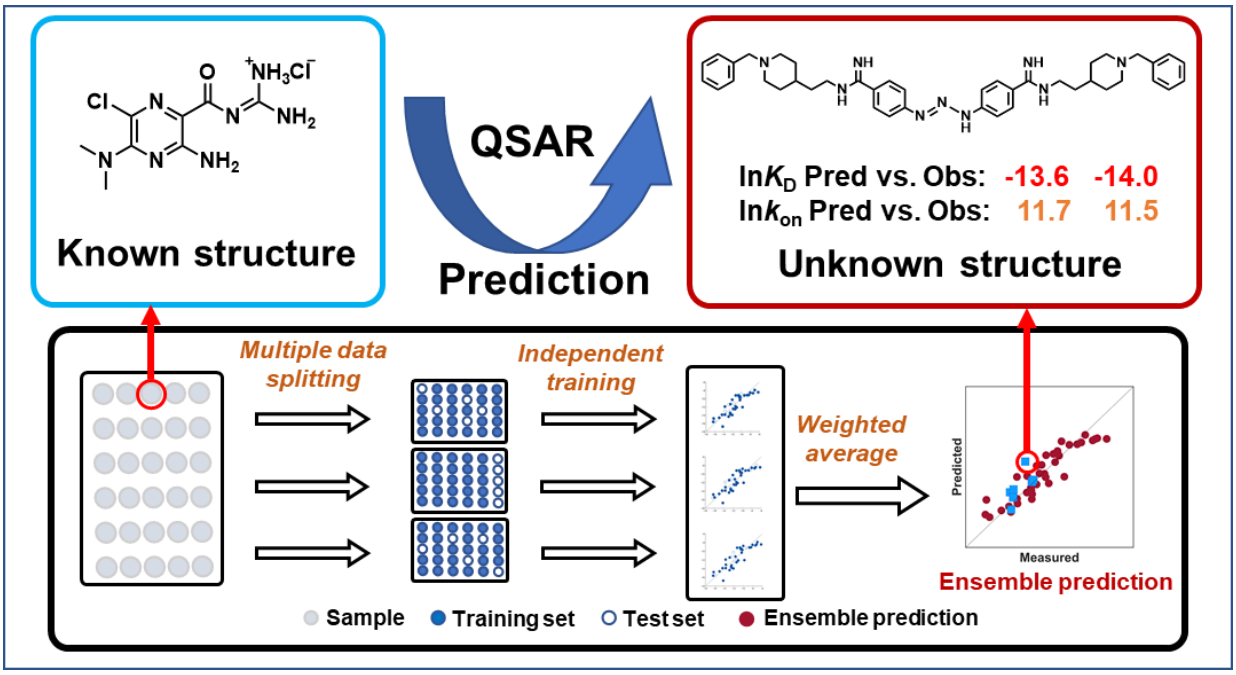
Ensemble learning-based quantitative structure-activity relationship platform predicts binding behavior of RNA-targeted small molecules

Zhengguo Cai; Martina Zafferani; Amanda E. Hargrove*

Department of Chemistry, Duke University, 124 Science Drive, Durham, NC 27705, USA

*Corresponding Author; Contact: amanda.hargrove@duke.edu

TOC



Abstract

The diversity of RNA structural elements and their documented role in human diseases make RNA an attractive therapeutic target. However, progress in drug discovery and development has been hindered by a limited understanding of the parameters that drive RNA recognition by small molecules, including a lack of experimentally validated structure-activity relationships (QSAR). We developed an adaptable ensemble learning-based method that quantitatively predicts both affinity and kinetic-based binding parameters of small molecules against the HIV-1 TAR model RNA system. A training set of small molecules was screened against the HIV-1-TAR construct using surface plasmon resonance, which provided the binding kinetics and affinities. Introduction of ensemble learning on these data combined with structure-based molecular descriptors afforded

predictive models as well as explicit interpretation of the contributing parameters. The accuracy of the model was tested by external validation where binding properties of additional molecules outside training set were correctly predicted. The ensemble model presented herein is the first application of predictive and experimentally validated 2D-QSAR against an RNA target, in this case HIV-1-TAR RNA, and provides a platform to guide future synthetic efforts. Furthermore, we expect the workflow described herein to be applicable to other RNA structures, ultimately providing essential insight into the small molecule descriptors that drive selective binding interactions and, consequently, exponentially increasing the efficiency of ligand design and optimization without the need for high-resolution structures.

Keywords
RNA, small molecules, quantitative structure-activity relationship, rational ligand design, ensemble learning, model interpretation

Introduction

Initiated in 2003, the ENCODE project[1] revealed an unprecedented number of non-protein-coding RNAs (ncRNAs), and their roles in the regulation of transcription, translation, genetic modification and RNA degradation have been subject of intense study in relation to human disease.[2] ncRNAs have been found to be abnormally expressed in multiple disease phenotypes, including neurodegenerative diseases and metastatic cancers.[3-12] The implications of these RNAs in disease pathogenesis underscore their potential roles as drug targets. To date, various ncRNAs have been targeted by small molecules across many species such as mammals, viruses, bacteria, and fungi.[13-17] These RNAs include the bacterial ribosome, HIV leader sequence in the 5' untranslated region (UTR), pre-miRNA in nuclease processing site, Huntington's disease related r(CAG) exonic repeats, and alternative splicing site for spinal muscular atrophy (SMA).

While RNA is an attractive therapeutic target, some RNA properties pose intrinsic challenges including: 1) limited chemical diversity of RNA relative to proteins; 2) the highly negatively charged backbone of RNA, and 3) the dynamic nature of RNA, which allows it

to sample a wide population of conformers. In particular, the diverse and complex conformational dynamics of RNA increase the complexity of RNA structure determination, including that of RNA:ligand structures, ultimately hindering the development of predictive binding models as well as our understanding of the drivers of small molecule:RNA recognition. Currently, the most successful discovery method for bioactive RNA-targeted small molecules has been focused screens, which require synthetic library curation based on prior knowledge of the biased chemical space of RNA-targeted small molecules.[18] Additionally, the current paradigm of RNA-targeted small molecule often disregards binding kinetics, precluding a full understanding and optimization of binding behaviors of a compound. Indeed, many marketed drugs are characterized with slow dissociation processes and prolonged target occupancy, indicating the significance of binding kinetics in evaluating *in vivo* activity.[19] The design of compounds with kinetic selectivity will open a new avenue for RNA targeting and facilitate the hit-to-lead triage during hit optimization,[20, 21] though very few studies have so far demonstrated how to intentionally optimize RNA binding kinetics.[22] Overall, there are clear unmet needs in finding potential RNA-targeted chemical probes beyond screening and to rationally design small molecules with desired binding behaviors, including appropriate binding kinetics.

To fully access the numerous potentially-druggable RNA targets discovered through the 'RNA revolution', a rational tool for ligand design and comprehensive understanding of RNA:small molecule binding details is required. Recently, machine learning-aided mechanistic studies and ligand predictions have shown success in multiple complex tasks, including the design of enantioselective catalysts in organic synthesis and bioactive ligands for kinase inhibition.[23-26] Significant work has been done to explore key descriptors involved in RNA recognition.[27-29] Most of this work has utilized publicly available data, which offers a large amount of data but is limited by the diversity and inconsistency of screening methods employed. In addition, if the data originates from multiple RNA targets, only general guiding principles can be derived, which is not sufficient for precise ligand design for a specific RNA target. Among multiple computational tools, quantitative structure-activity relationship (QSAR) study can pinpoint guiding principles for a specific target by correlating the experimentally observed binding

properties with the molecular descriptors of the ligands.[30-32] A robust and predictive QSAR model has been proven to be an efficient tool to predict activities of small molecule candidates and to drive hit optimization. Despite its success in protein-based ligand design, however, few QSAR studies have been conducted for identifying RNA-targeted small molecules.[33-35]

Herein, we built a systematic workflow utilizing QSAR as an intermediate to connect molecular descriptors of a given ligand with its binding profiles against a specific RNA. The activities, including binding affinity (K_D) and kinetic rate constants (k_{on} and k_{off}), were measured for various molecules via surface plasmon resonance (SPR). To the best of our knowledge, this constitutes the first example of a systematic empirical QSAR study conducted for a specific RNA target; consequently, we considered several new strategies to build the model. To overcome the potential bias related to diverse scaffolds, we applied an ensemble strategy to model training. Ensemble models were trained for explaining binding behaviors with low root-mean-square-error (RMSE) and high coefficient of determination (R^2) and applied to yield a novel quantitative structure-kinetic relationship for RNA ligands. External validation using untested small molecules verified the accuracy of prediction of the model built on the training set. Importantly, our model interpretation derived significant descriptors for binding kinetics and affinity, plotting the landscape with which improvement of binding behaviors could be achieved via intentional structural modifications. We anticipate that this framework could be readily extended to different RNA contexts to facilitate the design and synthesis of novel RNA-targeted ligands. The workflow built in this study will contribute to improving the understanding of RNA:small molecule binding mechanisms and provide an efficient tool to rationally design new ligands for a given RNA target.

Results and Discussion

We chose the HIV-1 transactivation response (TAR) element (**Figure 1a**) as a suitable model system to develop our workflow as this well-validated antiviral target has been frequently screened against small molecules, providing us with numerous candidates for the training set.[36-39] Forty reported TAR ligands or compounds with known RNA-targeted scaffolds were selected for model construction. Overall, these ligands could be classified into 4 categories, namely aminoglycosides (AGs), dimethyl amilorides[40, 41] (DMAs), diphenyl furans[42, 43] (DPFs) and nucleic acid dyes (**Figure 1a**). These ligands covered a range of binding behaviors with the aim of building a model that can be applied to the prediction of ligands with diverse chemical architecture.

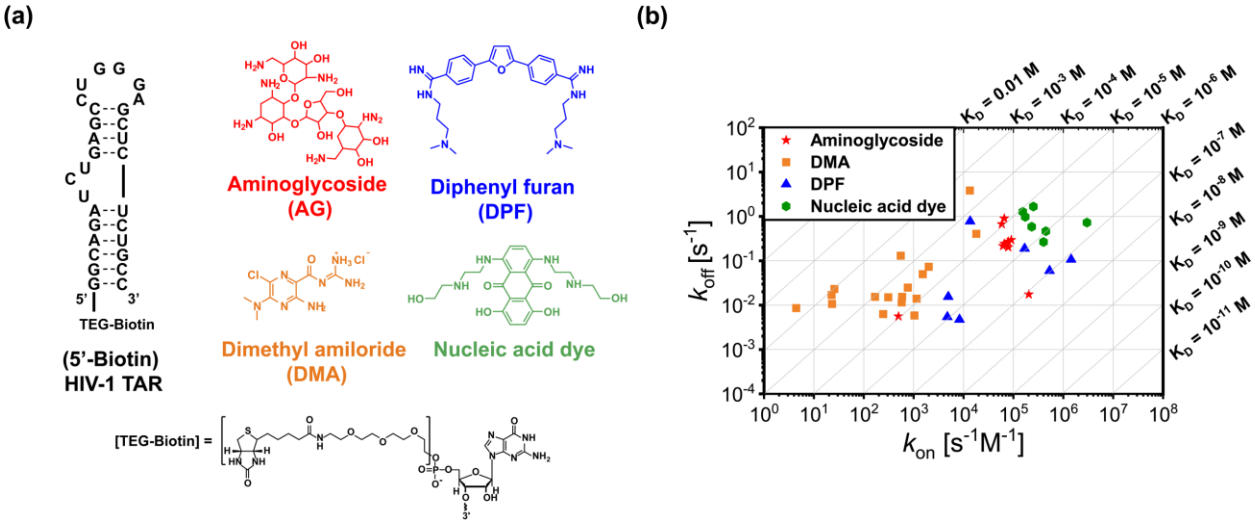
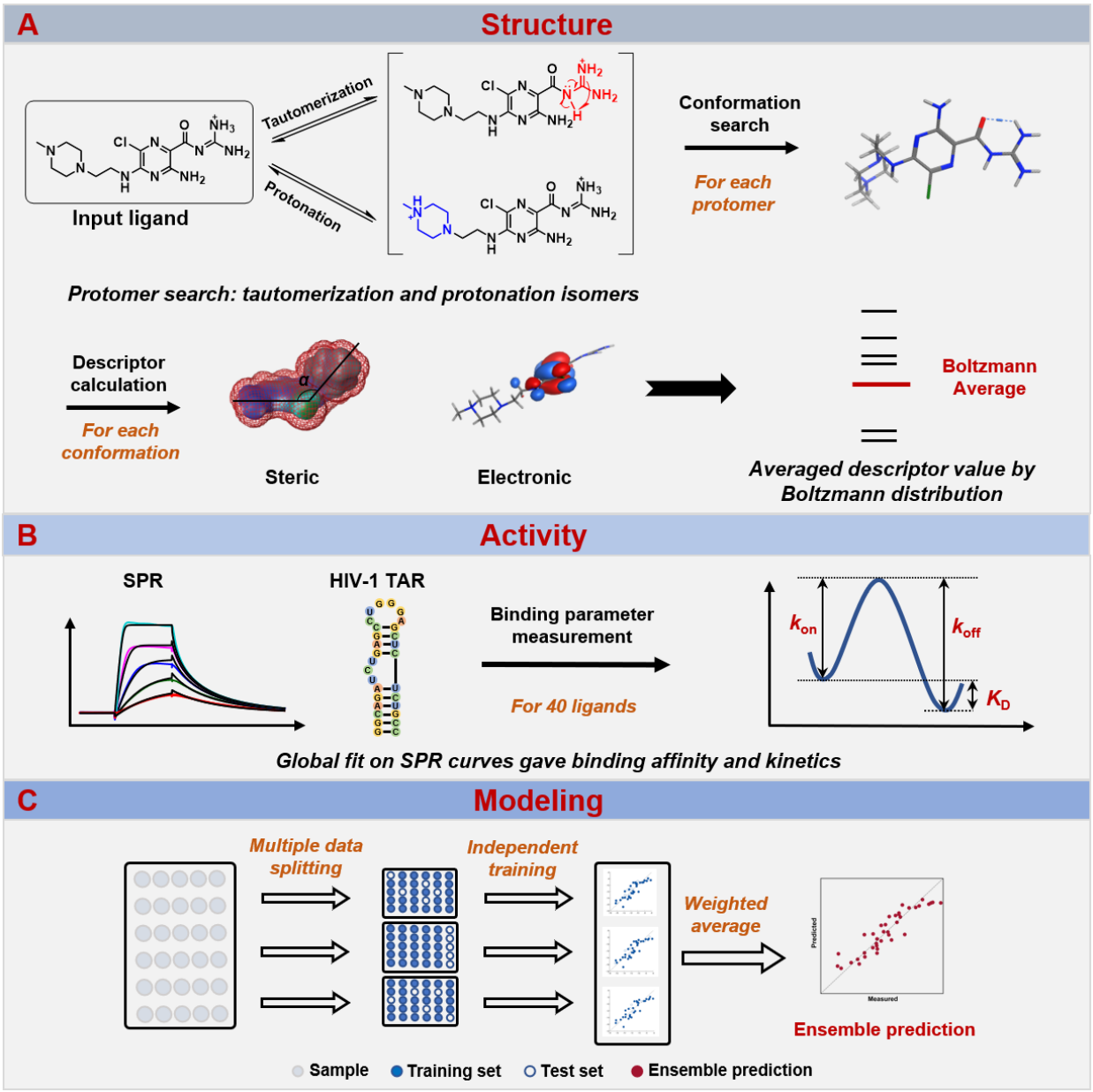


Figure 1 A. Sequence and structure of 5' biotinylated HIV-1 TAR and chemical structures of the scaffolds used in this work. **B**. Kinetics map of 40 tested ligands, represented on the 10-based logarithmic coordinates. The diagonal lines represent K_D values calculated from k_{off}/k_{on}. Units of three parameters are shown. The rest of study used values based on these units.

1. Calculations of molecular descriptors

We started this workflow by obtaining molecular information of each compound via quantitative calculation of their molecular descriptors. Each descriptor provides information on a physiochemical property of a compound, ranging from topological to electrostatic terms. For example, atomic connectivity that represents topological connections within a molecule was calculated upon the use of graph theory matrices, which lays the foundation of many other descriptors including related adjacency and distance matrices as well as chemical identity and hybridization states. According to

previous reports, many QSAR expressions suggest that ligand binding preferences originate from non-covalent interactions exerted in the micro-space of the ligand.[44] Hence conformation-dependent 3D descriptors were included to account for the spatial environment of the ligands, such as partial charges and potential energy. As demonstrated in **Panel A**, **Scheme 1**, to ensure the accurate calculation of these molecular descriptor values, we considered whether multiple species of a given molecule may exist at experimental conditions. Namely, many of the RNA-targeted ligands in the training set are predicted to be positively charged under the buffered condition (pH=7.4 in SPR), thus protonation and tautomerization states were considered for each ligand and described by a distribution coefficient. For each ligand state, potential conformations within 3 kcal/mol of the lowest energy conformation, as determined using the Molecular Operating Environment (MOE) software, were selected. The descriptor value of a specific ligand state was determined as the Boltzmann-weighted average of these conformations. Finally, the descriptor value of each ligand is the weighted average of the results from multiple states based on distribution coefficients mentioned before. In total, we calculated 435 descriptors of each ligand. For model training, we selected 304 by deleting descriptors that had more than 25% repeated entries since such descriptors would cause fitting error during model searching. While the presence of multiple species and/or conformations is often overlooked due to computational cost, accuracy of molecular descriptors is a prerequisite for reliable and robust QSAR models.



Scheme 1 Workflow of ensemble QSAR. **Structure**: input molecules were searched on their "protomers" and then searched on conformations of each protomer. Molecular descriptors were calculated for each conformation and averaged based on Boltzmann distribution. **Activity**: small molecules binding HIV-1 TAR were characterized via SPR and parameters including K_D, k_{on} and k_{off} were fitted globally. **Modeling**: with multiple data splitting and independent model training, the final prediction is given by the averaged predictions from multiple learners followed by model interpretation.

2. Measurement of binding parameters

To evaluate the binding parameters of the small molecules against HIV-1 TAR, we utilized SPR to measure the kinetic rate constants and binding affinities. Kinetic analyses for the observed SPR curves were performed globally for the entire concentration series (**Panel B**, **Scheme 1**). To ensure the robustness of the experimental data, we conducted multiple tests for each ligand to reach consistency between measurements. The kinetics map summarizes the distribution of k_{on}, k_{off} and K_D along logarithmic coordinates (**Figure 1b**). All three parameters have a wide range of values spanning at least 2 log units, supporting the appropriateness for reliable QSAR modeling from a response variable perspective.[45]

To validate our kinetics data, we compared our results to a previous survey that showed RNA ligand association was generally slower than that for protein.[46] The measured on and off rates values in our SPR data are similar in order of magnitude to the RNA:ligand values previously reported (**Table 1**).[46] The overall association rate constant of an RNA-ligand pair for all three RNA-ligand sets (median: ~10^4 $M^{-1}s^{-1}$) was not only far below the diffusion limit (centered at 10^9 $M^{-1}s^{-1}$) but also suggested a generally slower binding than protein-ligand pairs (median: 6.6×10^6 $M^{-1}s^{-1}$). This slow RNA recognition was expected due to the existence of multi-conformation distribution in unbound RNA states, though some variation was observed between ligand classes. Specifically, in our HIV-1 TAR-ligand set, most of the fast association rates were observed for aminoglycosides, nucleic acid dyes and DPFs (k_{on}: 10^4~10^5 $M^{-1}s^{-1}$), probably due to their strong electrostatic (aminoglycosides) or topologically matched pi-pi stacking interactions (dyes, DPFs). As moderate and weak binders in this set, DMAs were characterized by fewer potential protonation sites or less planar structure than other molecules, leading to overall slower binding rates. Rates of dissociation were comparable among the three RNA-ligand sets, with median values around 10^{-2} s^{-1}. Comparing binding strengths in **Table 1**, it was expected that RNA-ligand pairs with *in vitro* selected RNAs (e.g. aptamers) and naturally occurring RNAs that have evolved to bind small molecules (e.g. riboswitches and ribozyme) would have tighter binding than the ones in our dataset (**Table 1**). In our QSAR study, we covered a range of binding affinities to achieve a generalizable scope and aid the discovery of decisive descriptors for binding of diverse small molecules.

Table 1 Median values of binding parameters from three sets of RNA-ligand interaction, values for in vitro-selected and naturally occurring RNA-ligands from ref 46.

	k_{on} (M^{-1} s^{-1})	k_{off} (s^{-1})	K_d (M)
RNA (in vitro-selected) - ligand (N=13)[46]	**8.1×10^4**	**6.3×10^{-2}**	**4.3×10^{-7}**
RNA (naturally occurring) - ligand (N=24)[46]	**5.5×10^4**	**1.9×10^{-2}**	**3.0×10^{-7}**
HIV-1 TAR - ligand (N=40)	**1.3×10^4**	**9.0×10^{-2}**	**9.8×10^{-6}**

3. QSAR modeling: ensemble learning

Ensemble model construction

A key consideration in ensemble construction for QSAR is the continuity of the energy landscapes created by the ligands, i.e. whether gradual changes in ligand topology and electrostatic properties are smoothly plotted along the target activity function.[32, 47] While QSAR has been classically applied to molecules from the same scaffold (congeneric sets) to alleviate these concerns, several studies have reported successful continuous fields even with the use of diverse scaffolds.[48-50] Appropriate splitting of the training and test sets is critical to achieving a smooth landscape that avoids local minima where the model would explain only a subset of the compound pool.[51] A conventional single data split approach bears high risk of overfitting, especially with diverse scaffolds in the substrate pool. There are multiple sampling methods that can maximize the consistency between the selected subsets and the whole dataset, however, such as stratified sampling for a sample with multiple subpopulations.[52]

To reduce the potential bias and variance on QSAR prediction relating to single data splitting, we applied stratified sampling and independently repeated this procedure multiple times. As a result, the prediction is given by the ensemble of independent QSAR models derived from each sampling, reducing dependence on the representativeness of a single data splitting (**Panel C**, **Scheme 1**). Through multiple samplings and ensemble learning, we aimed to reduce the variance of the prediction.[53]

To test whether this ensemble learning strategy can achieve a high precision of prediction on the diverse substrates, we started modeling using ensemble learning workflow and compared the result with the model trained from a single data splitting. For the 40-compound dataset, we first classified the molecules into different categories or "strata" by K-means clustering. Using the 304 descriptors calculated for each ligand, we initially set K=4 based on the four scaffolds as previously mentioned, and the converged results matched well with the expected, with the exception of mitoxantrone joining the DMA set rather than the set of nucleic acid dyes (**Table S1**). Considering the large size of the DMA group and structural diversity within the DMA set, we decided to set K=5 where the converged clustering result further split the DMA set (**Table 2**). The resulting five categories explored the chemical diversity of this scaffold more comprehensively, while also generating more equally distributed subgroups and favoring representative data splitting when compared to K=4 clustering. To maximize the applicable domain of the final model, the ligands DPF p1 and DMA-186, representing the upper and lower limits of K_D, respectively, were forced to be in the training set. A similar strategy was also applied when conducting k_{on} and k_{off} modeling.

Table 2 Classification of parent dataset into five categories: Aminoglycosides, DMA set1, DMA set 2, DPFs and nucleic acid dyes, referring K-means clustering result (K=5). Note: DMA-1~DMA-164 are from ref 40, DMA-180~DMA-194 from ref 41, DCC compounds from ref 54, DPF x1~DPF x10 from ref 42 (x = m or p), DPF p15 from ref 43. The rest of compounds are commercially available.

Aminoglycosides (N=9)	DMA set1 (N=5)	DMA set2 (N=12)	DPFs (N=7)	Nucleic Acid Dyes (N=7)
Neomycin B	DMA-1	DMA-180	DPF m1	Acridine Orange
Paromomycin	DMA-148	DMA-187	DPF p1	TO-PRO-1
Sisomycin	DMA-156	DMA-190	DPF m3	Furamidine
Streptomycin	DMA-164	DMA-191	DPF m9	Ethidium Bromide
Tobramycin	DMA-186	DMA-193	DPF m10	Mitoxantrone
Gentamicin		DMA-194	DPF p6	Thiazole Orange
Neamine		DCC-3k	DPF p15	H-33258
Kanamycin		DCC-3l		
Amikacin		DCC-3u		
		DCC-3v		
		DCC-3r		
		DCC-3q		

Exhaustive searching for 3-variable linear models ($y \sim 1+x_1+x_2+x_3$) based on ordinary least squares (OLS) was performed independently for each sampling. Namely, the total descriptor space (with 304 variables) was searched for all 3-variable combinations and tested on their adjusted R^2 values in the regression model, resulting in 4,636,304 models for each data splitting. The number of variables in the model was determined as three to ensure a reasonably high R^2 and also to be far below the limit set by the Topliss rule[55] to avoid overfitting. Resulting models were evaluated by multiple fitting metrics, including adjusted R^2 (on training dataset), R^2_{LOOCV} (leave-one-out cross-validation on training set), Q^2_{F2} (R^2 on test set), etc. (see detailed equations for the fitting metrics in **Methods and scripts, SI**). We set different cut-off values of these metrics based on literature precedence to select eligible models for ensemble learning.[56, 57] Specifically, good models were first selected based on the adjusted R^2 (R^2 >0.75) (**Scheme S1**). The robustness of these models was then determined through a leave-one-out cross-validation (R^2_{LOOCV}>0.7) on the training set, and the predictiveness was validated by the R^2 of the test set (Q^2_{F2}>0.65 for K_D and Q^2_{F2}>0.75 for k_{on}). Models that met these criteria were determined top models.

The ensemble model was then constructed using the top models selected from each data splitting. To ensure that each data splitting was considered equally, the values for the top models from each corresponding splitting were equally weighted and normalized to the total number of data splittings (120) to create the ensemble model. We tested the model behavior by including more data splittings (**Table S2**), however the overall RMSE and R^2 of fitting changed only slightly, indicating that 120 runs are sufficient to obtain a converged result.

For the ensemble model of K_D, there were 785 top models involved in total (**Figure 2a**). The model resulting from the aforementioned ensemble method yielded overall lower RMSE (1.0226) and higher R^2 (0.83228) than any individual model. This result could be explained by the fact that in the ensemble method, all of the small molecules contributed to the training of the model thereby lowering the risk of extreme predictions that could result from a single model built on a static training set. Residuals remained unexplained

by the ensemble model, which is consistent with avoiding overfitting but may indicate that certain compounds have deviated behaviors relative to those predicted by the model. Similarly, for the modeling of k_{on} the ensemble model was an integrated result of 1429 top models (**Figure 2b**). The overall explanatory ability of the ensemble model on the parent dataset exceeds any individual model, characterized with an RMSE of 1.4215 and R^2 of 0.82019.

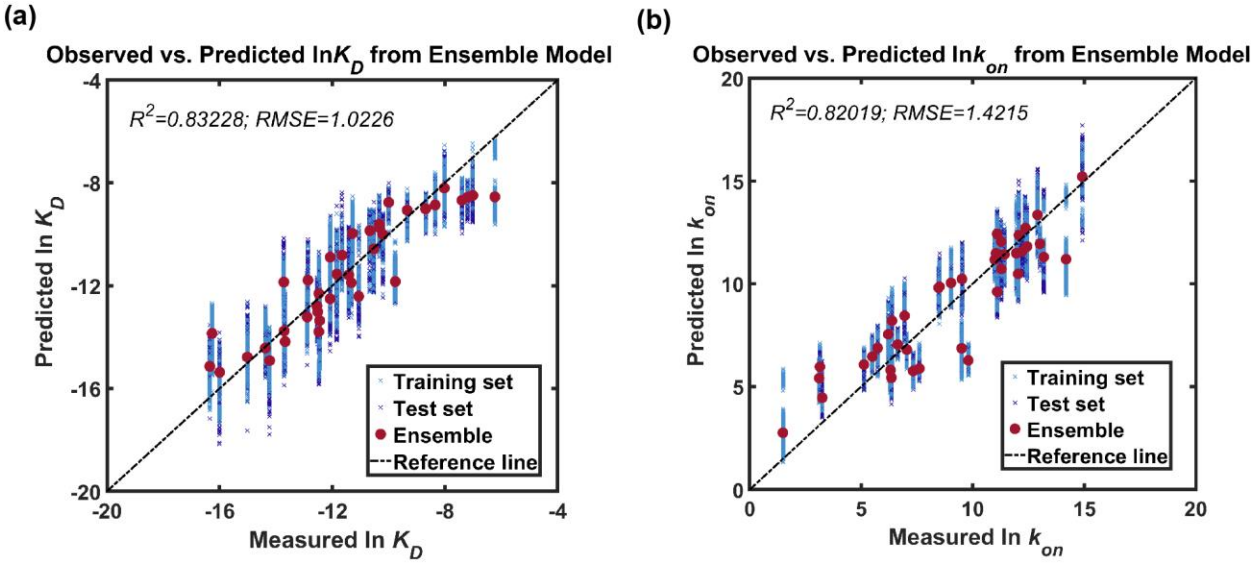


Figure 2 Ensemble model prediction. **(a)** lnK_D model constructed from 785 top models. **(b)** lnk_{on} model constructed from 1429 top models.

The model search for k_{off} resulted in poorly fitted models, indicating the complexity of the dissociation process and suggesting that more sophisticated regressions (i.e. nonlinear regression) may be needed for the explanation of this parameter, which also bears less variation than K_D and k_{on}. As a result, the analyses performed in the rest of the study focused on K_D and k_{on}.

Comparison with elastic net regularization and random forest

To further evaluate the suitability of the ensemble model, we compared behaviors of the ensemble model with models trained from two commonly used algorithms, i.e. elastic net regularization and random forest (RF). Elastic net regularization was chosen because the loss function (**Table S3**), which is a measure of goodness during model prediction, is very similar to OLS and it includes penalty terms for overfitting, an important consideration with our large number of variables. RF was chosen because it includes bagging as

construction method, which is similar to our ensemble approach, though the loss function is very different.

As mentioned above, for a high-dimension descriptor set, it is important to prevent over-fitting and to pick the most relevant descriptors. Elastic net regularization avoids over-fitting by adding two penalty terms (adjusted by α) on the basis of OLS in its loss function (see **Table S3**).[58] The result of lnK_D modeling showed that lasso regression gave a conserved prediction, namely that low lnK_D and high lnK_D values were predicted to be moderate lnK_D values (**Figure 3a**). Adjusting weights of two penalty terms in loss function gave similar conserved predictions, close to the average of the whole data set (**Figure 3b**). Similar results were found during lnk_{on} modeling (**Figure S1**). Therefore, elastic net regularization could not afford a comparable precision on the given dataset when compared to the ensemble least squares model we trained via OLS.

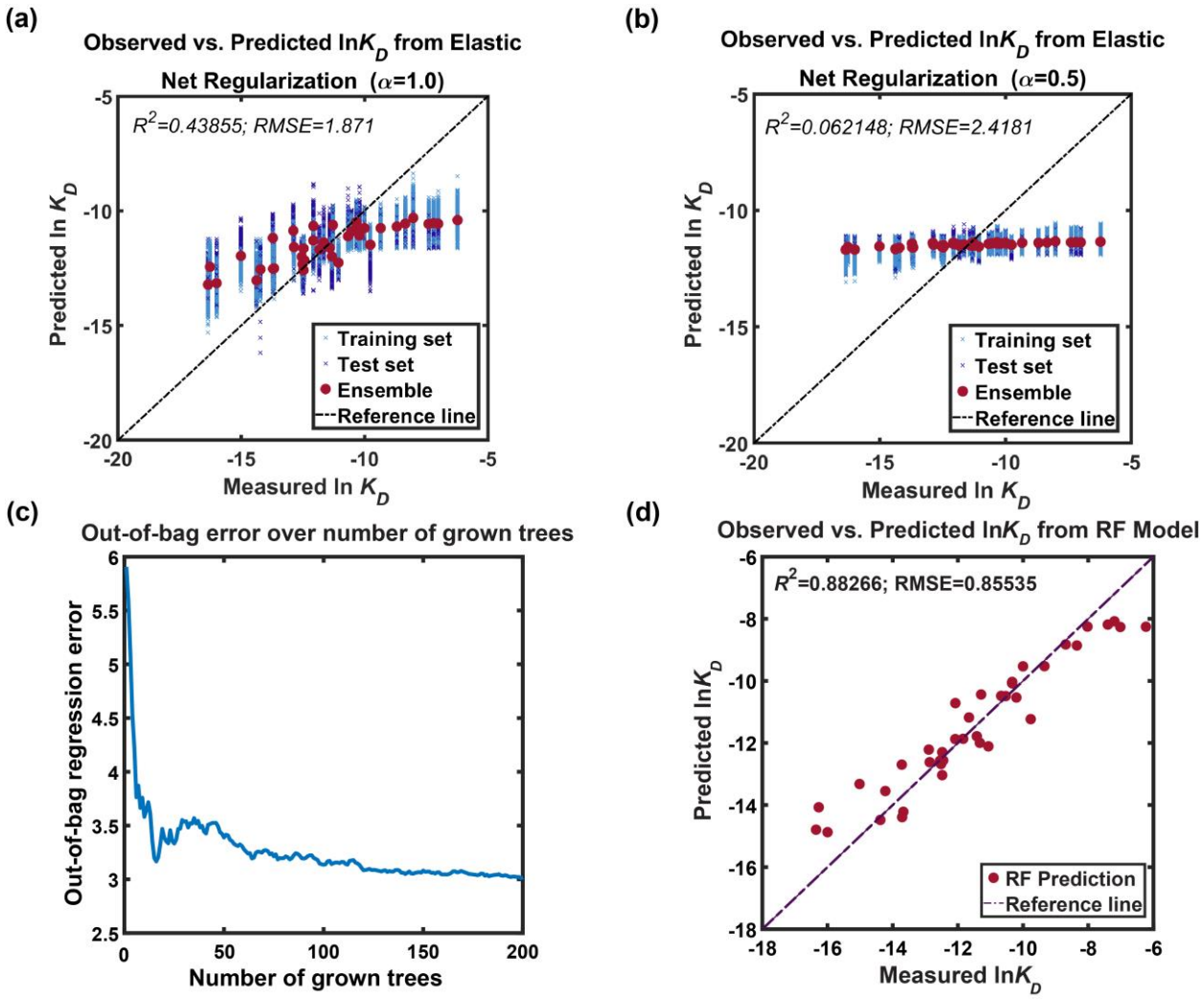


Figure 3. Comparison with elastic net regularization and random forest. **(a)**. Averaged prediction from 3000 independent runs from lasso regression (α = 1), predicted lnK_D was compared to the measured values using a reference line. **(b)**. Averaged prediction from 3000 independent runs from elastic net regularization (α = 0.5), predicted lnK_D was compared to the measured values using a reference line. See more results on elastic net regularizations (including ridge regression) in **Figure S1**. **(c)**. The out-of-bag error decreased with the number of grown trees in RF model. **(d)**. Overall prediction of RF model on the dataset, the predicted lnK_D was plotted over measured lnK_D, compared through a reference line.

Having diverse substrates as training set, it is quite important to reduce the variance of the prediction. To evaluate how the QSAR ensemble model behaved on lowering variance and realizing precise prediction, we compared it with RF model. The random forest algorithm reduces variance via the application of bagging on sampling and random choice on sub-feature space, providing stable and accurate predictions.[59] When compared to our linear regression models, RF is constructed by multiple non-linear decision trees based on the information gain at each node, making it powerful at multiple machine learning tasks but also hard to interpret. Utilizing the *TreeBagger* function in

MATLAB, we built a regression forest consisting of 200 decision trees. The out-of-bag regression error was plotted along with the number of trees grown (**Figure 3c**). Results indicated that 200 trees are sufficient to reach convergency. The R^2 for out-of-bag prediction reached around 0.5, even though the overall prediction on the dataset reached 0.88 R^2, indicating the heterogeneity of the dataset and a certain extent of overfitting. We changed the size of subspace used in each node construction to see the effect on model quality, and the summarized results (**Table S4 and Figure S2**) suggested that the size of the subspace makes little difference on the model's behavior.

Overall, our ensemble model reached a comparable level of precision on the prediction to the RF model (**Figure 3d**), indicating its utility as a way to use linear regression with diverse substrates and keep explicit model interpretation, which will be discussed later.

4. External validation: prediction of the unknowns

We next tested whether the ensemble model could be used to predict the binding parameters of untested compounds and further guide ligand synthesis. We performed external validation using nine untested compounds after model construction. Five of the untested molecules were diphenylfuran analogs related to others that had been included in the parent dataset (8 out of 40). The remaining four molecules were in-house synthesized molecules that contain a diminazene (DMZ) scaffold, which has been shown as a potential RNA-targeted scaffold[60] but has been less explored so far. Importantly, no DMZ-based molecules were in our training set, and thus this scaffold represents a rigorous test of model performance. The SPR results indicated that eight of nine molecules bound HIV-1 TAR RNA, with only DMZ-O6 showing a weak SPR signal, either due to weak binding or poor SPR response properties. The lnK_D and lnk_{on} parameters of the eight molecules were predicted using our ensemble models and compared with the predictions from the random forest model. As shown in **Figure 4**, both ensemble (**Figure 4a**) and RF models (**Figure 4c**) achieved high precision on the lnK_D prediction of the external set, represented by the low RMSE values, where the prediction of RF model reached slightly lower RMSE. The prediction interval (PI) defined by standard error of observations describes the range a new observation would fall into, which was used here to map the precision of new predictions from the two models. In lnK_D predictions, only one

compound (DMZ-M3, **Figure 5**) predicted from the ensemble model fell outside the 95% prediction interval (PI) range, while predictions from the RF model all fell within the 95% PI range. Similarly, for $\ln k_{on}$ predictions, the RF model (**Figure 4d**) outperformed the OLS model (**Figure 4b**), indicated by the lower RMSE value. Only one point (DPF p13, **Figure 5**) predicted by RF model fell outside the 95% PI range, which was also the only compound that was not well predicted by the OLS model.

Overall, the OLS ensemble model reached high precision for external prediction. Specifically, it correctly predicted binding parameters of the DMZs whose structures were not present in the training set, showing strong potential for generalizability. The ensemble model could achieve comparable performance to the well-established RF algorithm-trained model, supporting, once again, the applicability of this ensemble method for QSAR studies with diverse substrates.

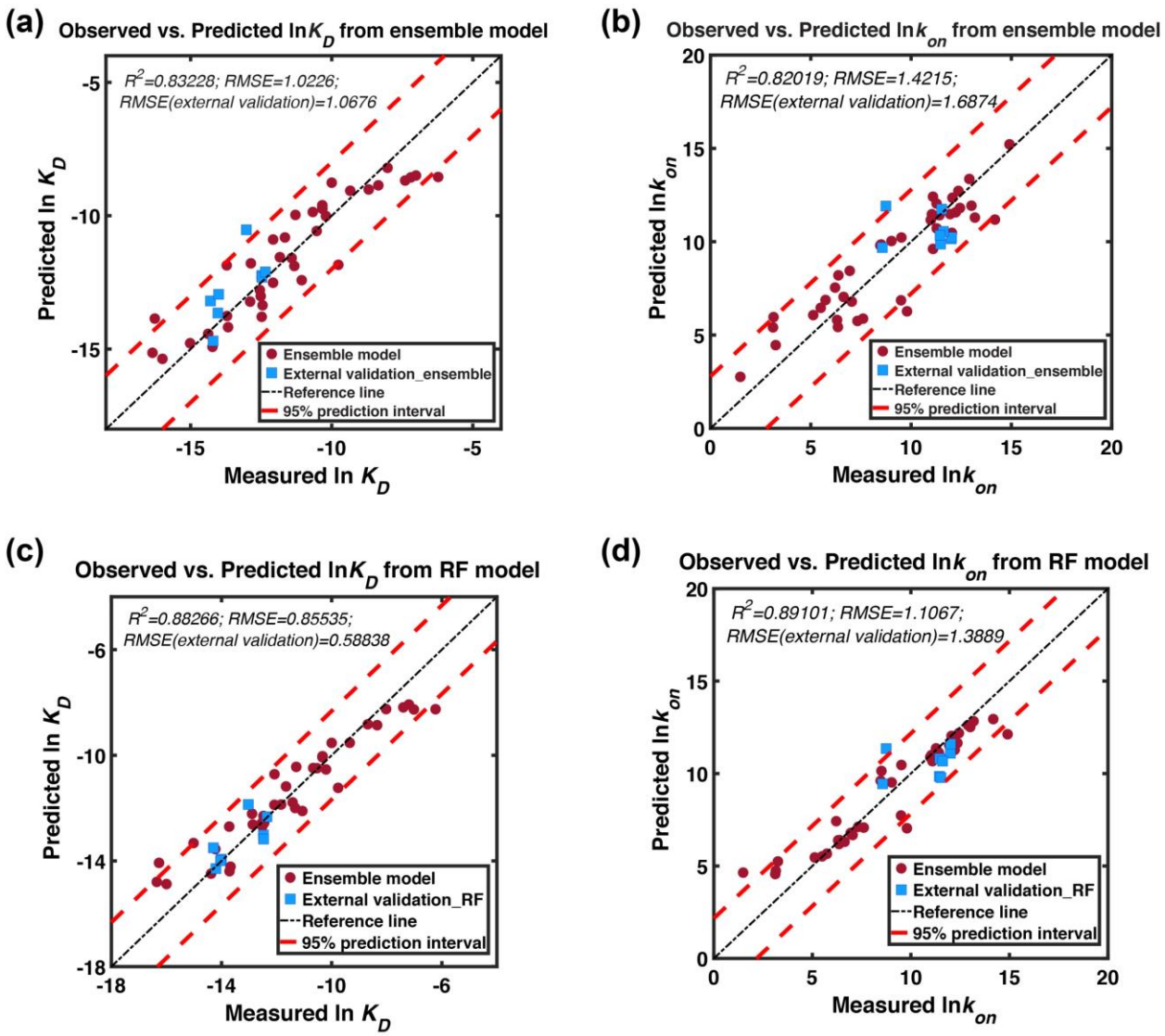
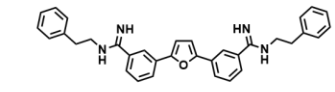
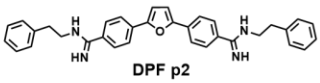


Figure 4. External validation on ensemble models of **(a)** lnK_D and **(b)** lnk_{on}, compared with RF models of **(c)** lnK_D and **(d)** lnk_{on} . The 95% PI was defined by the t-multiplier and standard error (see equations in **Table S5**), as represented by the dashed red line.



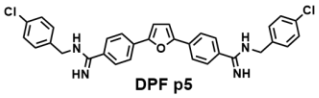
DPF m2

	lnK_D	lnk_{on}
Predicted	-12.2	10.2
Observed	-12.5	12.0
95% PI	[-14.3, -10.1]	[7.2, 13.1]



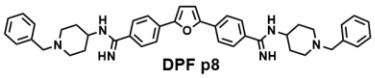
DPF p2

	lnK_D	lnk_{on}
Predicted	-12.3	10.2
Observed	-12.5	12.0
95% PI	[-14.4, -10.2]	[7.3, 13.1]



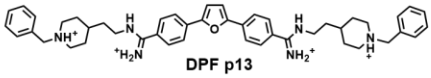
DPF p5

	lnK_D	lnk_{on}
Predicted	-12.1	9.9
Observed	-12.4	11.5
95% PI	[-14.2, -10.0]	[7.0, 12.8]



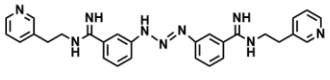
DPF p8

	lnK_D	lnk_{on}
Predicted	-13.2	10.6
Observed	-14.3	11.6
95% PI	[-15.3, -11.1]	[7.6, 13.5]



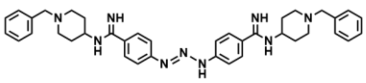
DPF p13

	lnK_D	lnk_{on}
Predicted	-14.7	11.9
Observed	-14.2	8.7*
95% PI	[-16.8, -12.6]	[9.0, 14.8]



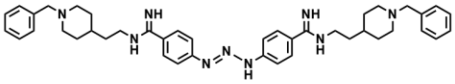
DMZ m3

	lnK_D	lnk_{on}
Predicted	-10.5	9.7
Observed	-13.0*	8.6
95% PI	[-12.6, -8.4]	[6.8, 12.6]



DMZ p8

	lnK_D	lnk_{on}
Predicted	-12.9	10.3
Observed	-14.0	11.4
95% PI	[-15.0, -10.8]	[7.4, 13.2]



DMZ p13

	lnK_D	lnk_{on}
Predicted	-13.6	11.7
Observed	-14.0	11.5
95% PI	[-15.7, -11.6]	[8.8, 14.6]

Figure 5. Chemical structures of molecules tested in external validation as well as their predicted, observed and 95% predictive interval of lnK_D and lnk_{on}. Asterisk (*) indicates the outliner by 95% PI.

5. Model interpretation

After QSAR model construction, we investigated the relationship between small molecule binding parameters and the predictors that yielded the best models. One way to obtain such information is to extract the descriptor importance from the model and assess the most significant descriptors. As a parallel comparison, we appraised interpretation of both the ensemble model and the RF model. We expected the ensemble model to be more explicit as the random forest model is often considered a “black box” method due to the inclusion of large number of deep decision trees.

To interpret descriptor importance from the ensemble model, all the descriptors that contributed to the model were recorded as well as their coefficients. The coefficients for each individual descriptor, weighted as described above in the ensemble model construction, were added, and the summed result of each descriptor was deemed descriptor importance. As shown in **Figure 6a** and **6b**, the ensemble models for lnK_D and

lnk_{on} were comprised of 49 and 67 descriptors (more details in **Table S6** and **Table S7**), respectively. In the model of lnK_D, the top five descriptors occupied 67.2% of the descriptor weight, among which PEOE_RPC- (#144) and PEOE_PC+ (#141) are descriptors related to partial charge, b_1rotR (#45) and b_rotR (#50) both describe molecular flexibility, and GCUT_PEOE_2 (#92) is a hybrid descriptor recording charge information derived from a graphic representation of the molecule (**Table 3**).

Table 3 Top 5 descriptors from lnK_D model and their physical meanings

Descriptor ID	Coefficients	Physical meaning
#144: PEOE_RPC-	-1.76	Relative negative partial charge: the smallest negative charge divided by the sum of the negative charge
#45: b_1rotR	-0.65	Fraction of rotatable single bonds: b_1rotN divided by b_heavy.
#50: b_rotR	0.61	Fraction of rotatable bonds: b_rotN divided by b_heavy
#92: GCUT_PEOE_2	-0.52	Calculated from the eigenvalues of a modified graph distance adjacency matrix
#141: PEOE_PC+	-0.39	Total positive charge

Through discerning these five descriptors' coefficients (**Table 3**) from lnK_D model, the following directions were proposed to discover more high-affinity ligands:

1. Increase PEOE_RPC- values since its coefficient is negative. We could increase the ratio of the smallest negative charge to the total negative charge in a molecule. This interpretation indicated that to favor binding, the partial negative charge in a molecule should be concentrated at discrete positions rather than distributed dispersedly. One interpretation is that a dispersed partial negative charge inhibits binding via electrostatic repulsion with the RNA backbone while discrete negative charges can be accommodated or even be beneficial through metal coordination or hydrogen bonding. GCUT_PEOE_2, an adjacency and distance matrix descriptor related to partial charges, also supports an important role for partial charge distribution on apparent binding, though not as directionally as PEOE_RPC-.

1. b_1rotR (#45) and b_rotR (#50) both relate to molecular flexibility. Although they have opposite signs, if we dissect the detailed equation (**Figure S3**), structural differences

can be interpreted. Specifically, the ratio of the conjugated single bonds, i.e. single bonds connected to atoms in double bonds, relative to total bonds between heavy (non-hydrogen) atoms should be decreased. Groups such as esters or amides tend to increase the ratio of conjugated single bonds, as do aromatic rings with attached sp3-hybridized atoms. In our case, the percentage of guanine or amidine groups relative to the rest of the molecule should be kept low. This is consistent with our observations that most of the strong ligands, such as aminoglycosides, DPFs and dye molecules, are characterized by a lower ratio of these functionalities than weak binders, namely DMAs (see the detailed ratio of conjugated single bonds for each compound in **Figure S4**). DMAs have fewer total heavy bonds but multiple single bonds adjacent to double bonds, including in the acylguanidine substituent and the exhaustive substitution of the pyrazine ring.

3. Increase the molecules total positive charge as is indicated by the negative coefficient of PEOE_PC+ descriptor. Unsurprisingly given the negatively charged RNA backbone, the total positive charge can significantly impact affinity. For instance, aminoglycosides are characterized by a high PEOE_PC+ values, gaining moderate to high affinity even though they have low PEOE_RPC- values.

In the model of lnk_{on}, we found that two descriptors predominated: b_1rotR (#45) and b_rotR (#50), which totaled 91.2% of the whole weight. Similarly, as to lnK_D, the signs of the two descriptors are opposite, even though there is overlap in the associated physical properties. The detailed dissection of these two terms could afford a similar but more decisive conclusion when compared to the lnK_D model: to increase the association rate constant, the ratio of the conjugated single bonds to bonds between heavy atoms should be decreased. This finding might suggest a novel strategy for improving RNA-ligand association kinetics via rational structural design, which is different from a previous report that utilized the electrostatic anchor.[22] Combined results from the descriptor analysis of lnK_D and lnk_{on} reveals that molecular flexibility has an impact on the RNA recognition mostly via the effect on the association process, also suggesting a quantitative direction for ligand optimization using this model.

In contrast, the descriptor importance of the RF model could be extracted by permuting out-of-bag observations among the trees.[61] This approach assumes that if the prediction is highly dependent on a specific descriptor, then by shuffling the values of that descriptor, the prediction error will increase significantly. Via descriptor permutation, we generated the descriptor importance for lnK_D and lnk_{on} from RF models (**Figure 6c and 6d**). This analysis revealed that almost all of descriptors had contributions to the model, with multiple descriptors having similar impact on the model prediction. Therefore, less information could be derived to enable rational design of new small molecule structures when compared to the ensemble model.

Model interpretation is as important as model performance since it provides directions of how to improve for small molecules. Even though the RF model showed excellent precision on the data fitting and prediction, the non-trivial model interpretation limits its practical use, especially on structure design. In contrast, the ensemble model equipped us with lens to investigate molecular factors having quantitative impact on binding. Informed by the physiochemical meaning of the decisive descriptors, ligand and library synthesis decisions can be rationally guided.

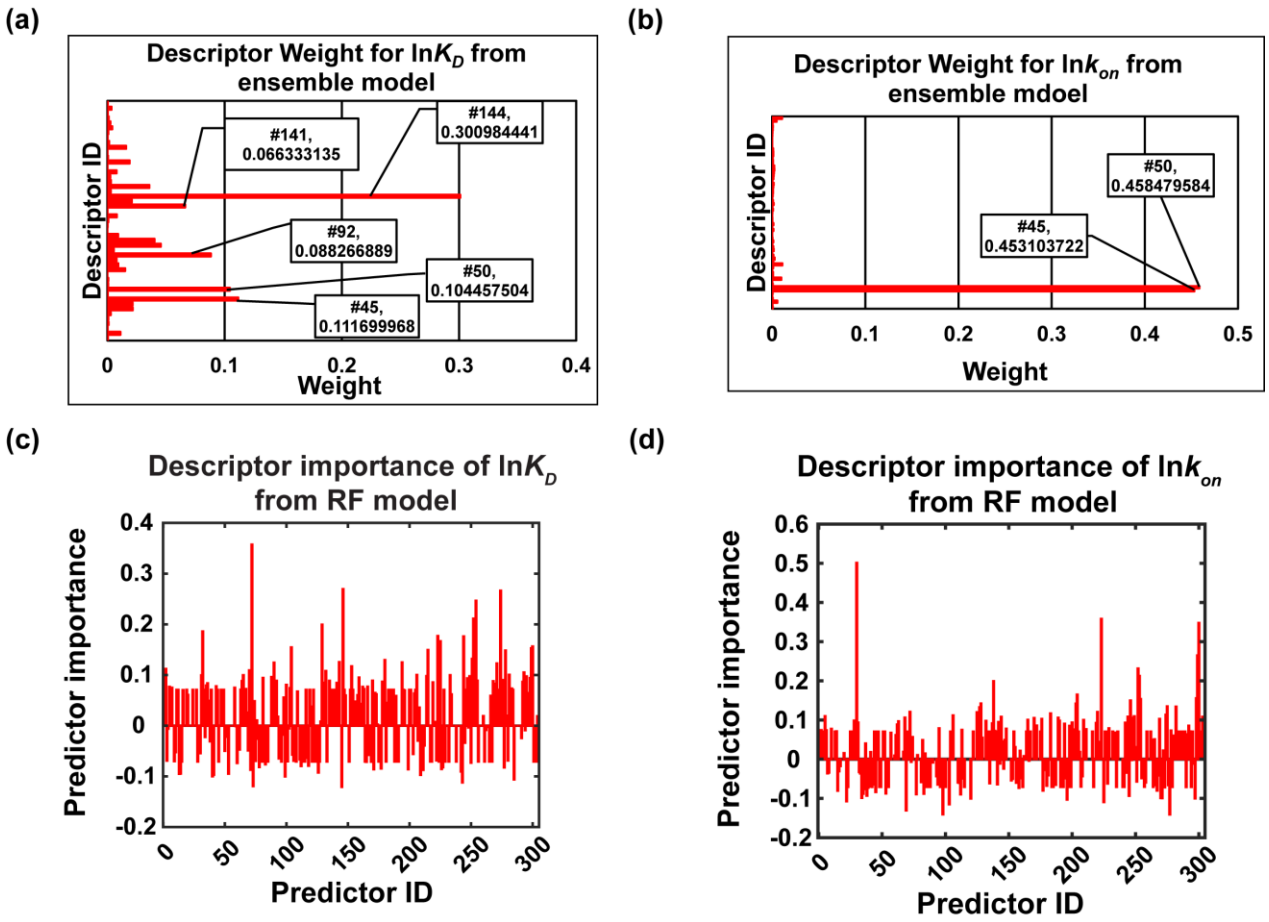


Figure 6. Descriptor weight for **(a)** lnK_D model and **(b)** lnk_{on} model calculated from equation coefficients. Data label were created for top 5 descriptors in lnK_D and top 2 descriptors in lnk_{on}, respectively. Descriptor importance was calculated for **(c)** lnK_D and **(d)** lnk_{on} from RF models by permutation tests.

6. Surfing the structure-activity landscape

The structure-activity landscape integrates structural similarity and activity function relationships between compounds, revealing QSAR continuity regions where activity function changes gradually with structures as well as discontinuity regions where small changes in chemical structure lead to significant changes in activity.[51] Since the compounds involved in the QSAR study were diverse, the surface of the structure-activity landscape would be expected to be rugged and deviate from the continuous region of QSAR. To investigate how a superficially rugged structure-activity landscape was able to afford models with high R^2, we first calculated the structure-activity landscape index (SALI) as:

$$SALI_{i,j} = \frac{|A_i - A_j|}{1 - sim(i,j)}$$

where A represents the potency or activity function of the compound (i.e. lnK_D or lnk_{on}), and sim(i,j) is the similarity coefficient between two compounds.[62] The SALI matrix could easily reveal "activity cliffs" in the landscape where highly similar compounds show compelling difference in activities. In the SALI calculation, it is worth noting that the choice of the molecular descriptors in the similarity calculation has been identified to impact the results more than the choice of the similarity metric.[62] Therefore, in our calculations, we consistently used Tanimoto coefficients[63] as the similarity metric but changed the descriptor set during calculation to compare the output.

We first mapped out the similarity matrix using the full 304-descriptor set, i.e. without input from the modeling, and ranked the index based on lnK_D and lnk_{on}, respectively (**Figure 7a** and **Figure S5a**). The similarity heat map was consistent with the most prominent similarity existing within scaffolds, including aminoglycosides, DPFs and DMAs. In contrast, low similarity was identified from scaffold to scaffold. Nucleic acid dyes also share low similarity scores between each other, probably due to their diverse structural features. Using these similarity matrices, the corresponding SALI matrices were generated to visualize the potential "activity cliffs". The resulting heat maps (**Figure 7a** and **Figure S5b**) indicated that most of the potency differences observed in lnK_D and lnk_{on} could be smoothly explained by the variation in descriptors, even though different scaffolds co-existed in one dataset. More explicitly, the structurally dissimilar compounds in the dataset were also characterized with dissimilar binding behaviors, thus providing a path to construct the QSAR model along which the descriptor spaces are proportional. Despite most of the continuous landscape, a few sporadic "activity cliffs" were found within a scaffold, for example Neomycin B and Paromomycin in lnK_D along with DCC-3u and DCC-3k in lnk_{on}. These "activity cliffs" might be the cause of the unexplained residuals in our model, which are important for avoiding overfitting and maintaining the needed precision balance for external generalization.

The similarity matrices calculated on full descriptors are a good indicator for “dataset modelability” before model construction.[64] At the post-modeling stage, to investigate how completed models encoded for the compound activities, we shrunk the descriptor space to only descriptors that described the final models, namely a set of 49 descriptors for lnK_D and 67 descriptors for lnk_{on}. The overall similarity scores were increased as expected, leading to an overall increase of the SALI index as indicated by the scale bar (**Figure 7b** and **Figure S5c**). Nevertheless, the pattern of the SALI matrices remained (**Figure 7b** and **Figure S5d**), where most of the compound potency changed smoothly as the structural features changed. The use of more specified descriptor space enhanced the index for identifying “activity cliffs”, revealing unexplored pairs that might be an initiation of activity optimization. However, the “activity cliffs” might also arise from extremely similar compounds (e.g. DPF m1 and DPF p1) that, due to the high similarity value, might be the false “activity cliffs”. Indeed, most of the “activity cliffs” found above are between similar congeners. In the future, omitting highly similar pairs will further improve the suitability for SALI analysis on QSAR.

In summary, the employment of SALI revealed good modelability of our dataset, as well as few observed “activity cliffs” that would translate into unpredictable residuals in final models. The existence of the “activity cliffs” could be double-edged, however, as it deviates the model from the continuous region but also could be the starting point for design of a new scaffold complemented by the “mechanism hopping”.[65]

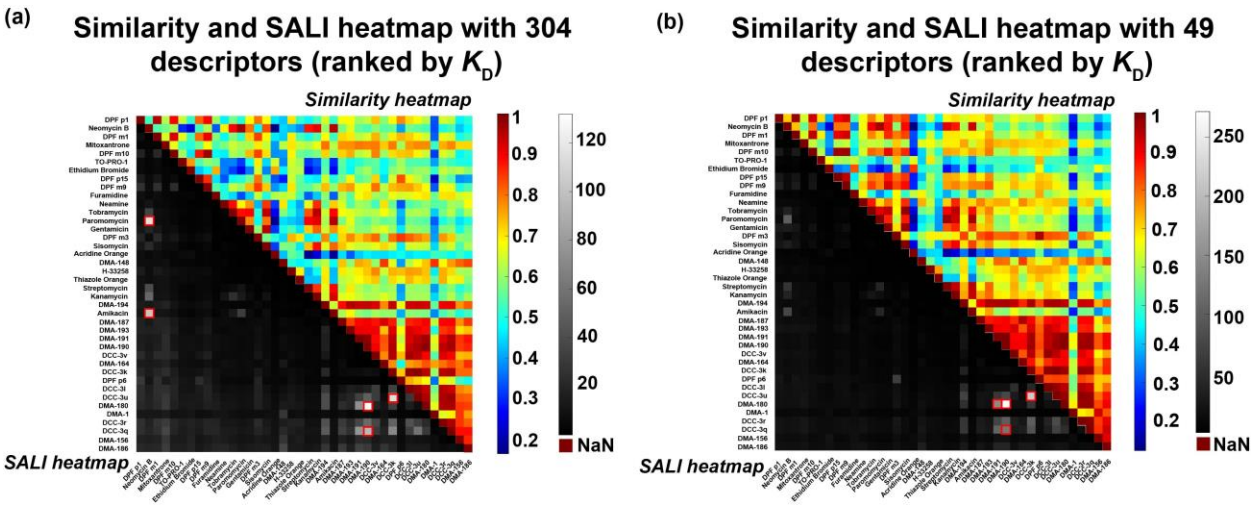


Figure 7. (a) Similarity and SALI heatmap calculated from full 304 descriptors, ranked by lnK_D, prominent "activity cliffs" were highlighted by red boxes **(b)** Similarity and SALI heatmap calculated from 49 descriptors used in ensemble lnK_D model, ranked by lnK_D., prominent "activity cliffs" were highlighted by red boxes.

Conclusion

Discovery of novel RNA-targeted chemical probes is pivotal for connecting basic understanding of RNA regulation in biology and its potential therapeutic application. Numerous ncRNAs as potential drug targets are emerging following the RNA revolution. However, difficulties in obtaining accurate 3D structures and conformational landscape for a given RNA hinders efficiency of rational design of the RNA-targeted ligand from a structure-based approach. Additionally, lack of appreciation of binding kinetics in hit discovery compromised an alternative path towards ligand optimization via kinetic selectivity. Consequently, a novel method that can bypass the structural information and comprehensively evaluate binding parameters, from affinities to kinetics, is greatly needed. Benefiting from the era of Big Data and well-developed algorithms for multiple machine learning tasks, we developed a strategy for computation-aided compound design. Specifically, a systematic QSAR workflow for RNA ligand discovery was built using HIV-1 TAR as a model system to demonstrate the potential application of this method on a broad scope of ligands. To the best of our knowledge, this is the first time that 2D-QSAR has been used to predict binding parameters of RNA-targeted ligands beyond explaining the observed data and also the first time that ensemble learning has

been applied to linear regression to ensure a precise and interpretable model with a large number of variables.

By applying an ensemble learning strategy, we trained models from 40 diverse small molecules as the basis of our understanding of RNA ligand chemical space. The trained models afforded satisfactory explanations for both binding affinities and kinetics data from SPR. The subsequent external validation of eight previously untested compounds revealed similar precision as compared to the well-established random forest algorithm, supporting the power of our ensemble models to inform compound design. Notably, our model was able to accurately predict the affinity and k_{on} of three untested small molecules synthesized from a distinct scaffold not present in the training set, underscoring the breadth of application of the method to a diverse small molecule library. The detailed analysis of the descriptor space highlighted by the best models revealed important roles of molecular flexibility and potential charge, both localized and general, in RNA recognition of small molecules. Moreover, the ensemble model provided information on the modification of these descriptors to better aid molecular design and lead optimization based on the coefficients in the model equations. Further investigation on the structure-activity landscape suggested a wide range of smoothly transited regions where compound potency could be modeled via gradual changes in structural features.

We anticipate that the method applied here will be an efficient tool in hit identification and lead optimization for a wide range of specific RNA targets. The knowledge gained from known ligands during training can now be efficiently transformed into quantitative models for generalization, i.e. prediction of binding affinity and kinetics. Additionally, this proof-of-concept study could be feasibly extended to other biomacromolecules targets with little structural characterization, including other ncRNAs and proteins. Various parameters could be investigated as well, such as binding entropy and enthalpy. We anticipate the workflow set forth here to significantly facilitate rational decision-making in medicinal chemistry, overcoming one of the current bottlenecks in RNA-targeted small molecule development.

Acknowledgements

We acknowledge past and present Hargrove Lab members for their assistance with project conceptualization and manuscript editing. We particularly thank former lab members Dr. Neeraj Patwardhan, Ph.D., Dr. Anita Donlic, Ph.D. and Dr. Aline Umuhire Juru, Ph.D. for donating the synthesized DMA, DPF and DCC molecules used here. Surface plasmon resonance analyses were performed in the Duke Human Vaccine Institute's Biomolecular Interaction Analysis Shared Resource Facility (Durham, NC) under the direction of Dr. S. Munir Alam and Dr. Brian E. Watts.

This work was supported by Duke University, U.S. National Institutes of Health (U54 AI150470), the Alfred P. Sloan Foundation, and an award from Duke University School of Medicine Core Facilities for use of the BIA Core. Z.C. was supported in part by a Kathleen Zielik Fellowship from the Duke University Chemistry Department.

References

1. ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **2007,** *447* (7146), 799-816.
2. Cech, Thomas R.; Steitz, Joan A., The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell* **2014,** *157* (1), 77-94.
3. Ji, Q.; Zhang, L.; Liu, X.; Zhou, L.; Wang, W.; Han, Z.; Sui, H.; Tang, Y.; Wang, Y.; Liu, N.; Ren, J.; Hou, F.; Li, Q., Long non-coding RNA MALAT1 promotes tumour growth and metastasis in colorectal cancer through binding to SFPQ and releasing oncogene PTBP2 from SFPQ/PTBP2 complex. *British Journal of Cancer* **2014,** *111* (4), 736-748.
4. Gupta, R. A.; Shah, N.; Wang, K. C.; Kim, J.; Horlings, H. M.; Wong, D. J.; Tsai, M.-C.; Hung, T.; Argani, P.; Rinn, J. L.; Wang, Y.; Brzoska, P.; Kong, B.; Li, R.; West, R. B.; van de Vijver, M. J.; Sukumar, S.; Chang, H. Y., Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **2010,** *464* (7291), 1071-1076.
5. Esteller, M., Non-coding RNAs in human disease. *Nat Rev Genet* **2011,** *12* (12), 861-874.
6. Gibb, E. A.; Brown, C. J.; Lam, W. L., The functional role of long non-coding RNA in human carcinomas. *Molecular Cancer* **2011,** *10* (1), 38.

7. Chen, X.; Yan, C. C.; Zhang, X.; You, Z.-H., Long non-coding RNAs and complex diseases: from experimental results to computational models. *Briefings in Bioinformatics* **2017,** *18* (4), 558-576.
8. Taft, R. J.; Pang, K. C.; Mercer, T. R.; Dinger, M.; Mattick, J. S., Non-coding RNAs: regulators of disease. *The Journal of Pathology* **2010,** *220* (2), 126-139.
9. Shi, X.; Sun, M.; Liu, H.; Yao, Y.; Song, Y., Long non-coding RNAs: A new frontier in the study of human diseases. *Cancer Lett* **2013,** *339* (2), 159-166.
10. Johnson, R., Long non-coding RNAs in Huntington's disease neurodegeneration. *Neurobiology of Disease* **2012,** *46* (2), 245-254.
11. Gutschner, T.; Diederichs, S., The hallmarks of cancer. *RNA Biology* **2012,** *9* (6), 703-719.
12. Brown, J. A.; Bulkley, D.; Wang, J.; Valenstein, M. L.; Yario, T. A.; Steitz, T. A.; Steitz, J. A., Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. *Nat Struct Mol Biol* **2014,** *21* (7), 633-640.
13. Rizvi, N. F.; Smith, G. F., RNA as a small molecule druggable target. *Bioorg Med Chem Lett* **2017,** *27* (23), 5083-5088.
14. Matsui, M.; Corey, D. R., Non-coding RNAs as drug targets. *Nat Rev Drug Discov* **2017,** *16* (3), 167-179.
15. Wahlestedt, C., Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat Rev Drug Discov* **2013,** *12* (6), 433-446.
16. Ling, H.; Fabbri, M.; Calin, G. A., MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nat Rev Drug Discov* **2013,** *12* (11), 847-865.
17. Thomas, J. R.; Hergenrother, P. J., Targeting RNA with Small Molecules. *Chemical Reviews* **2008,** *108* (4), 1171-1224.
18. Morgan, B. S.; Forte, J. E.; Hargrove, A. E., Insights into the development of chemical probes for RNA. *Nucleic Acids Research* **2018,** *46* (16), 8025-8037.
19. Walkup, G. K.; You, Z.; Ross, P. L.; Allen, E. K. H.; Daryaee, F.; Hale, M. R.; O'Donnell, J.; Ehmann, D. E.; Schuck, V. J. A.; Buurman, E. T.; Choy, A. L.; Hajec, L.; Murphy-Benenato, K.; Marone, V.; Patey, S. A.; Grosser, L. A.; Johnstone, M.; Walker, S. G.; Tonge, P. J.; Fisher, S. L., Translating slow-binding inhibition kinetics into cellular and in vivo effects. *Nat Chem Biol* **2015,** *11* (6), 416-423.
20. Schoop, A.; Dey, F., On-rate based optimization of structure–kinetic relationship – surfing the kinetic map. *Drug Discovery Today: Technologies* **2015,** *17*, 9-15.
21. Schneider, E. V.; Böttcher, J.; Huber, R.; Maskos, K.; Neumann, L., Structure–kinetic relationship study of CDK8/CycC specific compounds. *Proceedings of the National Academy of Sciences* **2013,** *110* (20), 8081.
22. Sengupta, R. N.; Herschlag, D., Enhancement of RNA/Ligand Association Kinetics via an Electrostatic Anchor. *Biochemistry* **2019,** *58* (24), 2760-2768.
23. Guo, J.-Y.; Minko, Y.; Santiago, C. B.; Sigman, M. S., Developing Comprehensive Computational Parameter Sets To Describe the Performance of Pyridine-Oxazoline and Related Ligands. *Acs Catal* **2017,** *7* (6), 4144-4151.
24. Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E., Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019,** *363* (6424), eaau5631.
25. de Ávila, M. B.; Xavier, M. M.; Pintro, V. O.; de Azevedo, W. F., Supervised machine learning techniques to predict binding affinity. A study for cyclin-dependent

kinase 2. *Biochemical and Biophysical Research Communications* **2017,** *494* (1), 305-310.
26. Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B., Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018,** *23* (8), 1538-1546.
27. Jamal, S.; Periwal, V.; Consortium, O.; Scaria, V., Computational analysis and predictive modeling of small molecule modulators of microRNA. *Journal of Cheminformatics* **2012,** *4* (1), 16.
28. Rizvi, N. F.; Santa Maria, J. P.; Nahvi, A.; Klappenbach, J.; Klein, D. J.; Curran, P. J.; Richards, M. P.; Chamberlin, C.; Saradjian, P.; Burchard, J.; Aguilar, R.; Lee, J. T.; Dandliker, P. J.; Smith, G. F.; Kutchukian, P.; Nickbarg, E. B., Targeting RNA with Small Molecules: Identification of Selective, RNA-Binding Small Molecules Occupying Drug-Like Chemical Space. *SLAS DISCOVERY: Advancing the Science of Drug Discovery* **2019,** *25* (4), 384-396.
29. Morgan, B. S.; Forte, J. E.; Culver, R. N.; Zhang, Y.; Hargrove, A. E., Discovery of Key Physicochemical, Structural, and Spatial Properties of RNA-Targeted Bioactive Ligands. *Angewandte Chemie International Edition* **2017,** *56* (43), 13498-13502.
30. Babu, P. A.; Smiles, D. J.; Narasu, M. L.; Srinivas, K., Identification of Novel CDK2 Inhibitors by QSAR and Virtual Screening Procedures. *QSAR & Combinatorial Science* **2008,** *27* (11-12), 1362-1373.
31. Tugcu, G.; Koksal, M., A QSAR Study for Analgesic and Anti-inflammatory Activities of 5-/6-Acyl-3-alkyl-2-Benzoxazolinone Derivatives. *Molecular Informatics* **2019,** *38* (8-9), 1800090.
32. Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A., QSAR without borders. *Chemical Society Reviews* **2020,** *49* (11), 3525-3564.
33. Maciagiewicz, I.; Zhou, S.; Bergmeier, S. C.; Hines, J. V., Structure–activity studies of RNA-binding oxazolidinone derivatives. *Bioorg Med Chem Lett* **2011,** *21* (15), 4524-4527.
34. Sekhar, Y. N.; Nayana, M. R. S.; Sivakumari, N.; Ravikumar, M.; Mahmood, S. K., 3D-QSAR and molecular docking studies of 1,3,5-triazene-2,4-diamine derivatives against r-RNA: Novel bacterial translation inhibitors. *Journal of Molecular Graphics and Modelling* **2008,** *26* (8), 1338-1352.
35. Setny, P.; Trylska, J., Search for Novel Aminoglycosides by Combining Fragment-Based Virtual Screening and 3D-QSAR Scoring. *Journal of Chemical Information and Modeling* **2009,** *49* (2), 390-400.
36. Mei, H.-Y.; Cui, M.; Heldsinger, A.; Lemrow, S. M.; Loo, J. A.; Sannes-Lowery, K. A.; Sharmeen, L.; Czarnik, A. W., Inhibitors of Protein−RNA Complexation That Target the RNA: Specific Recognition of Human Immunodeficiency Virus Type 1 TAR RNA by Small Organic Molecules. *Biochemistry* **1998,** *37* (40), 14204-14212.
37. Zeng, L.; Li, J.; Muller, M.; Yan, S.; Mujtaba, S.; Pan, C.; Wang, Z.; Zhou, M.-M., Selective Small Molecules Blocking HIV-1 Tat and Coactivator PCAF Association. *J Am Chem Soc* **2005,** *127* (8), 2376-2377.
38. Abulwerdi, F. A.; Shortridge, M. D.; Sztuba-Solinska, J.; Wilson, R.; Le Grice, S. F. J.; Varani, G.; Schneekloth, J. S., Development of Small Molecules with a

Noncanonical Binding Mode to HIV-1 Trans Activation Response (TAR) RNA. *Journal of Medicinal Chemistry* **2016,** *59* (24), 11148-11160.
39. Sztuba-Solinska, J.; Shenoy, S. R.; Gareiss, P.; Krumpe, L. R. H.; Le Grice, S. F. J.; O'Keefe, B. R.; Schneekloth, J. S., Identification of Biologically Active, HIV TAR RNA-Binding Small Molecules Using Small Molecule Microarrays. *J Am Chem Soc* **2014,** *136* (23), 8402-8410.
40. Patwardhan, N. N.; Ganser, L. R.; Kapral, G. J.; Eubanks, C. S.; Lee, J.; Sathyamoorthy, B.; Al-Hashimi, H. M.; Hargrove, A. E., Amiloride as a new RNA-binding scaffold with activity against HIV-1 TAR. *MedChemComm* **2017,** *8* (5), 1022-1036.
41. Patwardhan, N. N.; Cai, Z.; Umuhire Juru, A.; Hargrove, A. E., Driving factors in amiloride recognition of HIV RNA targets. *Org Biomol Chem* **2019,** *17* (42), 9313-9320.
42. Donlic, A.; Morgan, B. S.; Xu, J. L.; Liu, A.; Roble Jr, C.; Hargrove, A. E., Discovery of Small Molecule Ligands for MALAT1 by Tuning an RNA-Binding Scaffold. *Angewandte Chemie* **2018,** *130* (40), 13426-13431.
43. Donlic, A.; Zafferani, M.; Padroni, G.; Puri, M.; Hargrove, Amanda E., Regulation of MALAT1 triple helix stability and in vitro degradation by diphenylfurans. *Nucleic Acids Research* **2020,** *48* (14), 7653-7664.
44. Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A., QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry* **2014,** *57* (12), 4977-5010.
45. Gedeck, P.; Rohde, B.; Bartels, C., QSAR − How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *Journal of Chemical Information and Modeling* **2006,** *46* (5), 1924-1936.
46. Gleitsman, K. R.; Sengupta, R. N.; Herschlag, D., Slow molecular recognition by RNA. *RNA* **2017,** *23* (12), 1745-1753.
47. Stumpfe, D.; Bajorath, J., Exploring Activity Cliffs in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012,** *55* (7), 2932-2942.
48. González-Díaz, H.; Bonet, I.; Terán, C.; De Clercq, E.; Bello, R.; García, M. M.; Santana, L.; Uriarte, E., ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *European Journal of Medicinal Chemistry* **2007,** *42* (5), 580-585.
49. Devillers, J., QSAR Modeling of Large Heterogeneous Sets of Molecules. *SAR and QSAR in Environmental Research* **2001,** *12* (6), 515-528.
50. Lagunin, A. A.; Geronikaki, A.; Eleftheriou, P.; Pogodin, P. V.; Zakharov, A. V., Rational Use of Heterogeneous Data in Quantitative Structure–Activity Relationship (QSAR) Modeling of Cyclooxygenase/Lipoxygenase Inhibitors. *Journal of Chemical Information and Modeling* **2019,** *59* (2), 713-730.
51. Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H., Navigating structure–activity landscapes. *Drug Discovery Today* **2009,** *14* (13), 698-705.
52. Neyman, J., On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. In *Breakthroughs in Statistics: Methodology and Distribution*, Kotz, S.; Johnson, N. L., Eds. Springer New York: New York, NY, 1992; pp 123-150.

53. Polikar, R., Ensemble Learning. In *Ensemble Machine Learning: Methods and Applications*, Zhang, C.; Ma, Y., Eds. Springer US: Boston, MA, 2012; pp 1-34.
54. Umuhire Juru, A.; Cai, Z.; Jan, A.; Hargrove, A. E., Template-guided selection of RNA ligands using imine-based dynamic combinatorial chemistry. *Chemical Communications* **2020,** *56* (24), 3555-3558.
55. Topliss, J. G.; Edwards, R. P., Chance factors in studies of quantitative structure-activity relationships. *Journal of Medicinal Chemistry* **1979,** *22* (10), 1238-1244.
56. Alexander, D. L. J.; Tropsha, A.; Winkler, D. A., Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling* **2015,** *55* (7), 1316-1322.
57. Gramatica, P.; Sangion, A., A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. *Journal of Chemical Information and Modeling* **2016,** *56* (6), 1127-1131.
58. Zou, H.; Hastie, T., Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2005,** *67* (2), 301-320.
59. 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. *Journal of Chemical Information and Computer Sciences* **2003,** *43* (6), 1947-1958.
60. Zhou, J.; Le, V.; Kalia, D.; Nakayama, S.; Mikek, C.; Lewis, E. A.; Sintim, H. O., Diminazene or berenil, a classic duplex minor groove binder, binds to G-quadruplexes with low nanomolar dissociation constants and the amidine groups are also critical for G-quadruplex binding. *Molecular BioSystems* **2014,** *10* (10), 2724-2734.
61. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T., Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **2007,** *8* (1), 25.
62. Guha, R.; Van Drie, J. H., Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *Journal of Chemical Information and Modeling* **2008,** *48* (3), 646-658.
63. Bajusz, D.; Rácz, A.; Héberger, K., Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015,** *7* (1), 20.
64. Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A., Data Set Modelability by QSAR. *Journal of Chemical Information and Modeling* **2014,** *54* (1), 1-4.
65. Iyer, P.; Stumpfe, D.; Bajorath, J., Molecular Mechanism-Based Network-like Similarity Graphs Reveal Relationships between Different Types of Receptor Ligands and Structural Changes that Determine Agonistic, Inverse-Agonistic, and Antagonistic Effects. *Journal of Chemical Information and Modeling* **2011,** *51* (6), 1281-1286.