

Successive Statistical and Structure-Based Modeling to Identify Chemically Novel Kinase Inhibitors

Lindsey Burggraaff, Eelke B. Lenselink, Willem Jaspers, Jesper van Engelen, Brandon J. Bongers, Marina Gorostiola González, Rongfang Liu, Holger H. Hoos, Herman W. T. van Vlijmen, Adriaan P. IJzerman, and Gerard J. P. van Westen*



Cite This: *J. Chem. Inf. Model.* 2020, 60, 4283–4295



Read Online

ACCESS |



Metrics & More

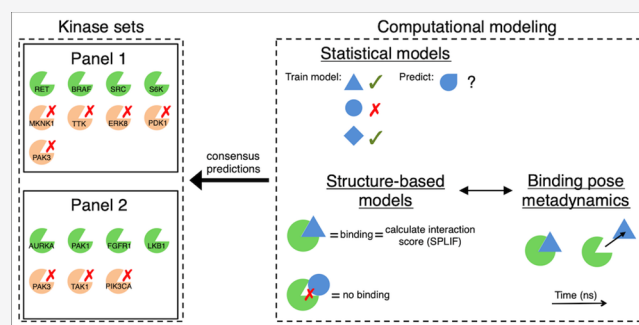


Article Recommendations



Supporting Information

ABSTRACT: Kinases are frequently studied in the context of anticancer drugs. Their involvement in cell responses, such as proliferation, differentiation, and apoptosis, makes them interesting subjects in multitarget drug design. In this study, a workflow is presented that models the bioactivity spectra for two panels of kinases: (1) inhibition of RET, BRAF, SRC, and S6K, while avoiding inhibition of MKNK1, TTK, ERK8, PDK1, and PAK3, and (2) inhibition of AURKA, PAK1, FGFR1, and LKB1, while avoiding inhibition of PAK3, TAK1, and PIK3CA. Both statistical and structure-based models were included, which were thoroughly benchmarked and optimized. A virtual screening was performed to test the workflow for one of the main targets, RET kinase. This resulted in 5 novel and chemically dissimilar RET inhibitors with remaining RET activity of <60% (at a concentration of 10 μ M) and similarities with known RET inhibitors from 0.18 to 0.29 (Tanimoto, ECFP6). The four more potent inhibitors were assessed in a concentration range and proved to be modestly active with a pIC_{50} value of 5.1 for the most active compound. The experimental validation of inhibitors for RET strongly indicates that the multitarget workflow is able to detect novel inhibitors for kinases, and hence, this workflow can potentially be applied in polypharmacology modeling. We conclude that this approach can identify new chemical matter for existing targets. Moreover, this workflow can easily be applied to other targets as well.



INTRODUCTION

Compound promiscuity can be leveraged to develop multitarget drugs. Such multitarget drugs can replace existing multidrug treatments, while maintaining the therapeutic effect.¹ One of the advantages of a multitarget drug or single-drug treatment is that no drug–drug interactions occur, making the treatment less risky and harmful for the patient.² However, since multitarget drugs are designed to bind to multiple proteins, they may tend to be more promiscuous as well. Therefore, when developing multitarget compounds, off-target binding and pathways should also be considered.

In the light of the “Multi-Targeting Drug” DREAM challenge,³ bioactivities were computationally modeled for two panels of kinases. The first panel was based on treatment of medullary thyroid carcinoma, where kinases RET, BRAF, SRC, and S6K, should be inhibited and MKNK1, TTK, ERK8, PDK1, and PAK3, should not be affected.^{4,5} RET was considered the main on-target kinase in this panel and, thus, was prioritized over other kinases in the panel. The second panel was based on tauopathies in neurodegenerative disease: compounds should inhibit AURKA, PAK1, FGFR1, and LKB1 and not bind to PAK3, TAK1, and PIK3CA.³ Since the main on-targets, AURKA and PAK1, and additional on-targets

FGFR1 and LKB1, are targeted in the central nervous system, compounds for panel 2 kinases should additionally be able to pass the blood–brain barrier.

This study describes a rigorous workflow to model the bioactivity spectra of compounds in kinases (Figure 1) and identify *novel* inhibitors. Every step in the workflow was extensively benchmarked, and each model was validated prior to virtual screening. A consensus scoring approach was used to rank virtual screening compounds, and only compounds with a Tanimoto similarity (ECFP6 (extended-connectivity bit string, diameter 6))⁶ < 0.4 were considered to make sure that existing active molecules would not be “rediscovered”. This consensus approach encompassed statistical modeling techniques, such as quantitative structure–activity relationship (QSAR) models and proteochemometric (PCM) modeling.^{7–9} Moreover, structure-based docking and pose metadynamics were

Special Issue: New Trends in Virtual Screening

Received: December 30, 2019

Published: April 28, 2020



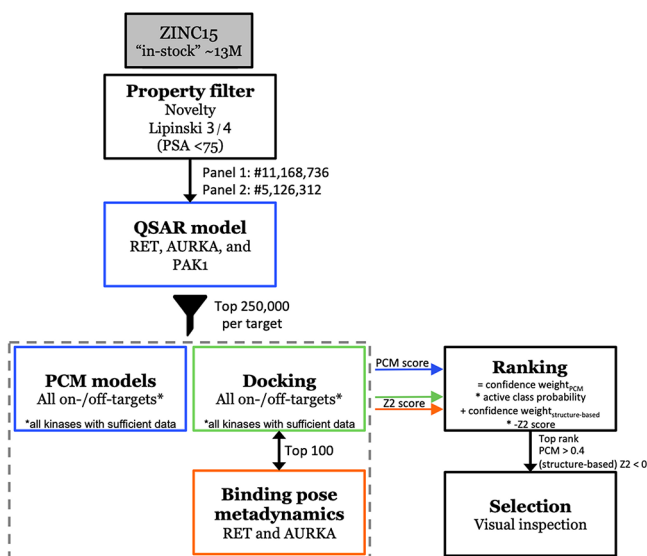


Figure 1. Virtual screening workflow for the two panels of kinases. Statistical models (blue), structure-based models (green), and molecular dynamics (orange) were applied to rank the virtual screening compounds.

applied.¹⁰ Next to compound ranking, this approach was also used to exclude inactive-predicted compounds from virtual screening along the way. Fast machine learning models were applied to discard compounds early in the workflow. Subsequently, slower, but information rich, structure-based models were applied that consequently only had to score a smaller fraction of compounds compared to the entire initial virtual screening set. In this way, millions of compounds were screened rapidly and scored accurately.

In addition to the extensive benchmark validations, the workflow was also validated in vitro for one of the main on-target kinases. Forty-six compounds were experimentally validated for RET, of which 15 were selected based on the consensus approach, 6 based on statistical models only, and 25 compounds based on structure-based modeling. This resulted

in a total of 5 inhibitors causing RET activity < 60% at a concentration of 10 μ M. The four most potent compounds were further inspected and their IC_{50} were determined. Although the most potent RET inhibitor was modestly active with a pIC_{50} of 5.1 ± 0.1 , all compounds were chemically distinct from known RET inhibitors with Tanimoto (ECFP6) similarities smaller than 0.29. Therefore, the identified inhibitors in this study provide a new starting point for hit/lead optimization based on novel scaffolds.

RESULTS

Data Curation and Filtering. Compound bioactivity data was derived from different sources to increase data availability for the benchmark sets in model training and validation. Kinase compound data was retrieved from the following sources: ChEMBL database (version 23),¹¹ publicly available sets from Eidogen,¹² and ExCAPE-DB.¹³ The data was curated by standardizing chemical structures and bioactivity values (see *Methods* section for details). In this way, a data set was constituted that contained compound bioactivity information for 512 kinases (data set available at [10.4121/uuid:6af1d9de-281f-4221-b7e1-e7c01b90dfe0](https://doi.org/10.4121/uuid:6af1d9de-281f-4221-b7e1-e7c01b90dfe0)). This data set was used for training and testing of statistical models. Since validation was performed using cross-validation, the data set was divided into subsets; five subsets were created based on chemical clustering of the compounds per target (see *Methods* section for details). The biggest subset was used in training of every model, whereas the remaining four subsets were used rotationally for testing. The subsets that were not used in testing in the specific iteration were added to the training set.

Additionally, separate active/inactive/decoy data sets were constructed for validation of structure-based models. These benchmark sets were constructed for each panel kinase separately. Data for these benchmark sets were derived from ChEMBL (version 23), which was curated and filtered to construct a benchmark set for each kinase containing inactive and chemically diverse active compounds: only compounds with measured pChEMBL¹⁴ values were included in the benchmark sets and a restrain was set for the number of active

Table 1. PCM Performance Per Target^a

target	PCM			QSAR			number of compounds	
	MCC	BEDROC ($\alpha = 20$)	ROC	MCC	BEDROC ($\alpha = 20$)	ROC	active (pChEMBL ≥ 6.5)	inactive (pChEMBL < 6.5)
RET	0.15	0.64	0.76	0.23	0.63	0.75	1492	416
BRAF	0.18	0.74	0.56	0.20	0.75	0.54	1119	1359
SRC	0.28	0.47	0.72	0.26	0.47	0.72	4642	2238
S6K	0.38	0.85	0.79	0.45	0.85	0.78	1662	685
MKINK1	0.09*	0.42	0.61	0.01	0.32	0.50	549	51
TTK1	0.22	0.45	0.78	0.26	0.44	0.75	663	276
ERK8	****	0.05	0.48	-0.12	0.02	0.35	302	30
PDK1	0.27*	0.85	0.72	0.31	0.86	0.71	579	536
PAK3	0.25	0.72	0.91	0.07	0.27	0.71	1204	53
AURKA	0.37	0.65	0.78	0.38	0.47	0.77	3165	1674
PAK1	0.32	0.74	0.86	0.28	0.66	0.77	712	114
FGFR1	0.41	0.70	0.85	0.77	0.71	0.82	2477	928
LKB1	0.57**	0.53	0.76	0.26*	0.45	0.63	429	47
TAK1	0.15***	0.27	0.68	0.12*	0.33	0.69	1204	53
Average	0.25	0.58	0.74	0.25	0.52	0.68	295	56

^aMean over 4-fold cross-validation. Asterisks indicate that no value could be determined due to lack of predicted (positive/negative) classes: *, 1 cross-validation failed; **, 2 cross validations failed; ***, 3 cross validations failed; ****, 4 cross validations failed. Indicated in bold in each column is the best performing model for that given parameter.

compounds. Up to a maximum of 100 actives per kinase were included, whereas the number of inactive compounds was not limited. The 100 active compounds were selected based on chemical diversity by clustering the compounds with pChEMBL value >6.5 into 100 clusters and selecting only the cluster centers. Additionally, DUD-E¹⁵ decoys were added to each benchmark set. These decoys have similar physicochemical properties as active ligands but differ in chemical structure. The structure-based benchmark sets were smaller than the training and test sets of statistical models, but big enough to allow for validation of the models. The smaller size of the structure-based benchmark sets allowed for quicker model evaluation, resulting in validation of many protein structures.

The virtual screening set that was screened using both statistical- and structure-based models was derived from the ZINC15¹⁶ database. All “in stock” compounds were filtered on drug-like properties by discarding compounds that did not adhere to 3 of 4 Lipinski rules.¹⁷ Furthermore, compounds were filtered on novelty: compounds with Tanimoto similarity (ECFP6) > 0.4 compared to existing actives (pChEMBL > 5) on the kinases in the respective panel were excluded from the virtual screening set. The virtual screening set was additionally filtered for panel 2 kinases by including the likelihood of compound passing the central nervous system by only keeping compounds with polar surface area $< 75 \text{ \AA}^2$. This resulted in a virtual screening set for panel 1 of 11 168 736 compounds and for panel 2 of 5 126 312 compounds.

Statistical Models. Separate quantitative structure–activity relationship (QSAR) models were constructed for the main on-targets RET, AURKA, and PAK1, as a first filter for bioactive compounds. The models were validated with 4-fold cross-validation using standardized benchmark sets (see [Methods](#) for details). The benchmark sets were constructed per target and were extracted from the main statistical benchmark set containing 512 kinases. Chemical descriptors were calculated for every compound: FCFP4 (feature-connectivity bit string, diameter 4) fingerprints and physicochemical descriptors (listed in [Table S2](#)). These descriptors describe the compounds and were used in training the models. The RET, AURKA, and PAK1 QSAR models were predictive with ROC (receiver operating characteristic) scores higher than random ([Table 1](#)). The ROC of the QSAR models was comparable between targets (ROC 0.76 ± 0.01), whereas the Matthews correlation coefficient (MCC) varied slightly more (MCC 0.30 ± 0.08).

The performances of the QSAR models were sufficient as a first filter for bioactive compounds, discarding the least active compounds and steering clear of the decision boundary. Virtual screening set 1 was screened using the RET QSAR, and virtual screening set 2 was screened using the AURKA and PAK1 QSAR models separately. Using the active class probability score, the most promising compounds (250 000 compounds per RET/AURKA/PAK1) were selected to be further processed in the structure-based approaches. This prescreening of compounds with a simple, but fast model decreased the number of compounds effectively. As a result, subsequent screening and scoring steps proceeded quicker, since fewer compounds had to be screened. The subsequent steps, proteochemometrics (PCM)⁹ and structure-based modeling, were carried out in parallel. Additionally, QSAR models were constructed for the remaining kinases simultaneously. These QSAR models were compared to the more

advanced PCM models to select the best approach for scoring. The performances of these QSAR models are included in [Table 1](#).

The PCM models that were created were applied solely for the purpose of scoring and not as a filtering step. PIK3CA was excluded from modeling, as insufficient data was available to build and validate the models. The PCM models were constructed for each kinase separately and were based on the particular kinase and its L4 level family members as annotated in ChEMBL.¹⁸ In addition to the compound descriptors used in QSAR modeling (FCFP4 fingerprints and physicochemical properties), the PCM models included protein descriptors that were based on a full kinome sequence alignment (see [Methods](#) for details). Initial PCM models were trained using default random forest settings: 300 trees, $\log_2(m)$ features at every node in the tree, no maximum tree depth, a minimum of 2 samples to consider a node for splitting, and bootstrap enabled. This resulted in an overall performance of MCC 0.22 and ROC 0.69 (average over all panel kinases). Since the PCM models were intended for scoring of compounds, the models were optimized to enhance predictive performance. Optimal settings were explored by tuning hyperparameters with random search, a basic approach to automated machine learning.¹⁹ Approximately 500 random forest models were trained during optimization, resulting in the best model with performance MCC 0.25 and ROC 0.74 and the following settings: 300 trees (fixed) with 43 features at every node in the tree, a maximum tree depth of 99, a minimum of 12 samples to consider a node for splitting, and bootstrapping disabled. The performance of the QSAR and PCM models per kinase are shown in [Table 1](#).

Predictions for most kinases were comparable between QSAR and PCM modeling. However, for target PAK3, PCM clearly outperformed QSAR with MCC difference 0.18 and ROC difference 0.20. MCC could not be calculated for ERK8 because of a small data set and, consequently, a lack of a predicted class (total number of compounds for ERK8 is 332). Using QSAR, the MCC displays a negative correlation for ERK8 (-0.12), which is also reflected by the ROC score that is worse than random (ROC < 0.5). PDK1 has high early enrichment (BEDROC (Boltzmann-enhanced discrimination of the receiver operating characteristic)) with both QSAR and PCM. Although overall enrichment (ROC) for PDK1 is lower than the early enrichments, it is still good with ROC 0.71 (QSAR) and 0.72 (PCM). The average over all targets shows that PCM predicts slightly better than QSAR, with a similar MCC score of 0.25, but higher (BED)ROC scores: differences BEDROC = 0.06 and ROC = 0.06. Moreover, the nature of PCM models allows for extrapolation of bioactivities from related kinases to the target of interest. Therefore, it was hypothesized that the PCM models will perform better than QSAR models when applied to a more diverse chemical data set such as the virtual screening sets. The performances of the PCM models of the main on-targets AURKA and PAK1 were higher than the average performance over all targets. However, main on-target RET had a PCM MCC value (0.15) that was lower than average (0.25). Nevertheless, (BED)ROC scores were higher than the average: RET BEDROC 0.64, RET ROC 0.76, average BEDROC 0.58, and average ROC 0.74. Moreover, all RET scores were better than random indicating the predictive power of the model.

The settings that corresponded to the best performing model in 4-fold cross-validation were applied to train PCM models on the full data set per kinase (including test set in

training). The models were applied to score all the virtual screening compounds that passed the filtering step using QSAR models.

Structure-Based Models. Structure-based scoring was performed in addition to scoring of compounds using the PCM models. For many kinases, multiple crystal structures have been deposited in the PDB, and it is often not obvious which crystal structure should be used prospectively. Therefore, we performed a rigorous benchmark to determine the best enriching crystal structure. The number of validated crystal structures for all targets ranged from 1 to 135, excluding ERK8 and PAK3 for which no crystal structures were available at the time (December 2017). The crystal structures that were deemed suitable for virtual screening were X-ray protein structures that contained a cocrystallized orthosteric ligand. A total of 499 crystal structures were benchmarked using corresponding compound benchmark sets that were composed for each target separately. The benchmark sets containing active compounds, inactive compounds, and decoys, were docked into the orthosteric binding pockets, from which a docking score was derived for each compound. The decoys were considered as inactive compounds when calculating actives enrichment for each crystal structure. Enrichment was calculated based on best docking score per compound. It was observed that actives enrichment varied greatly between different structures of the same kinase: ROC ranging from 0.58 (PDB 2IVS) to 0.77 (2IVU) for RET, 0.47 (SOSF) to 0.80 (2XNG and 4O0W) for AURKA, and 0.65 (3Q4Z and 4O0T) to 0.95 (5IME) for PAK1 (performances for every kinase, see Table S3). The models per target were ranked based on the sum of the overall (ROC) and early enrichment (BEDROC, $\alpha = 160.9$). Models with a score ROC+BEDROC ≥ 1 (e.g., ROC 0.40 + BEDROC 0.60) were considered sufficient for prediction of active compounds.

Kinases LKB1 and MKNK1 only had a few crystal structures available (1 and 2 structures, respectively) of which performance was low (ROC + BEDROC: 0.50 and 0.72, respectively). Moreover, kinase S6K of which 18 structures were available, only reached maximum performance of ROC+BEDROC 0.80. Therefore, additional protein structures were created for these kinases by application of induced-fit docking. In contrast to docking, induced-fit docking accommodates the ligand and additionally reorientates the side chains of binding pocket residues. This allows the residues to change conformation in order for the ligand to fit into the binding pocket. Five ligands for each LKB1, MKNK1, and S6K, were docked with induced-fit docking into the respective binding pocket. Subsequently, the ligands were removed from the protein, keeping only the altered protein structures. This resulted in 95 additional protein structures for LKB1, 44 for MKNK1, and 74 for S6K. These additional structures were validated using the same benchmark sets as used for the initial structures. The protein structures that were created using induced-fit docking varied in performance from ROC+BEDROC 0.27 to 1.07. For all three kinases, an induced-fit structure was generated that outperformed the initial crystal structure. The best performance for LKB1 was ROC+BEDROC 0.99, for MKNK1 it was 1.07, and for S6K it was 0.97. The protein structures resulting from induced-fit docking are available at dx.doi.org/10.4121/uuid:9e61b6a6-88e5-4a18-ba19-dbf1bbdd656a.

On the basis of the docking performances of all models, five protein structures per target were selected for structural protein–ligand interaction fingerprints (SPLIF)²⁰ calculations.

With SPLIF, interactions between the binding pocket residues and (docked) ligand are calculated, resulting in a SPLIF score that indicates the similarity between the interactions of the (docked) ligand compared to a reference compound. In this case, the reference compound consists of the cocrystallized ligand in the corresponding structure. The structures for SPLIF calculations were selected based on diversity of the cocrystallized ligands per kinase. The diversity of these ligands was assessed by clustering the ligands using affinity propagation clustering.²¹ The cluster centers of the structures with highest ROC+BEDROC in docking were selected as reference. Consequently, the corresponding protein structures and docked benchmark set poses were thus used in SPLIF calculations. The SPLIF similarity scores for each benchmark compound were used to calculate actives enrichment based on SPLIFs. Especially for MKNK1 actives enrichment increased significantly compared to enrichment based on docking scores: ROC+BEDROC SPLIF 1.57 versus ROC+BEDROC docking 1.07. On the basis of early enrichment, BEDROC ($\alpha = 160.9$), the performance using SPLIF scores increased for 9 of 13 kinases (Figure 2). However, for targets BRAF, PDK1, PAK1, and FGFR1 docking scores enriched (BEDROC ($\alpha = 160.9$)) better than SPLIF scores.

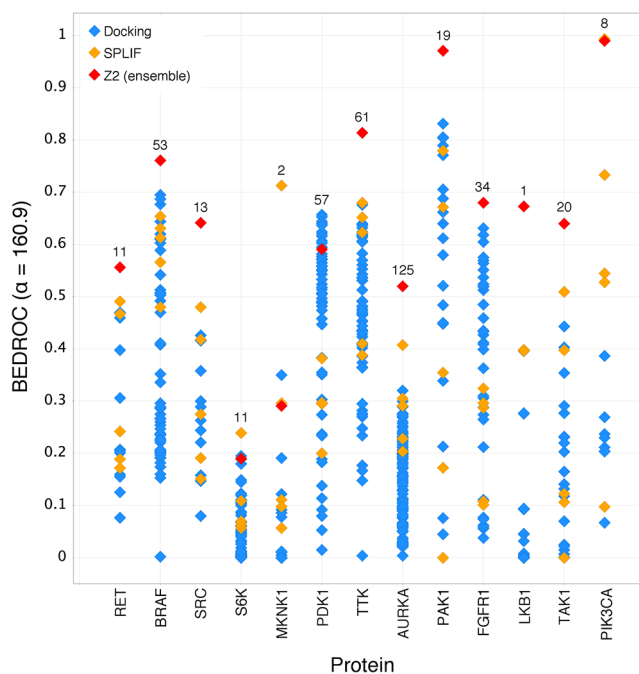


Figure 2. Early enrichment of actives, BEDROC ($\alpha = 160.9$), per crystal structure for each target. Enrichment reached by docking scores (blue), SPLIF scores (orange), and Z2-scores (red), are shown for all 13 kinases that had a crystal structure available. Numbers on top indicate the number of crystal structures for each kinase.

An ensemble score was constituted, which combined docking and SPLIF scores. This ensemble score, the Z2 score,²² was based on only docking scores, only SPLIF scores, or a combination of docking and SPLIF scores (see Methods for details). Prior to calculation of Z2-scores, the docking and SPLIF scores were normalized to Z-scores (more negative Z-score means better binding). This normalization was done with respect to the actives from the benchmark set for a given kinase. The early enrichments of active compounds based on Z2-scoring increased performance for 10 of the 13 targets

Table 2. Weights Assigned to Each Target Per Modeling Technique

	panel 1 kinases								
	on-target				off-target				
	RET	BRAF	SRC	S6K	MKNK1	TTK	ERK8	PDK1	PAK3
PCM	0.54	0.14	0.90	0.67	0.13	0.41	−0.02	0.35	0.70
structure based	1.44	1.66	1.56	0.48	1.42	1.79	n.a.	1.45	n.a.
total	1.98	1.80	2.46	1.15	1.55	2.20	−0.02	1.80	0.70
	panel 2 kinases								
	on-target				off-target				
	AURKA	PAK1	FGFR1	LKB1	TAK1	PIK3CA	PAK3		
PCM	0.93	0.50	0.98	0.28	0.16	n.a.		0.70	
structure based	1.34	1.23	1.62	0.04	0.46	0.97	n.a.		
total	2.27	1.73	2.60	0.32	0.62	0.97		0.70	

(Figure 2). The best (ensemble of) models per target are listed in Table S5. The best performances of the main on-targets RET, AURKA, and PAK1 were reached by Z2-scoring. The performance of PAK1 was very good with BEDROC ($\alpha = 160.9$) 0.97. However, it should be noted that the number of compounds in this benchmark set was low due to lack of data (63 active compounds, 10 inactive compounds and 3886 decoys). Therefore, this model may not be representative when applied to a virtual screening set. The best BEDROC performances of RET and AURKA were 0.56 and 0.52, respectively. Furthermore, the overall performance for these targets were good with ROC 0.85 for RET and ROC 0.82 for AURKA.

Prospective Structure-Based Docking. The virtual screening compounds resulting from QSAR filtering were docked into the protein structures that resulted in the best performances in benchmarking. Not the entire virtual screening set could be docked on all kinases because of time constraints. Therefore the top 250 000 compounds for RET, AURKA, and PAK1, with highest active-class probabilities, were selected for docking. This included all active-predicted compounds for RET and AURKA (167 828 for RET and 315 for AURKA), and additional, consecutively ranked compounds that did not reach the QSAR activity threshold (active class probability > 0.5). For PAK1, the top 250 000 compounds were selected from a total of 298 163 active-predicted compounds. Subsequently, the compounds were assigned a structure-based Z2 score, with the exception of S6K, MKNK1, and PDK1 for which SPLIF or docking score gave the best BEDROC ($\alpha = 160.9$) performance and thus Z2-scores were replaced with the respective score. The best scoring method for S6K underperformed with a BEDROC of 0.24. Although this model was not discarded immediately, the poor performance was taken into account later when reliability weights were assigned to the corresponding method and kinase.

Binding Pose Metadynamics. Binding pose metadynamics^{23,24} was performed to rescore the docked poses and scores of compounds. Since binding pose metadynamics is a time-consuming modeling technique, only the main on-targets (RET, AURKA, and PAK1) were subjected to this method. Binding pose metadynamics measures the persistence of ligand-protein interactions and the movements of the ligand's heavy atoms, which are sampled during variation of the complex's free energy states throughout the simulation. The result of binding pose metadynamics is a metadynamics-composition score. This composition score was calculated for the top 100 compounds (based on docking score) from the

benchmark set of each included kinase. The docking scores from which these top 100 compounds were chosen were derived from a single protein structure per target to allow for easy and direct comparison. The selected protein structures had the best actives enrichment based on docking and included 2IVU for RET, 2BMC for AURKA, and 4EQC for PAK1. The composition score was added to the docking score, resulting in a combined score. The actives enrichment for the targets was re-evaluated using this new score. It was observed that performance (based on the top 100 compounds) of PAK1 did not increase (ROC+BEDROC difference 0.01). However, actives enrichment for RET and AURKA increased with performances (ROC+BEDROC) 1.27 for RET and 1.76 for AURKA, compared to initial (docking) performance of 1.26 for RET and 0.77 for AURKA. Therefore, for RET and AURKA the docking scores of virtual screening compounds were rescored with metadynamics-composition scores, resulting in an indirect reranking of the top 100 virtual screening compounds, which was based on Z2 scores.

Polypharmacology Ranking of Compounds. The virtual screening compounds were scored with both a statistical model score (PCM score) and structure-based score. A final compound rank per target was constituted by adding a weight to the predictions made by statistical models and structure-based models (Table 2). The ranking order of compounds for off-targets, that is, kinase compounds should not interact with, was reversed: ranking compounds with high predicted activity as lowest, and compounds with worst predicted activity as highest. The weights of the statistical PCM model were based on the ROC score corrected for the size of the training data set per target. The structure-based models were considered to be better suited to select novel chemistry and were therefore assigned higher weights than the PCM model weights: structure-based models were attributed more weight compared to equally performing PCM models (same ROC). The weights of the structure-based models were calculated by taking the sum of BEDROC+ROC. These weights were reduced by penalizing the models when induced-fit structures were used and when the numbers of compounds in the benchmark sets were insufficient (for details see Methods). The weights were subsequently used to rank the compounds per target: Structure-based Z2-scores were multiplied by the structure-based weight, PCM predicted class probability was multiplied by the statistical model weight, and as final step the derived scores were summed up to retrieve a final rank per target.

On the basis of the overall target weights (Table 2), predictions for kinases LKB1, ERK8, PAK3, TAK1, and

PIK3CA, were not very reliable. However, predictions for the more important main on-targets RET, AURKA, and PAK1 were considered to be reliable. The workflow was evaluated by the selection and experimental validation of compounds for one of the main on-targets: RET. The consensus approach was used to select 15 highly ranked virtual screening compounds for RET (all compounds had PCM score >0.4 and structure-based score $Z2 < 0$), which were validated in vitro. Moreover, the performance of the different approaches was compared on the RET models by additionally selecting compounds based on the predictions of only statistical models, and only structure-based models.

Experimental Validation. The models were evaluated by experimental validation of predicted actives for RET kinase. A total selection of 46 compounds was purchased and validated for RET inhibition. These compounds were selected by different criteria: 6 compounds based on predicted activity by QSAR and PCM modeling, 25 compounds based on structure-based docking (including rescoring by binding pose metadynamics) and SPLIF scoring, and 15 compounds based on consensus scoring of statistical and structure-based models.

Compounds that were selected using only statistical models satisfied the set active-class threshold criteria of both QSAR (active probability > 0.6) and PCM (active probability > 0.5). The structure-based thresholds were docking score < -8 and SPLIF score >0.25 . Additionally, structure-based compounds with a docking score < -10 were selected that did not necessarily adhere to the SPLIF score criteria. Furthermore, compounds that were selected based on consensus scoring fitted the following thresholds: statistical predictions PCM > 0.4 , and structure-based score $Z2 < 0$. The compounds that were selected based on structure-based predictions or with consensus scoring were additionally inspected visually by checking the 3D docking pose and interaction with the hinge region.

The 46 compounds were first tested using single point measurements at two concentrations: 10 and 0.1 μM (Table S6). The top ten compounds showing RET inhibition (RET activity $< 80\%$, concentration 10 μM), were based on either consensus scoring or structure-based scoring. None of these compounds was from statistical scoring only. Five compounds showed inhibitory activity (RET activity $< 60\%$) for RET at a concentration of 10 μM , of which one compound also showed slight RET inhibition at concentration 0.1 μM (RET activity $< 100\%$). The activities of the four most potent hits were assessed more accurately by a potency determination in triplicate, yielding pIC_{50} values. The inhibitors were modestly active. ZINC33008650 was the best inhibitor with a pIC_{50} of 5.1 ± 0.1 , followed by ZINC72312837 (pIC_{50} 4.6 ± 0.2), ZINC12324934 (pIC_{50} 4.6 ± 0.2), and ZINC9518200 (pIC_{50} 4.0 ± 0.2). The docked pose of ZINC12324934 in RET suggests that the compound is an orthosteric binder with potential to bind to the hinge region, as evidenced by a hydrogen bond interaction with the backbone of Ala807 (Figure 3). Additional hydrogen bonds are observed between the ligand and Glu775, Ser891, and Lys728. The docked poses of all four inhibitors are included in Supplementary file S7. Although the inhibitors showed modest activity, their chemical diversity compared to known RET inhibitors was high since the compounds were prefiltered on novelty by selecting on Tanimoto similarity < 0.4 (compared to inhibitors with pChEMBL value ≥ 5).

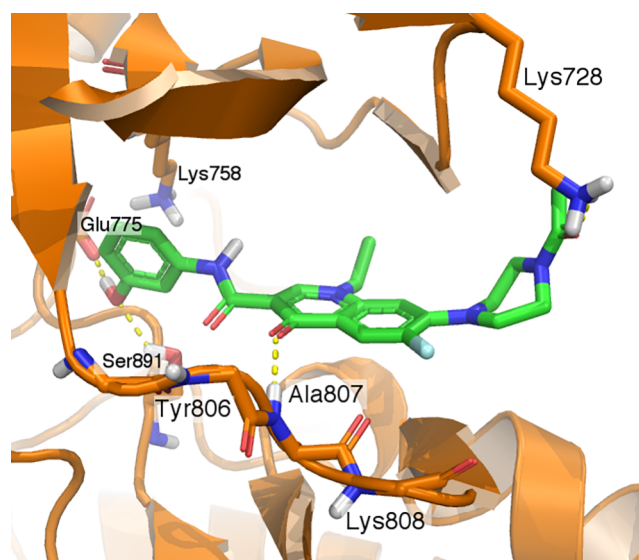


Figure 3. ZINC12324934 (green) docked into the orthosteric binding pocket of RET (orange) (PDB 2IVU). Hydrogen bonds are displayed as yellow dotted lines.

Bioactivity Spectra Prediction. The bioactivity profile of the five most active inhibitors for RET was predicted for all kinases in the panel. The bioactivity spectra for both on- and off-targets are shown in Table 3, where positive values indicate binding and values below and including zero indicate no binding. On the basis of the entire predicted bioactivity spectra, the most potent RET inhibitors comply better with inactivity on off-targets than activity on on-targets. However, the predictions are not equally reliable for each kinase (based on the previously assessed weights shown in Table 2). Nevertheless, assuming that the predictions of the kinases that scored equally well or better compared to RET (weight 1.98) are the most accurate, conclusions can be drawn for on-target SRC and off-target TTK (weights 2.46 and 2.20, respectively). Although the RET inhibitors were not predicted active for SRC, inactivity was predicted for off-target TTK. The compounds were selected based on bioactivity for RET, and as a result may not show optimal bioactivity spectra. However, the novelty filter to which the compounds were subjected prior to virtual screening included all of the kinases in the panel (as opposed to only RET). Therefore, all predicted interactions might indicate novel starting points for future research.

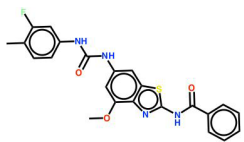
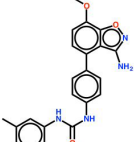
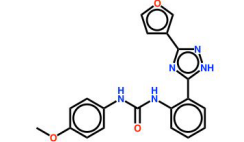
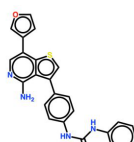
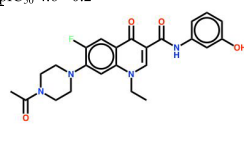
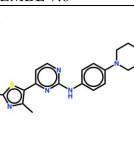
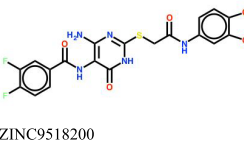
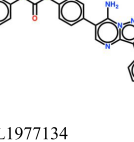
Manual Inspection of Hits. The four most potent hit compounds resulting from virtual screening and experimental validation were inspected based on their novelty and patentability. Novelty was re-evaluated by similarity searching (Tanimoto ECFP6) on a more strict threshold for RET compounds ($\text{pChEMBL} \geq 4$) in the most recent version of ChEMBL (version 25) and patentability was checked in SureChEMBL²⁵ (similarity > 0.9). The most similar RET actives were CHEMBL1979093 for ZINC33008650 (similarity 0.24, RET pChEMBL 7.2), CHEMBL1983715 for ZINC12324934 (similarity 0.18, RET pChEMBL 6.9), CHEMBL1977134 for ZINC9518200 (similarity 0.20, RET pChEMBL 8.4), and CHEMBL1965570 for ZINC72312837 (similarity 0.29, RET pChEMBL 7.6) (Table 4). None of the compounds was listed as patented in SureChEMBL.

Table 3. Predicted Bioactivity Spectra for the Panel 1 Kinases of the Five Most Potent RET Inhibitors

compound	on-target				off-target				
	RET	BRAF	SRC	S6K	MKNK1	TTK	ERK8 ^a	PDK1	PAK3 ^a
ZINC33008650	1	0	0	0	-2	-2	0	-2	0
ZINC12324934 ^b	0	0	0	0	0	0	0	0	0
ZINC9518200	2	0	0	0	-1	-1	0	-1	0
ZINC72312837	2	0	0	0	-1	-1	0	-2	0
ZINC65184824	1	-1	0	0	-2	-1	0	-2	0

^aPredictions based on limited structure-based data (no crystal structure available). ^bCompound was selected on RET docking score only; therefore, no Z2 score was available for this compound and no structure-based weight could be assigned.

Table 4. Novel RET Inhibitors and Their Most Similar RET Actives in ChEMBL Based on Tanimoto (ECFP6)

Compound	Closest similar (pChEMBL ≥ 4)	Similarity (Tanimoto or ECFP6)
 ZINC33008650 pIC ₅₀ 5.1 ± 0.1	 CHEMBL1979093 RET pChEMBL 7.2	0.24
 ZINC72312837 pIC ₅₀ 4.6 ± 0.2	 CHEMBL1965570 RET pChEMBL 7.6	0.29
 ZINC12324934 pIC ₅₀ 4.6 ± 0.2	 CHEMBL1983715 RET pChEMBL 6.9	0.18
 ZINC9518200 pIC ₅₀ 4.0 ± 0.2	 CHEMBL1977134 RET pChEMBL 8.4	0.20

Additionally, novelty of the chemical scaffolds was checked, which were compared based on Bemis-Murcko²⁶ scaffold trees with known active compounds for RET in ChEMBL. None of the four hit compounds was clustered in the same scaffold cluster as a known active, supporting the novelty of the inhibitors. Although the hit compounds were identified as modestly active binders, further SAR around these inhibitors may yield more potent derivatives. The novel scaffolds may be explored along different vectors to enhance affinity of the ligands and may reveal a new chemical space for RET inhibitors.

DISCUSSION

An elaborate virtual screening process was constituted in which both statistical and structure-based models were applied that were all carefully tuned and validated. Statistical models were validated using as many compound data as possible, with

attention paid to chemical diversity of the subsets in cross-validation. Structure-based models were slower than the statistical models and therefore smaller benchmark sets were used for validation of these models. These benchmark sets were composed of diverse actives to cover a large chemical space, irrespective of the smaller size of the data set. By application of statistical QSAR models and structure-based docking successively, computational time was used efficiently in virtual screening: compounds that had the least predicted activity probability for the main on-targets based on the QSAR models were excluded from docking. The consensus scoring approach, in which statistical PCM scores were combined with structure-based predictions, resulted in sets of active-predicted compounds per kinase. However, compounds were excluded from bioactivity modeling if the respective kinase had a poor reliability weight. This reliability weight was based on the performances per model per target derived from the benchmarking steps.

It was observed that the lack of compound data resulted in lower performance of statistical models in particular, as seen for kinase ERK8. Although structure-based models also performed worse for targets with less compound data and less crystal structures, this could partially be resolved by creation of additional protein structures using induced-fit docking and alternative scoring with SPLIF and Z2. However, alternate structures could also have been created using more elaborate ligand-protein sampling methods such as molecular dynamics.²⁷ Interestingly, performance between different structures of the same kinase varied greatly, rendering the benchmarking of multiple protein structures necessary.

On the basis of the final reliability scores, kinases RET, BRAF, SRC, MKNK1, TTK, PDK1, AURKA, PAK1, and FGFR1, are best used in bioactivity modeling, whereas predictions for kinases S6K, ERK8, PAK3, LKB1, TAK1, and PIK3CA, may not be accurate enough and thus should be excluded for now. Although the models of the first kinases are considered reliable, identification of compounds that fit the desired bioactivity profiles on these targets remains challenging. Since hit rates are rarely 100%, compounds with a well-predicted profile would need experimental validation. The hit rate for the experimental validation for RET in this study was 11% over 46 compounds, using pIC₅₀ ≥ 4 (or RET activity < 60% at a concentration of 10 μ M) as a threshold for active compounds. On the hypothesis that similar hit rates will be achieved for the other kinases, the probability of a compound being active on a number of kinases thus decreases with every additional target. Therefore, the number of experimentally validated compounds should be expanded, to increase the chance of finding a compound that fully fits the desired bioactivity profile.

The five hits that proved active on RET kinase were found with either consensus scoring or structure-based modeling. Although the number of compounds validated using statistical models is smaller, it is plausible that statistical modeling by itself is not a strong enough predictor to identify really novel compounds.²⁸ Since the virtual screening compounds were prefiltered on novelty (Tanimoto ECFP6 < 0.4) with known actives, the models were challenged with identification of chemical dissimilar actives.²⁹ This is a difficult task for statistical models as they rely on chemical patterns of known actives, and therefore, dissimilar compounds may lie outside the applicability domain of these models.²⁹ To resolve this issue, the chemical space can be expanded by addition of chemical diversity.⁸ Although this requires biological experiments to validate (in)activity of new compounds, an iterative screening approach may be applied to expand the chemical space efficiently and cost-effectively.³⁰ Nevertheless, identification of novel chemistry without the need for addition of much experimental data may be achieved through structure-based modeling, as these models are not particularly biased toward known chemistry. Yet, bias in structure-based models may indirectly be present, as the best performing protein structures were based on enrichment of known actives. Nevertheless, this is a minor issue compared to the relatively limited scope and bias of chemical space of statistical models.

The applied workflow in this study employs a filtering step through which compounds are discarded that were categorized as inactive by statistical QSAR models. On the basis of the previous statement on applicability domain, it is assumed good inhibitors can be wrongly categorized as “inactive” and neglected by the statistical model. One might reason that docking of all compounds may be a more effective method to identify novel inhibitors, as docking of 170 million compounds in proteins AmpC β -lactamase and the D4 dopamine receptor resulted in novel chemical scaffolds.³¹ However, docking of millions of compounds on multiple proteins is a very time-consuming task (~170 000 compounds per day for 1 RET crystal structure (24 cores)), which is unfeasible without significant computational resources. Thus, considering the scope of the task, the possibility that the QSAR models discard “good” compounds is accepted, as the QSAR models decrease the runtime of the workflow and make the task comprehensive.

As a final remark, it should be mentioned that the workflow applied to kinases did not implement orthosteric or allosteric binding of compounds. Therefore, inhibitors were not tuned to bind to the DFG-in or DFG-out kinase conformation, a conformational change that influences inhibitor binding greatly.³² Although structure-based docking was focused on the ATP-binding pocket, the most optimal crystal structures were selected based on benchmark sets that were not filtered for DFG-out and allosteric binders. As a consequence, crystal structures may have been selected that enriched DFG-out binders better than DFG-in binders. Moreover, since the statistical models were trained on sets that were not filtered for DFG-out binders, these models were also not able to distinguish DFG-in from DFG-out binders. As a result, it is undetermined whether the five hits from the screening workflow bind to the DFG-in conformation of RET. The docked poses of the hits do not constitute optimal hinge-binding, suggesting that it is plausible that they may be DFG-out binders. Moreover, two of the five most potent hit compounds contained a urea-motif, a motif that is often associated with DFG-out binders.^{32,33} To capture the binding

type of compounds, machine learning models could be used to predict the type of compound as an additional score, something that will be considered for future work.^{34,35}

CONCLUSION

An extensive workflow was designed to predict compound activity in kinases and to model the compounds' bioactivity spectra in kinases. The workflow can easily be expanded to other targets as well. By combining statistical and structure-based modeling, processing speed was optimized, while the accuracy of predictions was preserved. Every single kinase target was validated separately, which enabled reliability weights to be assigned to the predictions for every target. The workflow was experimentally validated by testing predictions made for RET kinase, a target with a good reliability score. A selection of 46 compounds was tested in vitro, of which 5 compounds showed RET inhibition (activity < 60%) at a concentration of 10 μ M. The four most potent inhibitors had pIC₅₀ values ranging from 4.0 to 5.1. The Tanimoto similarities (ECFP6) of these inhibitors with known RET actives was ≤ 0.29 . Moreover, the compounds contained unique chemical scaffolds, underscoring the true novelty of these inhibitors.

METHODS

Data Set Statistical Models. Training and validation sets for statistical models were constituted from compound information derived from ChEMBL (version 23),¹¹ publicly available sets from Eidogen,¹² and ExCAPE-DB.¹³ The compounds with experimental bioactivity for any kinase were filtered on molecular weight < 700, duplicates were removed, and compounds were standardized using BIOVIA Pipeline Pilot 2016.³⁶ Salts were removed, largest fragment was kept, stereochemistry and π -systems were standardized, and charges were neutralized. The resulting set contained 512 kinases and 123 246 bioactivities. For training and validation of the classification models the threshold for “active” compounds was set at pChEMBL/pK_i/pIC₅₀/pEC₅₀ ≥ 6.5 as we did previously.³⁷ Compounds that did not reach this threshold were termed “inactive”. All compounds per target were divided into five subsets using clustering with the Cluster Molecules component (FCFP4 clustering) in Pipeline Pilot 2016. Compounds were clustered into five clusters per each kinase family (same L2 level in ChEMBL). First, compounds were separated based on activity (active when pChEMBL > 6.5) and then clustered into five clusters, after which active and inactive compounds were combined again. To ensure that every panel kinase was represented in each cluster, clustering into five clusters was done in two steps: the panel kinase was clustered first, followed by coclustering of compounds from related kinases into the same cluster. The biggest cluster, or subset, was used as fixed training set, while the other four subsets were rotationally used as test or training set in 4-fold cross-validation. Data sets used in QSAR modeling only contained information on the respective kinase. In PCM modeling, additionally data of related kinases was included based on L4 classification in ChEMBL.

Data Set Structure-Based Docking. For each kinase target, with the exception of PAK3 and ERK8 because of lack of a crystal structure, a benchmark set was created to validate structure-based docking for each target. The benchmarking set for structure-based docking included active compounds,

inactive compounds, and decoy compounds, which were derived from ChEMBL (version 23) and the DUD-e Web server.¹⁵ Compounds with reported pChEMBL value for the challenge kinases, with confidence score 9, assay type B, and molecular weight <550, were standardized using BIOVIA Pipeline Pilot 2016³⁶ by removing salts and keeping the largest fragment. An activity gap was realized to better distinguish active and inactive compounds: compounds with a pChEMBL value >6.5 were assigned the label “active”, and compounds with a pChEMBL value >4 and <5.5 were labeled “inactive”. Compounds with pChEMBL value ≤ 4 were excluded as inactive compounds to limit the number of inactives. Exceptions in the thresholds were made for activity labeling of compounds for targets RET and MKNK1: for RET compounds were labeled as inactive when pChEMBL value >4.5 and <6.5, and for MKNK1 compounds were defined as active when pChEMBL value >6 to increase the fraction of active compounds. The thresholds used for RET were not intended to alter the number of (in)active compounds; RET was used as exploratory case and therefore for this kinase “initial” thresholds were used. However, the “general” threshold for all kinases was adapted since this yielded increased performance.

Scaffold trees were generated for the compounds based on Bemis–Murcko scaffolds²⁶ and the compound with the highest pChEMBL value per scaffold class, per activity class, per kinase, was kept. Compounds labeled as active were clustered into 100 clusters using the Cluster Molecules component (*k*-means) in BIOVIA Pipeline Pilot 2016. From the resulting clusters, only the cluster centers were kept. Kinases that had less than 100 active compounds available were excluded from this clustering step. The resulting active compound sets of maximum 100 compounds per kinase were used to generate decoys for each target kinase using the DUD-e Web server. For target PIK3CA decoys were generated based on only 80% of the active molecules as DUD-e decoy generation failed for the remaining fraction. The decoys were considered as inactive compounds when used in model validation. The collected actives, inactives, and decoys were prepared for docking using LigPrep from Schrödinger.²³

Screening Data Set. Compounds for virtual screening were extracted from the ZINC15 compound library.¹⁶ These compounds were selected based on “in stock” status and were filtered on a maximum of one Lipinski’s rule of five violation. Additionally, the screening set was filtered on novelty by only including compounds with Tanimoto similarity (ECFP6) < 0.4 compared to known “active” compounds (pChEMBL value ≥ 5) in ChEMBL (version 23) for the kinases of the respective panel. For panel 2, covering AURKA, PAK1, FGFR1, LKB1, PAK3, TAK1, and PIK3CA, the screening set was additionally filtered on PSA < 75 to allow permeability of the blood–brain barrier. The compounds in the screening data set were additionally prepared for structure-based docking by using LigPrep.

Statistical Modeling—QSAR. Single-target QSAR models were constructed for the main on-target kinases RET, AURKA, and PAK1. QSARs were built using BIOVIA Pipeline Pilot 2016 (version 18.0.1.1604). Models were trained on compound descriptors FCFP4 (3000 most frequent bits) and 86 physicochemical descriptors (S1). The following settings were applied in training categorical Random Forest QSAR models: 1000 trees, $\log_2(m)$ number of descriptors, equalized

class sizes, and seed 12345. The classification threshold for active compounds was set at pChEMBL > 6.5.

Statistical Modeling—PCM. PCM modeling was performed on all 512 kinases in the statistical modeling data set. A multitarget PCM model was constructed using a random forest classifier in scikit-learn.³⁸ Compound descriptors were the same as used in QSAR modeling. Protein descriptors were calculated based on full sequence alignment derived from kinase.com (with a total of 1567 alignment positions including gaps).^{39,40} The residues in the alignment were converted to three z-scales and an additional mean value for each of the three z-scales was added per sequence, resulting in a total of 4704 protein descriptors per kinase. Gaps were included in these descriptors, and were assigned a value of “0” for all three z-scales as was done previously.^{41,42}

The random forest model was of high complexity because of the high dimensionality of the data, which includes the compounds’ physicochemical properties, FCFP4 descriptors, and protein descriptors. The complexity of the random forest models was reduced by imposing constraints on parameters, such as the number of trees and maximum depth of the trees. The hyperparameters of the random forest model were optimized by utilizing random search, which evaluates the performance of the algorithm using different and randomly chosen configurations with cross-validation. Random search can be considered a simple form of automated machine learning^{19,43} and has been shown to outperform other basic methods of hyperparameter optimization, such as grid search.⁴⁴ Approximately 500 random configurations were evaluated, resulting in an improvement of on average 7% AUROC over the default configuration (300 trees, $\log_2(m)$ features at every node in the tree, no maximum tree depth, a minimum of 2 samples to consider a node for splitting, and bootstrap enabled) in 4-fold cross-validation. The settings that corresponded with the best model in random parameter optimization were applied in model training on the full data set, which was used in virtual screening. The final model consisted of 300 trees (fixed) with 43 features at every node in the tree, a maximum tree depth of 99, a minimum of 12 samples to consider a node for splitting, and bootstrapping disabled.

Structure-Based Docking. X-ray structures of all challenge proteins were extracted from the PDB⁴⁵ (except for ERK8 and PAK3 as no structure was available at the time). Crystal structures lacking cocrystallized orthosteric small-molecule ligands were discarded. The remaining structures were prepared for docking with the Protein Preparation tool from the Schrödinger 2017-4 suite after removing any other components than the protein, orthosteric ligand, and binding pocket ions. The “add missing side chains” option was used, waters were removed, hydrogens were added, and disulfide bonds were created. The crystal structures were superposed per target and cocrystallized ligands were removed from the binding site. The grid for docking was determined for each target by the center of one of the cocrystallized ligands (box size $xyz = 35$ Å). Compounds were docked into the binding pocket using the Schrödinger 2017-4 suite²³ and the OPLS3 force field⁴⁶ with standard precision (SP) and standard settings. A maximum of ten poses per compound (per target) was generated.

Induced-Fit Docking. Induced-fit docking, as implemented in the Schrödinger 2017-4 suite,²³ was applied to kinases LKB1, S6K, and MKNK1. These kinases showed poor

enrichment of actives when docked into the available crystal structures (sum of ROC and BEDROC < 1). For S6K five active compounds, for MKNK1 ten active compounds, and for LKB1 six active compounds were docked using induced-fit docking (see Table S8 for list of compounds). The crystal structures selected for induced-fit docking were 2HW6 for MKNK1, 2WTK for LKB1, and 3WF5, 3A62, and 4RLO for S6K. The resulting protein conformations were used as alternative protein structures in addition to the original crystal structures.

Structural Protein–Ligand Interaction Fingerprints (SPLIFs). Structural protein–ligand interaction fingerprints (SPLIFs)⁴⁷ were calculated for a maximum of ten poses per compound that were retrieved from structure-based docking. The cocrystallized ligands of the protein structures were used as reference in SPLIF calculations to derive Tanimoto-like SPLIF scores. Five protein structures per kinase were selected for SPLIF generation. These structures were selected based on diversity of their cocrystallized ligands and best active-enrichment based on docking scores. The diversity of the cocrystallized ligands was assessed using affinity propagation clustering based on FCFP6 similarity.⁶ One protein structure was selected from every corresponding cluster based on best docking score performance until a maximum of five protein structures was reached. Subsequently, Tanimoto-like SPLIF scores were calculated for the compounds in the benchmark sets for each of the selected protein structures. For each compound (maximum 10 poses) the best SPLIF score was used to calculate actives enrichment based on SPLIF scores.

For RET and PIK3CA exceptions were made in the selection of benchmark proteins for SPLIF because only four and three clusters were generated, respectively. Therefore, similar cocrystallized ligands and their corresponding proteins were also selected to get a total of five protein structures per target. Furthermore, for RET structure 2IVV instead of the cocrystallized ligand, ChEMBL3775169 was used as a reference.

Binding Pose Metadynamics. A metadynamics-composition score¹⁰ was calculated with the binding pose metadynamics tool in the Schrödinger 2017-4 suite,²³ for compounds and poses derived from structure-based docking. The top 100 compounds for RET (PDB 2IVU) and AURKA (PDB 2BMC) were selected, based on best docking scores. The protein structures were prepared for metadynamics simulations by capping the termini and the run time of each simulation was set at 10 ns. The resulting metadynamics-composition scores were added to the existing docking scores to rerank the top 100 compounds in the benchmark set and to rerank the top 100 compounds in the screening set.

Ensemble Scoring in Structure-Based Modeling. For every target, protein models for virtual screening were selected based on ensemble scoring of docking and SPLIF scores. The combination, or ensemble, of protein models that resulted in the best actives enrichment was used in virtual screening. The validated ensembles were based on docking scores of the top five enriching models and all five SPLIF models per target. Different ensembles were tested for each target including: docking scores only, SPLIF scores only, and docking and SPLIF scores combined. The performances of the ensembles were evaluated using Z2-scoring by averaging over the top two Z-scores.⁴⁷ Prior to ensembling, the Z-scores per compound were normalized toward the actives from the respective benchmark set. Docking scores were normalized to Z-scores

by subtracting the mean docking score (mean over docking scores from all actives in the benchmark set) from the docking score of the test compound, and subsequently dividing by the standard deviation of the (benchmark) actives' docking scores. The same approach was used in normalization of SPLIF scores to Z-scores. However, the SPLIF scores were first multiplied by -1 to change the vector into the same direction as the docking scores (better binder, more negative score).

Compound Ranking Per Target. The compounds in virtual screening were ranked for ensembles of targets based on the scores derived from PCM modeling and structure-based ensemble score. The PCM predictions were given a confidence weight between 0 and 1 based on the prediction performance per target (with 0 being no confidence and 1 being highest confidence). The confidence score for the PCM model could be calculated using the following equation:

$$\text{confidence weight}_{\text{PCM}} = (\text{ROC} * 2 - 1) * \frac{\sqrt{\text{number of samples}}}{\sqrt{\text{max number of samples}}}$$

The PCM confidence score was based on the number of compounds (of the corresponding target) in training and the ROC score derived from 4-fold cross-validation: $\text{ROC} * 2 - 1$ (to normalize that 0.5 corresponds with a weight of 0 and 1 corresponds with a weight of 1), multiplied by the weight based on the number of samples for the target in the training data (calculated as $\sqrt{\text{number of samples}} / \sqrt{\text{maximum number of samples}}$).

The weights of the structure-based predictions were calculated by taking the total sum of ROC and BEDROC per target,⁴⁸ consequently giving the structure-based predictions more weight than the statistical PCM model predictions. However, structure-based models were penalized (by multiplying with 0.5) when induced-fit models were created for the kinase. Moreover, kinases containing less than 100 actives in the benchmark set were penalized by multiplying the structure-based weight with the fraction of number of actives (with a fraction of 1 being 100 actives, and a fraction of 0.1 being 10 actives).

Finally, the obtained weights for PCM and structure-based models were applied in calculating the compound rank per kinase.

$$\text{compound rank} = \text{confidence weight}_{\text{PCM}} * \text{active class probability} + \text{confidence weight}_{\text{structure-based}} * - Z2 \text{ score}$$

Compound predictions were multiplied by the subtotal weights and summed up (statistical model weight * statistical PCM prediction + structure-based model weight * structure-based Z2 score), resulting in a final compound rank per target.

Protein Kinase Assay. A selection of 46 compounds was experimentally validated for inhibitory activity against RET. The compounds were ordered via Mcule Inc.⁴⁹ (Budapest) and purchased from ChemBridge, ChemDiv, ChemScene, Enamine, Life Chemicals, and Vitas M Chemical Limited. The assays were performed by ProQinase GmbH,⁵⁰ Germany. First, the compounds were tested ($n = 1$) at final assay concentrations 10 and 0.1 μM . The five resulting hits (<60% remaining RET activity at concentration 10 μM) were re-evaluated by determination of IC_{50} values ($n = 3$). RET WT

activity was measured using a radiometric protein kinase assay (³³PanQinase Activity Assay). All kinase assays were performed in 96-well FlashPlates from PerkinElmer (Boston, MA, USA) in a 50 mL reaction volume. The reaction cocktail was pipetted in 4 steps in the following order: (1) 20 mL of assay buffer, (2) 5 mL of ATP solution (in H₂O), (3) 5 mL of test compound (in 10% DMSO), and (4) 20 μL enzyme/substrate mix. The assay for all protein kinases contained 70 mM HEPES–NaOH pH 7.5, 3 mM MgCl₂, 3 mM MnCl₂, 3 mM Na-orthovanadate, 1.2 mM DTT, 50 μg/mL PEG20000, 1 μM ATP, [γ -³³P]-ATP (approximately 1.91 × 10⁰⁵ cpm per well), 40 ng/50 μL of protein kinase, and 0.125 μg/50 μL of poly(Glu, Tyr)4:1 substrate. The reaction cocktails were incubated at 30 °C for 60 min. The reaction was stopped with 50 mL of 2% (v/v) H₃PO₄; the plates were aspirated and washed two times with 200 mL 0.9% (w/v) NaCl. Incorporation of ³³Pi was determined with a microplate scintillation counter (Microbeta, Wallac).

■ ASSOCIATED CONTENT

■ Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.9b01204>.

- List of 86 physicochemical descriptors (XLSX)
- Structure-based benchmark performance for crystal structures of every panel kinase (XLSX)
- Selected structure-based models per target (XLSX)
- Bioactivities of 46 compounds validated for RET activity (XLSX)
- Docked poses of the four most potent hit compounds (ZIP)
- List of compounds used to generate structures for S6K, MKNK1, and LKB1, with induced-fit docking (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Gerard J. P. van Westen – Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands; orcid.org/0000-0003-0717-1817; Email: gerard@lacdr.leidenuniv.nl

Authors

Lindsey Burggraaff – Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands; orcid.org/0000-0002-2442-0443

Eelke B. Lenselink – Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands

Willem Jespers – Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands; Department of Cell and Molecular Biology, Uppsala University, Uppsala 75124, Sweden; orcid.org/0000-0002-4951-9220

Jesper van Engelen – Department of Computer Science, Leiden Institute of Advanced Computer Science, Leiden University, 2333 CA Leiden, The Netherlands

Brandon J. Bongers – Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands

Marina Gorostiola González – Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands

Rongfang Liu – Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands

Holger H. Hoos – Department of Computer Science, Leiden Institute of Advanced Computer Science, Leiden University, 2333 CA Leiden, The Netherlands

Herman W. T. van Vlijmen – Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands; Janssen Research & Development, 2340 Beerse, Belgium; orcid.org/0000-0002-1915-3141

Adriaan P. IJzerman – Division of Drug Discovery & Safety, Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jcim.9b01204>

Author Contributions

L.B. wrote the manuscript. L.B., E.L., and W.J. performed the structure-based modeling. J.v.E. constructed the PCM models. B.J.B. and M.G.G. constructed QSAR models. R.L. assisted in the examination of the biological experiments. L.B., G.J.P.v.W., H.H.H., H.W.T.v.V., and A.P.IJ. contributed to the discussion of the work. All authors have read and approved the final version of the manuscript.

Notes

The authors declare no competing financial interest. Standardized data set of 512 kinases used in statistical modeling is available at DOI: [10.4121/uuid:6af1d9de-281f-4221-b7e1-e7c01b90dfe0](https://doi.org/10.4121/uuid:6af1d9de-281f-4221-b7e1-e7c01b90dfe0). Protein structures for LKB1, MKNK1, and S6K, which were constructed using induced-fit docking are available at DOI: [10.4121/uuid:9e61b6a6-88e5-4a18-ba19-dbf1bbdd656a](https://doi.org/10.4121/uuid:9e61b6a6-88e5-4a18-ba19-dbf1bbdd656a).

■ ACKNOWLEDGMENTS

G.J.P.v.W. thanks the Dutch Scientific Council (NWO) and Applied and Engineering Sciences (AES) for funding (VENI no. 14410).

■ ABBREVIATIONS

BEDROC, Boltzmann-enhanced discrimination of the receiver operating characteristic; ECFP, extended-connectivity bit string; FCFP, feature-connectivity bit string; MCC, Matthews correlation coefficient; PCM, proteochemometrics; QSAR, quantitative structure–activity relationship; ROC, receiver operating characteristic; SPLIF, structural protein–ligand interaction fingerprints

■ REFERENCES

- (1) de Lera, A. R.; Ganesan, A. Epigenetic Polypharmacology: From Combination Therapy to Multitargeted Drugs. *Clin. Epigenet.* **2016**, *8*, 105.
- (2) van Leeuwen, R. W. F.; Jansman, F. G. A.; van den Bemt, P. M. L. A.; de Man, F.; Piran, F.; Vincenten, I.; Jager, A.; Rijnveld, A. W.; Brugma, J. D.; Mathijssen, R. H. J.; van Gelder, T. Drug–Drug Interactions in Patients Treated for Cancer: A Prospective Study on Clinical Interventions†. *Ann. Oncol.* **2015**, *26*, 992–997.
- (3) Schlessinger, A.; Abagyan, R.; Carlson, H. A.; Dang, K. K.; Guinney, J.; Cagan, R. L. Multi-Targeting Drug Community Challenge. *Cell Chem. Biol.* **2017**, *24*, 1434–1435.

- (4) Dar, A. C.; Das, T. K.; Shokat, K. M.; Cagan, R. L. Chemical Genetic Discovery of Targets and Anti-Targets for Cancer Polypharmacology. *Nature* **2012**, *486*, 80.
- (5) Sonoshita, M.; Scopton, A. P.; Ung, P. M. U.; Murray, M. A.; Silber, L.; Maldonado, A. Y.; Real, A.; Schlessinger, A.; Cagan, R. L.; Dar, A. C. A Whole-Animal Platform to Advance a Clinical Kinase Inhibitor into New Disease Space. *Nat. Chem. Biol.* **2018**, *14*, 291.
- (6) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (7) van Westen, G. J. P.; Wegner, J. K.; Ijzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. *MedChemComm* **2011**, *2*, 16–30.
- (8) Burggraaff, L.; Oranje, P.; Gouka, R.; van der Pijl, P.; Geldof, M.; van Vlijmen, H. W. T.; Ijzerman, A. P.; van Westen, G. J. P. Identification of Novel Small Molecule Inhibitors for Solute Carrier SGLT1 Using Proteochemometric Modeling. *J. Cheminf.* **2019**, *11*, 15.
- (9) Christmann-Franck, S.; van Westen, G. J. P.; Papadatos, G.; Beltran Escudie, F.; Roberts, A.; Overington, J. P.; Domine, D. Unprecedentedly Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design? *J. Chem. Inf. Model.* **2016**, *56*, 1654–1675.
- (10) Clark, A. J.; Tiwary, P.; Borrelli, K.; Feng, S.; Miller, E. B.; Abel, R.; Friesner, R. A.; Berne, B. J. Prediction of Protein–Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations. *J. Chem. Theory Comput.* **2016**, *12*, 2990–2998.
- (11) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.
- (12) *Eidogen Sertanty*. <http://www.eidogen-sertanty.com/> (accessed Jan 22, 2018).
- (13) Sun, J.; Jeliakova, N.; Chupakhin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliakov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. *J. Cheminf.* **2017**, *9*, 17.
- (14) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 885–896.
- (15) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.
- (16) Sterling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
- (17) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.
- (18) 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*, D1083–90.
- (19) Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
- (20) Da, C.; Kireev, D. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *J. Chem. Inf. Model.* **2014**, *54*, 2555–2561.
- (21) Frey, B. J.; Dueck, D. Clustering by passing Messages Between Data Points. *Science (Washington, DC, U. S.)* **2007**, *315*, 972–976.
- (22) Sastry, G. M.; Inakollu, V. S. S.; Sherman, W. Boosting Virtual Screening Enrichments with Data Fusion: Coalescing Hits from Two-Dimensional Fingerprints, Shape, and Docking. *J. Chem. Inf. Model.* **2013**, *53*, 1531–1542.
- (23) Schrödinger. *Schrödinger Maestro*, release 2017-4; Schrödinger, LLC: New York, NY, 2017.
- (24) Tiwary, P.; Parrinello, M. A Time-Independent Free Energy Estimator for Metadynamics. *J. Phys. Chem. B* **2015**, *119*, 736–742.
- (25) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2016**, *44*, D1220–8.
- (26) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (27) Atzori, A.; Bruce, N. J.; Burusco, K. K.; Wroblowski, B.; Bonnet, P.; Bryce, R. A. Exploring Protein Kinase Conformation Using Swarm-Enhanced Sampling Molecular Dynamics. *J. Chem. Inf. Model.* **2014**, *54*, 2764–2775.
- (28) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
- (29) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.
- (30) Paricharak, S.; Ijzerman, A. P.; Bender, A.; Nigsch, F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-House HTS Data. *ACS Chem. Biol.* **2016**, *11*, 1255–1264.
- (31) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O’Meara, M. J.; Che, T.; Alga, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566*, 224–229.
- (32) Liu, Y.; Gray, N. S. Rational Design of Inhibitors That Bind to Inactive Kinase Conformations. *Nat. Chem. Biol.* **2006**, *2*, 358–364.
- (33) Vijayan, R. S. K.; He, P.; Modi, V.; Duong-Ly, K. C.; Ma, H.; Peterson, J. R.; Dunbrack, R. L., Jr; Levy, R. M. Conformational Analysis of the DFG-out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors. *J. Med. Chem.* **2015**, *58*, 466–479.
- (34) Miljković, F.; Rodríguez-Pérez, R.; Bajorath, J. Machine Learning Models for Accurate Prediction of Kinase Inhibitors with Different Binding Modes. *J. Med. Chem.* **2019**, DOI: 10.1021/acs.jmedchem.9b00867.
- (35) van Westen, G. J. P.; Gaulton, A.; Overington, J. P. Chemical, Target, and Bioactive Properties of Allosteric Modulation. *PLoS Comput. Biol.* **2014**, *10*, No. e1003559.
- (36) Dassault Systèmes BIOVIA. *BIOVIA Pipeline Pilot*; Dassault Systèmes: San Diego, 2016.
- (37) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; Ijzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using A ChEMBL Bioactivity Benchmark Set. *bioRxiv*, 2017. <https://www.biorxiv.org/content/10.1101/168914v1>.
- (38) 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. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
- (39) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science (Washington, DC, U. S.)* **2002**, *298*, 1912–1934.
- (40) <http://kinase.com> (accessed May 8, 2017).
- (41) van Westen, G. J.; Swier, R. F.; Wegner, J. K.; Ijzerman, A. P.; van Vlijmen, H. W.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 1): Comparative Study of 13 Amino Acid Descriptor Sets. *J. Cheminf.* **2013**, *5*, 41.
- (42) van Westen, G. J. P.; Swier, R. F.; Cortes-Ciriano, I.; Wegner, J. K.; Overington, J. P.; Ijzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 2): Modeling Performance of 13 Amino Acid Descriptor Sets. *J. Cheminf.* **2013**, *5*, 42.

(43) Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; Hutter, F. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015; pp 2962–2970.

(44) Kotthoff, L.; Thornton, C.; Hoos, H. H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA 2.0: Automatic Model Selection and Hyperparameter Optimization in WEKA. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.

(45) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(46) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281–296.

(47) Lenselink, E. B.; Jespers, W.; van Vlijmen, H. W. T.; Ijzerman, A. P.; van Westen, G. J. P. Interacting with GPCRs: Using Interaction Fingerprints for Virtual Screening. *J. Chem. Inf. Model.* **2016**, *56*, 2053–2060.

(48) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488.

(49) *mcule*. <https://mcule.com> (accessed Jun 13, 2019).

(50) ProQinase GmbH. *ProQinase* <https://www.proqinase.com> (accessed Aug 22, 2019).