ± 0.06	8
	MNK1	0.92 ± 0.06	0.78 ± 0.22	0.85 ± 0.20	0.92 ± 0.08	0.93 ± 0.03	21
	PIM3	0.91 ± 0.03	0.74 ± 0.05	0.94 ± 0.04	0.91 ± 0.03	0.95 ± 0.01	372
CMGC	DYRK1A	0.91 ± 0.03	0.76 ± 0.06	0.90 ± 0.02	0.89 ± 0.03	0.86 ± 0.01	360
	DYRK1B	0.90 ± 0.03	0.69 ± 0.08	0.89 ± 0.04	0.88 ± 0.02	0.86 ± 0.02	183
	DYRK4	0.93 ± 0.09	0.85 ± 0.20	0.80 ± 0.27	0.88 ± 0.14	0.59 ± 0.22	18
	GSK3B	0.94 ± 0.01	0.82 ± 0.03	0.93 ± 0.01	0.94 ± 0.01	0.91 ± 0.01	1266
	JNK1	0.92 ± 0.01	0.81 ± 0.02	0.94 ± 0.02	0.92 ± 0.01	0.94 ± 0.00	648
	p38a	0.92 ± 0.01	0.84 ± 0.01	0.92 ± 0.01	0.92 ± 0.00	0.92 ± 0.01	2580
TK	DDR2	0.90 ± 0.03	0.70 ± 0.08	0.94 ± 0.03	0.89 ± 0.03	0.93 ± 0.01	273
	EGFR	0.96 ± 0.01	0.89 ± 0.02	0.92 ± 0.01	0.96 ± 0.01	0.92 ± 0.00	1905
	ErbB2	0.96 ± 0.01	0.92 ± 0.03	0.90 ± 0.01	0.96 ± 0.01	0.90 ± 0.01	765
	JAK1	0.95 ± 0.02	0.89 ± 0.02	0.92 ± 0.04	0.95 ± 0.02	0.90 ± 0.02	308
	KIT	0.94 ± 0.02	0.83 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	0.95 ± 0.01	724
	TNK1	0.91 ± 0.10	0.80 ± 0.24	0.97 ± 0.07	0.88 ± 0.09	0.84 ± 0.08	32
	TYK2	0.93 ± 0.02	0.80 ± 0.04	0.93 ± 0.04	0.93 ± 0.02	0.93 ± 0.01	190
TKL	BRAF	0.97 ± 0.01	0.91 ± 0.04	0.91 ± 0.04	0.98 ± 0.01	0.92 ± 0.02	378
	LRRK2	0.93 ± 0.01	0.84 ± 0.06	0.89 ± 0.05	0.93 ± 0.02	0.87 ± 0.01	297
	TGFbR2	0.91 ± 0.11	1.00 ± 0.00	0.80 ± 0.16	0.94 ± 0.06	0.72 ± 0.10	15
STE	COT	0.91 ± 0.05	0.92 ± 0.05	0.78 ± 0.18	0.90 ± 0.03	0.78 ± 0.18	113
	PAK1	0.90 ± 0.06	0.64 ± 0.23	0.96 ± 0.08	0.91 ± 0.04	0.96 ± 0.02	25
other	AurA	0.92 ± 0.02	0.83 ± 0.02	0.88 ± 0.04	0.92 ± 0.01	0.88 ± 0.02	1062
	AurB	0.93 ± 0.02	0.82 ± 0.03	0.89 ± 0.01	0.93 ± 0.01	0.89 ± 0.01	973
	AurC	0.90 ± 0.02	0.73 ± 0.04	0.90 ± 0.01	0.89 ± 0.02	0.89 ± 0.01	680
	CDC7	0.95 ± 0.02	0.87 ± 0.05	0.91 ± 0.02	0.95 ± 0.01	0.91 ± 0.01	353
	NEK9	1.00 ± 0.00	0.70 ± 0.40	0.90 ± 0.20	0.95 ± 0.08	0.95 ± 0.02	10
	PLK1	0.93 ± 0.03	0.81 ± 0.03	0.97 ± 0.01	0.93 ± 0.02	0.97 ± 0.01	358
	PLK3	0.91 ± 0.03	0.75 ± 0.06	0.92 ± 0.04	0.91 ± 0.04	0.95 ± 0.01	125
	TTK	0.99 ± 0.00	0.96 ± 0.02	0.94 ± 0.03	0.98 ± 0.00	0.94 ± 0.01	213
	Wee1	0.98 ± 0.01	0.92 ± 0.06	0.94 ± 0.04	0.98 ± 0.01	0.94 ± 0.04	201

[a]Obtained via Random Forest. [b]Abbreviations: av, average; sens, sensitivity; spec, specificity; ext, external.

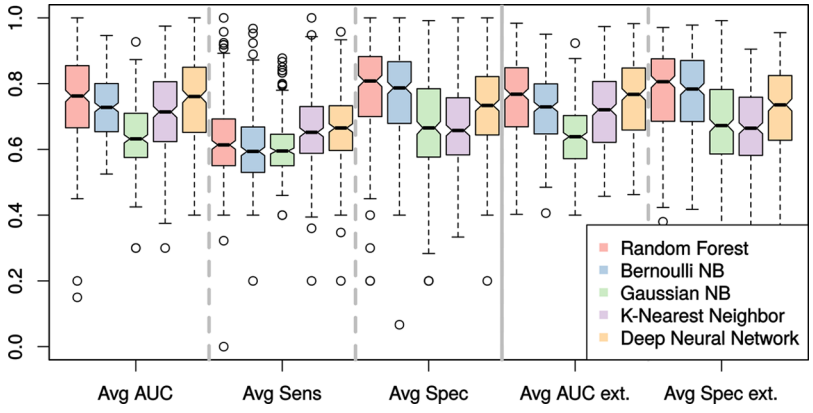


Figure 6. Boxplots of model quality measurements for Random Forest and alternative Machine Learning approaches. The plotted results are based on the *Combined* data set.

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485

The AUC values of the external test sets of each classifier are highly correlated (Supporting Information, Figure S2). However, this does not necessarily mean that an alternative method cannot perform better than RF on a single kinase (Figure 7). Notably, the alternative ML methods only

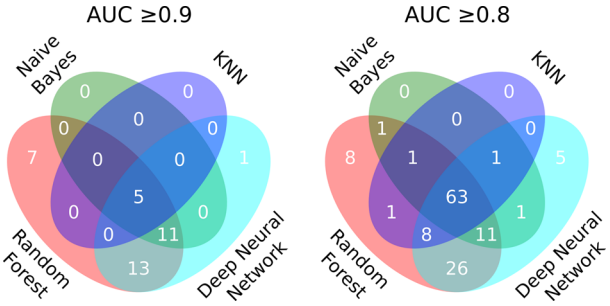


Figure 7. Overlap among various Machine Learning methods of high-quality models with an AUC of ≥0.9 and ≥0.8, respectively.

contribute one HQ models with an AUC ≥ 0.9, which is not also obtained with RF, although the majority of these models are commonly found by at least two of the four approaches. The unique HQ model was obtained by the DNN approach. On the other hand, the RF method is able to create seven unique HQ activity prediction models. Lowering the HQ cutoff to an AUC of 0.8 further confirms the superiority of RF and DNN over the other methods. While RF generates eight unique HQ models, seven additional kinases get HQ models using DNN (Figure 7 and Table 3). In many cases, however, the

Table 3. High-Quality Models Generated by Deep Neural Networks But Not by Random Forest

		AUC	AUC		
	kinase	Deep Neural Networks	Random Forest	ΔAUC	no. actives[a]
AUC ≥ 0.8	BRD2	0.85 ± 0.20	0.15 ± 0.30	0.70	7
	CDK8	0.83 ± 0.07	0.79 ± 0.05	0.04	101
	MLK2	0.81 ± 0.14	0.76 ± 0.13	0.05	20
	PDK1	0.80 ± 0.02	0.79 ± 0.02	0.01	501
	PHKg1	0.95 ± 0.10	0.75 ± 0.39	0.20	7
	PKACb	0.80 ± 0.40	0.60 ± 0.49	0.20	5
	PKR	0.80 ± 0.25	0.70 ± 0.29	0.10	8
AUC ≥ 0.9	PHKg1	0.95 ± 0.10	0.75 ± 0.39	0.20	7

[a]Number of active compounds in data set.

ΔAUC between DNN and RF is very small or the models were derived from only a small number of active compounds (Table 3). Accordingly, in this study, RF was still preferred over DNN, mainly due to lower computational cost of training and the access to feature importance information, which increases model interpretability.

Although RF generally outperforms the other classifiers, it can still be useful to assess and compare the performance of alternative methods. For instance, HQ models can be generated, which might not be accessible by solely using RF. Considering multiple independent classifiers might therefore boost the probability of success in a ligand-based VS endeavor.[51]

The Impact of the Balancing Technique. Besides *random* undersampling, various other techniques for data balancing have been explored (cf. Supporting Information for detailed information). In summary, two fundamentally different approaches for data balancing, undersampling and oversampling, might serve different purposes. Whereas undersampled models might be useful to capture a large number of potentially active molecules (high sensitivity), oversampling seems to be more applicable for detecting true negative compounds (high specificity). The balancing methods which yielded the best results were random undersampling, PCA-Centroids undersampling, and random oversampling (Supporting Information, Figures S3 and S4).

Bioactivity Fingerprints. Compound profiling data contain valuable information on the tested compounds on a variety of targets. Comparing compounds on the basis of their biological profiles instead of chemical similarity can provide complementary information and be a valuable source for hit expansion,[52] repurposing projects,[53] designing screening subsets[54] as well as to identify targets of phenotypic screenings.[55] Here, bioactivity FPs were generated based our *Proprietary* panel, where each bit was set to one if the pIC_{50} was ≥6.3 and to zero otherwise. To determine the 20 kinases with the largest information content for such predictions, we trained an RF regressor to predict the selectivity score *S* based on the entire experimental bioactivity profile of the *Proprietary* data set. *S* is defined as the number of kinases inhibited by the compound divided by the number of all tested kinases:[56]

$$S_C = \frac{\text{no. inhibited kinases}}{\text{no. kinases}} \quad (1)$$

Thus, the lower the value of *S*, the higher the selectivity of a given compound *C*. Importantly, the RF algorithm does not only predict *S* but also provides valuable information about the feature importance, which is a list of kinases in the present case. The identified kinases are spread across the entire kinome and include KDR, RSK2, AMPKa1, MARK2, SIK, IRAK1, CHK2, FGFR2, MELK, HIPK2, JAK2, FGFR3, MARK1, MSK1, CDK7, FGR, AMPKa2, FLT1, CDK5, and HIPK1. Next, bioactivity FPs of these 20 kinases were used for bioactivity prediction, resulting in good model performances with an average AUC of 0.79 ± 0.07 (Figure 8); thus, these models outperform those obtained based on chemical information. Increasing the number of kinases in the bioactivity fingerprint (FP length) further improves prediction results (average AUC for 30 kinases, 0.81 ± 0.07; for 40 kinases, 0.82 ± 0.07). Also, concatenating bioactivity and chemical FPs to one combined input results in excellent activity prediction models for a larger number of kinases (154 models with AUC ≥ 0.8) (Figure 8). The main improvement over using plain Morgan FPs is an increased average sensitivity. Thus, by determining the activity of a small panel of only 20 kinases in combination with chemical FPs, a bioactivity panel can be meaningfully extended to a large number of kinases using an RF classifier. It should be noted that the bioactivity part of the combined FP was processed by a modified implementation of Google's Winner-Takes-All hashing algorithm[57] to adjust the FP length of 20 bioactivity values to the length of the Morgan FP. This step is required to avoid highly imbalanced FP lengths, which lead to biased selection of features in the individual decision trees.

Selectivity Score and Identification of Off-Targets. So far we have investigated the performance quality when it comes to predicting the activity of compounds against a particular

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485

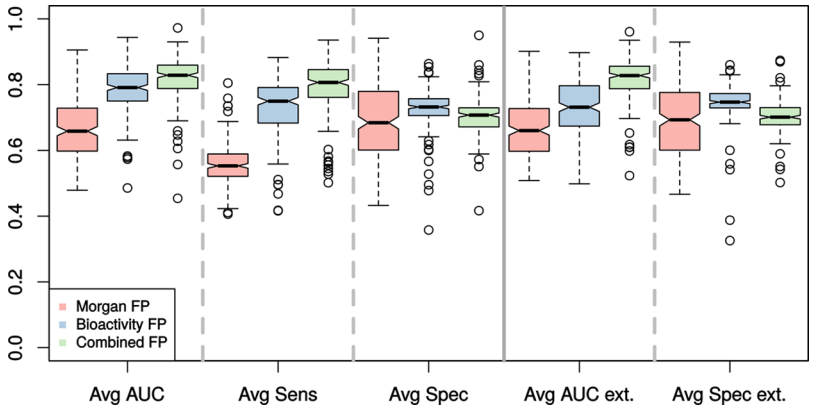


Figure 8. Boxplots of model quality measurements for Random Forest using chemical and/or bioactivity fingerprints based on *Proprietary* data. Adding bioactivity FPs mostly results in improved sensitivity. A concatenation of both FPs yields excellent performance in activity prediction. Plotted are results obtained from 5-fold CV.

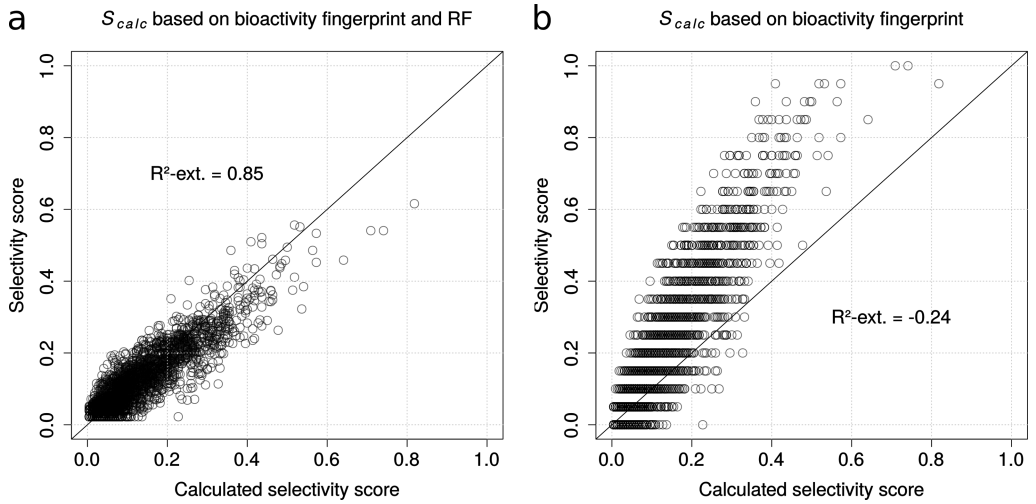


Figure 9. (a) Selectivity prediction based on bioactivity fingerprints obtained by Random Forest regression achieves an R^2_{ext} of 0.85 ± 0.01. (b) Plain averaging of experimentally obtained bioactivity fingerprints (S_{calc}) does not correlate with the experimental *S*. The calculation both in (a) and (b) are done based on 20 kinases.

Table 4. Sensitivity and Specificity of (Off-)Target Prediction[a]

		average		median	
fingerprint	no. kinases	sensitivity	specificity	sensitivity	specificity
Morgan FP	68	0.40 ± 0.39	0.90 ± 0.10	0.33 ± 0.33	0.93 ± 0.05
bioactivity FP	97	0.48 ± 0.41	0.72 ± 0.33	0.50 ± 0.50	0.86 ± 0.14
combined FP	154	0.54 ± 0.41	0.69 ± 0.36	0.60 ± 0.40	0.85 ± 0.15
bioactivity FP	68[b]	0.47 ± 0.42	0.72 ± 0.29	0.50 ± 0.50	0.83 ± 0.14
combined FP	68[b]	0.61 ± 0.41	0.70 ± 0.31	0.75 ± 0.25	0.85 ± 0.12

[a]Considered were only prediction models with an AUC ≥ 0.8 and which were in the *Proprietary* panel. [b]For comparison of the fingerprints, the bioactivity and combined FPs were also evaluated on the 68 HQ models based on Morgan FPs.

kinase. It is equally important to have a reliable estimation of the selectivity profile. The latter can be quantified by the selectivity score *S* (cf. above). Employing the 118 HQ-RF models (based on Morgan FPs of the *Combined* set) with AUC ≥ 0.8, we reconstructed a complete panel of predicted activity probabilities for the compounds of the *Proprietary* data (*N* = 4,712) from the respective test sets of a 10-fold cross validation. On the basis of the experimental pIC_{50} values (binarized at 6.3), *S* was calculated for each compound. Compounds with an *S*-value of 0 were removed due to inactivity across the entire kinase panel. From the 118 HQ models, 68 kinases were in the *Proprietary* panel, which consists of an almost complete matrix of bioactivity values. Hence, the calculated selectivity score S_{calc} is the average of the binary classification results (default probability cutoff: 0.5) obtained by these 68 kinase models. The calculated S_{calc} shows no predictive power of *S* with an $R^2_{ext} < 0$. Similarly, an RF regressor trained on the predicted panel of 68 kinases resulted only in an average R^2_{ext} of 0.17. However, an RF regressor trained on a bioactivity FP of 20 selected kinases (cf. above) resulted in very accurate selectivity regression model with an average R^2_{ext} of 0.85 ± 0.01 (Figure 9a). Notably, this is not trivial because the plain average of the 20 experimental

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485

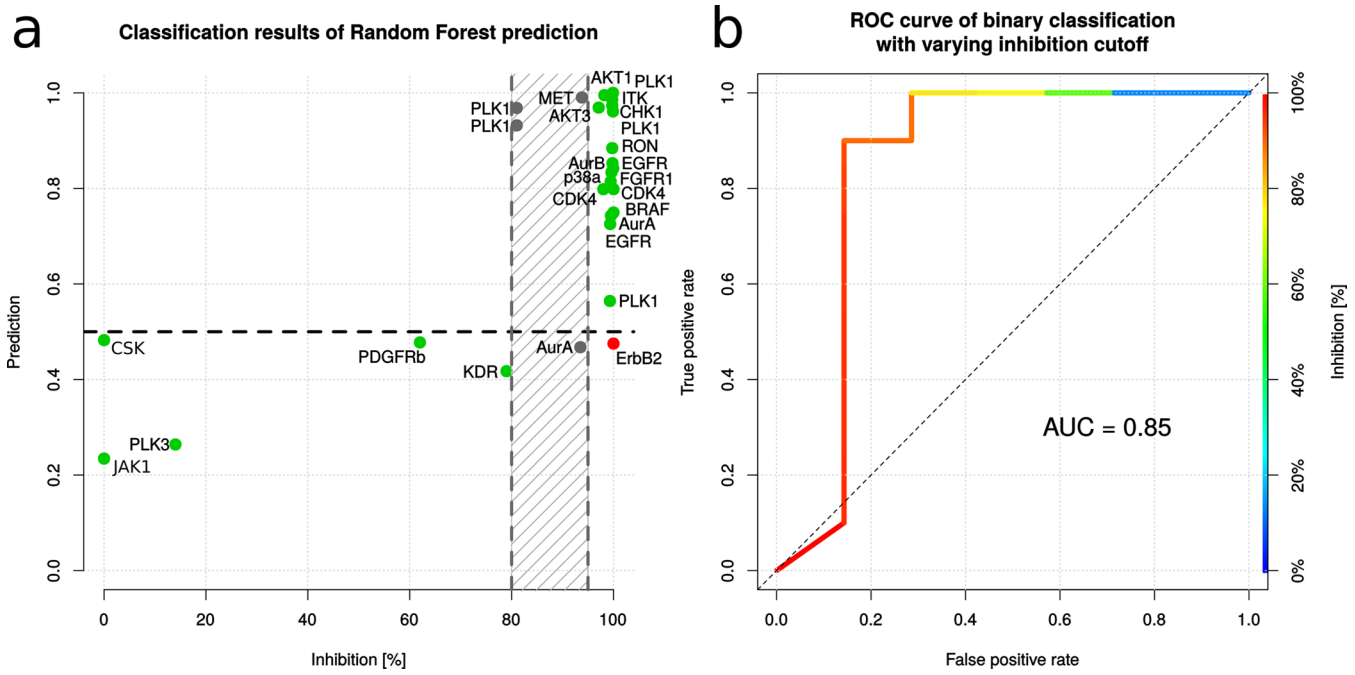


Figure 10. (a) Classification results of external data using the RF classifier. With a default probability cutoff of 0.5, only four wrong classifications at a 95% inhibition cutoff and two wrong classifications at an 80% inhibition cutoff appear for the RF classifier. Cohen's κ values of 0.65 and 0.79, respectively, suggest a very high predictive power. (b) ROC analysis of binary classification of the external data with varying experimental inhibition cutoff. An AUC of 0.85 confirms the successful application of the models on the external test set.

activity values does not yield an accurate estimate of the overall selectivity ($R^2_{ext} < 0$, Figure 9b).

Finally, the predicted panel was evaluated for correct kinase classifications per compound to assess the predictive power of our models for off-target identification and compound repurposing. Although the 68 HQ models do not allow a reliable estimation of selectivity scores (cf. above), they can in fact be used to reliably identify off-targets (Table 4). On average, 40% of all compound targets are correctly detected while only 10% false positive targets are among the predictions. Using bioactivity FPs for off-target prediction further increases the sensitivity of the predicted off-targets while the specificity drops slightly. Again, the best results are obtained with a combination of Morgan and bioactivity FPs. These findings strongly emphasize the applicability of the presented approach for unprecedented virtual kinase profiling at a large scale.

DISCUSSION AND CONCLUSION

Various methods for data set preparation and Machine Learning (ML) were employed to generate activity prediction models for over 280 kinases. The most important findings which improved the model quality were (1) the combination of open data with the proprietary data set from Merck KGaA, (2) choosing the RF classifier over alternative ML methods, and (3) balancing the data sets using random undersampling. An average ROC AUC $\geq$ 0.7 was achieved for ~200 kinases, of which 118 prediction models had an AUC $\geq$ 0.8. Tests on left-out data suggest a reliable applicability in virtual screening projects. Moreover, the models also enable reliable virtual kinase profiling and, thus, the detection of potential off-targets for compounds of interest.

The results are well in line with findings reported in recent literature, especially regarding the bias of the used training data. For instance, Bora et al. trained RF classifiers for 107 kinases, of which 100 achieved AUC values >0.9.[58] However, external validation of these models on the Metz and Anastassiadis data sets[5,9] strongly indicated an overfitting, as also experienced in our models trained on the *Open* data set only. Namely, although in our case 113 models with average AUC values >0.9 were derived based on *Open* data, external testing on the *Proprietary* panel resulted in rather low predictive power. On the other hand, as seen in model evaluation on left-out data (Figure 2a) and the low standard deviations in cross-validations (Figure 2b), our kinase activity models based on the *Combined* data do not show a strong dependency on the respective random training/test set split and are, thus, applicable on a larger chemical space.

For further external validation, we tested our models on 43 compounds, which were previously also used by Schürer and Muskal.[22] These compounds were not present in our training set. Because the underlying assay for these compounds (KINOMEScan[56,59]) was measured at a high concentration of 10 μM, a high cutoff of 95% inhibition was used for binary classification of these external test cases (this roughly corresponds to a pIC_{50} of 6.3, the cutoff used for model training). Without the necessity for adjusting the RF classifier probability cutoff for binary classification (default: 0.5), only three compounds were wrongly classified to be active and one compound wrongly classified as inactive (marked in gray and red, respectively; Figure 10a). Although the latter shows inhibition well above 95%, the predicted activity probability is slightly below the 0.5 cutoff. Encouragingly, a Cohen's κ value of 0.65 indicates a high predictive power of the tested models.[60] Because it is very difficult to derive an accurate estimation of the pIC_{50} from a single point measurement at activities above 80%, a second cutoff was set to 80% inhibition. In this case, only two false negatives appear, increasing Cohen's κ value to 0.79. To further minimize the error that might result from inhibition data conversion, the binary classification was evaluated using a varying threshold for experimental activity

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485

in a ROC analysis (Figure 10b). A very high AUC of 0.85 and a good early enrichment of true positives suggest that the classifiers can be applied successfully on the external data set regardless of the chosen cutoff for experimental activity.

Regarding the data balancing, random undersampling produced the models with the best performance, followed by the PCA-Centroids undersampling, which showed a higher sensitivity at the cost of a decline in specificity. A great improvement in the classification performance could be achieved by adding experimental activity values of a small kinase subset (20 kinases) to the chemical FP, yielding HQ models with an AUC $\geq$ 0.8 for 154 kinases of the 220 kinases in the *Proprietary* panel. This strongly suggests that bioactivity FPs from only a small number of kinases (in combination with Morgan FPs) contain enough information to accurately propagate bioactivity values to a kinase panel almost 10 times the size of the experimental subset. Furthermore, the combination of RF classification and bioactivity FPs of only 20 kinases allowed a reliable prediction of global compound selectivity *S*. Importantly, this cannot necessarily be derived directly from the raw experimental activity values.

In line with other studies, the RF classifier outperformed the Naive Bayes methods,[21] while the Deep Learning approach also generated models with excellent predictive power. It might be possible that a more extensive parameter study and the expansion of the Neural Network layers would further improve the performance of the Deep Learning activity prediction.[61] Notably, the Deep Learning approach also benefited from balanced training sets (data not shown).

Overall, the evaluation results of our prediction models strongly indicate a positive impact on future screening projects and off-target identification tasks. The former can be used to assess for a particular kinase of interest which compounds should be ordered for experimental verification (virtual screening) or to prioritize already investigated compounds from previous projects on new kinase targets (repurposing). On the other hand, the identification of off-targets is a prerequisite for the rational design of selective kinase inhibitors.[62] In ongoing research, it will be assessed how the addition of proteogenomic information and the combined usage of RF classifiers with Deep Learning networks can further improve the predictive power of the activity prediction models.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jmedchem.6b01611. The curated *Open* data set and source code for deriving prediction models can be downloaded from https://github.com/Team-SKI/Publications.

Additional external model validation, correlations between different ML methods, and the impact of various balancing techniques on the prediction performance. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49 6221 4261113. E-mail: fulle@bio.mx.

ORCID

Simone Fulle: 0000-0002-7646-5889

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Andrea Volkamer and Guillaume Roellinger (BioMed X Innovation Center, Heidelberg), Paul Czodrowski and Mireille Krier (Merck KGaA, Darmstadt), and Rebecca Wade (Heidelberg Institute for Theoretical Studies) for numerous valuable discussions. Furthermore, we are grateful to Günter Klambauer and Sepp Hochreiter (Johannes Kepler University Linz) for very helpful conversations regarding Deep Neural Networks.

■ ABBREVIATIONS USED

CV, cross-validation; DNN, Deep Neural Network; FP, fingerprint; HQ, high-quality; IQR, inter-quartile range; KNN, K-Nearest Neighbor; ML, Machine Learning; NB, Naïve Bayes; PCM, proteochemometrics; RF, Random Forest; ROC, receiver operating characteristic; SVM, Support Vector Machine; VS, virtual screening

■ REFERENCES

(1) Cohen, P. Protein kinases-the major drug targets of the twenty-first century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309−315.

(2) Cohen, P.; Alessi, D. R. Kinase drug discovery—what's next in the field? *ACS Chem. Biol.* **2013**, *8*, 96−104.

(3) Wu, P.; Nielsen, T. E.; Clausen, M. H. Small-molecule kinase inhibitors: an analysis of FDA-approved drugs. *Drug Discovery Today* **2016**, *21*, 5−10.

(4) Rask-Andersen, M.; Zhang, J.; Fabbro, D.; Schiöth, H. B. Advances in kinase targeting: current clinical use and clinical trials. *Trends Pharmacol. Sci.* **2014**, *35*, 604−620.

(5) Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the kinome. *Nat. Chem. Biol.* **2011**, *7*, 200−202.

(6) Paricharak, S.; Klenka, T.; Augustin, M.; Patel, U. A.; Bender, A. Are phylogenetic trees suitable for chemogenomics analyses of bioactivity data sets: the importance of shared active compounds and choosing a suitable data embedding method, as exemplified on kinases. *J. Cheminf.* **2013**, *5*, 49.

(7) Ferrè, F.; Palmeri, A.; Helmer-Citterich, M. Computational methods for analysis and inference of kinase/inhibitor relationships. *Front. Genet.* **2014**, *5*, 196.

(8) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011**, *29*, 1046−1051.

(9) Anastassiadis, T.; Deacon, S. W.; Devarajan, K.; Ma, H.; Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011**, *29*, 1039−1045.

(10) Tang, J.; Szwajda, A.; Shakyawar, S.; Xu, T.; Hintsanen, P.; Wennerberg, K.; Aittokallio, T. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inf. Model.* **2014**, *54*, 735−743.

(11) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.

(12) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361−367.

(13) Hillisch, A.; Heinrich, N.; Wild, H. Computational chemistry in the pharmaceutical industry: from childhood to adolescence. *ChemMedChem* **2015**, *10*, 1958−1962.

(14) Martin, E.; Mukherjee, P.; Sullivan, D.; Jansen, J. Profile-QSAR: a novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity. *J. Chem. Inf. Model.* **2011**, *51*, 1942−1956.

(15) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197−206.
(16) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175−181.
(17) Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Livingstone, D. J.; Ford, M. G.; Whitley, D. C. Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. *J. Chem. Inf. Model.* **2002**, *42*, 1256−1262.
(18) Ning, X.; Walters, M.; Karypisxy, G. Improved machine learning models for predicting selective compounds. *J. Chem. Inf. Model.* **2012**, *52*, 38−50.
(19) Unterthiner, T.; Mayr, A.; Klambauer, G.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Hochreiter, S. Deep learning as an opportunity in virtual screening. *Proceedings of the Deep Learning Workshop at NIPS, Montreal (QC), December 12, 2014*; 2014.
(20) Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* **2011**, *7*, 472.
(21) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Comparison of random forest and Pipeline Pilot naive bayes in prospective QSAR predictions. *J. Chem. Inf. Model.* **2012**, *52*, 792−803.
(22) Schürer, S. C.; Muskal, S. M. Kinome-wide activity modeling from diverse public high-quality data sets. *J. Chem. Inf. Model.* **2013**, *53*, 27−38.
(23) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5−32.
(24) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **2003**, *43*, 1947−1958.
(25) Goldstein, D. M.; Gray, N. S.; Zarrinkar, P. P. High-throughput kinase profiling as a platform for drug discovery. *Nat. Rev. Drug Discovery* **2008**, *7*, 391−397.
(26) Cortés-Ciriano, I.; Ain, Q. U.; Subramanian, V.; Lenselink, E. B.; Méndez-Lucio, O.; IJzerman, A. P.; Wohlfahrt, G.; Prusis, P.; Malliavin, T. E.; van Westen, G. J.; Bender, A. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *MedChemComm* **2015**, *6*, 24−50.
(27) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. QSAR modeling of imbalanced high-throughput screening data in PubChem. *J. Chem. Inf. Model.* **2014**, *54*, 705−712.
(28) Gao, Y.; Davies, S. P.; Augustin, M.; Woodward, A.; Patel, U. A.; Kovelman, R.; Harvey, K. J. A broad activity screen in support of a chemogenomic map for kinase signalling research and drug discovery. *Biochem. J.* **2013**, *451*, 313−328.
(29) Dranchak, P.; MacArthur, R.; Guha, R.; Zuercher, W. J.; Drewry, D. H.; Auld, D. S.; Inglese, J. Profile of the GSK published protein kinase inhibitor set across ATP-dependent and-independent luciferases: implications for reporter-gene assays. *PLoS One* **2013**, *8*, e57888.
(30) Knapp, S.; Arruda, P.; Blagg, J.; Burley, S.; Drewry, D. H.; Edwards, A.; Fabbro, D.; Gillespie, P.; Gray, N. S.; Kuster, B.; Lackey, K. E.; Mazzafera, P.; Tomkinson, N. C. O.; Willson, T. M.; Workman, P.; Zuercher, W. J. A public-private partnership to unlock the untargeted kinome. *Nat. Chem. Biol.* **2013**, *9*, 3−6.
(31) Elkins, J. M.; Fedele, V.; Szklarz, M.; Abdul Azeez, K. R.; Salah, E.; Mikolajczyk, J.; Romanov, S.; Sepetov, N.; Huang, X.-P.; Roth, B. L.; Al Haj Zen, A.; Fourches, D.; Muratov, E.; Tropsha, A.; Morris, J.; Teicher, B. A.; Kunkel, M.; Polley, E.; Lackey, K. E.; Atkinson, F. L.; Overington, J. P.; Bamborough, P.; Müller, S.; Price, D. J.; Willson, T. M.; Drewry, D. H.; Knapp, S.; Zuercher, W. J. Comprehensive characterization of the Published Kinase Inhibitor Set. *Nat. Biotechnol.* **2015**, *34*, 95−103.
(32) Volkamer, A.; Eid, S.; Turk, S.; Jaeger, S.; Rippmann, F.; Fulle, S. Pocketome of human kinases: prioritizing the ATP binding sites of (yet) untapped protein kinases for drug discovery. *J. Chem. Inf. Model.* **2015**, *55*, 538−549.
(33) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of mixed IC_{50} data-a statistical analysis. *PLoS One* **2013**, *8*, e61007.
(34) *R: a Language and Environment for Statistical Computing*; R Core Team, 2015; https://www.R-project.org/ (accessed February 18, 2016).
(35) Oksanen, J.; Blanchet, F. G.; Kindt, R.; Legendre, P.; Minchin, P. R.; O'Hara, R. B.; Simpson, G. L.; Solymos, P.; Stevens, M. H. H.; Wagner, H. *Vegan: Community Ecology Package*; 2016; R package version 2.3−4.
(36) Neuwirth, E. *RColorBrewer: Colorbrewer Palettes*; 2014; R package version 1.1−2.
(37) Wal, J. V. D.; Falconi, L.; Januchowski, S.; Shoo, L.; Storlie, C. *SDMTools: Species Distribution Modelling Tools; Tools for Processing Data Associated with Species Distribution Modelling Exercises*, 2014; R package version 1.1−221.
(38) *RDKit: Open-Source Cheminformatics Software*, 2016; http://www.rdkit.org, (accessed February 18, 2016).
(39) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(40) 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.; Duchesnay, E. *Scikit-learn*: machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(41) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a bayesian model. *J. Med. Chem.* **2004**, *47*, 4463−4470.
(42) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. *TensorFlow*: large-scale machine learning on heterogeneous systems. 2015; http://tensorflow.org/, accessed February 18, 2016.
(43) Nair, V.; Hinton, G. E. Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Isreal, June 21−24, 2010*, 2010; pp 807−814.
(44) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.
(45) Mani, I.; Zhang, I. kNN approach to unbalanced data distributions: a case study involving information extraction. *Proceedings of Workshop on Learning from Imbalanced Datasets, ICML, Washington DC, August 21, 2003*.
(46) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, 321−357.
(47) Japkowicz, N. The class imbalance problem: significance and strategies. *Proceedings of the 2000 International Conference on Artificial Intelligence, Las Vegas, Nevada, June 26−29, 2000*.
(48) Zhao, Z.; Wu, H.; Wang, L.; Liu, Y.; Knapp, S.; Liu, Q.; Gray, N. S. Exploration of type II binding mode: a privileged approach for kinase inhibitor focused drug discovery? *ACS Chem. Biol.* **2014**, *9*, 1230−1241.
(49) Fedorov, O.; Müller, S.; Knapp, S. The (un)targeted cancer kinome. *Nat. Chem. Biol.* **2010**, *6*, 166−169.
(50) Zhang, L.; Daly, R. J. Targeting the human kinome for cancer therapy: current perspectives. *Crit. Rev. Oncog.* **2012**, *17*, 233−246.
(51) Riniker, S.; Fechner, N.; Landrum, G. A. Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485

making by committee can be a good thing. *J. Chem. Inf. Model.* **2013**, *53*, 2829−2836.
(52) Riniker, S.; Wang, Y.; Jenkins, J. L.; Landrum, G. A. Using information from historical high-throughput screens to predict active compounds. *J. Chem. Inf. Model.* **2014**, *54*, 1880−1891.
(53) 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**, *9*, 1622−1631.
(54) 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.
(55) Cortes Cabrera, A.; Lucena-Agell, D.; Redondo-Horcajo, M.; Barasoain, I.; Diaz, F.; Fasching, B.; Petrone, P. Compound biological signatures facilitate phenotypic screening and target elucidation. *bioRxiv* **2016**, 041947.
(56) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2008**, *26*, 127−132.
(57) Yagnik, J.; Strelow, D.; Ross, D. A.; Lin, R.-s. The power of comparative reasoning. *International Conference on Computer Vision (IEEE), Barcelona, Spain, November 6−13, 2011*; pp 2431−2438.
(58) Bora, A.; Avram, S.; Ciucanu, I.; Raica, M.; Avram, S. Predictive models for fast and effective profiling of kinase inhibitors. *J. Chem. Inf. Model.* **2016**, *56*, 895−905.
(59) Fabian, M. A.; Biggs, W. H.; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lélias, J.-M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329−336.
(60) McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med.* **2012**, *22*, 276−282.
(61) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.
(62) Volkamer, A.; Eid, S.; Turk, S.; Rippmann, F.; Fulle, S. Identification and visualization of kinase-specific subpockets. *J. Chem. Inf. Model.* **2016**, *56*, 335−346.

DOI: 10.1021/acs.jmedchem.6b01611
J. Med. Chem. 2017, 60, 474−485