Quantitative Structure Activity Relationship (QSAR) study predicts small molecule binding to RNA structure

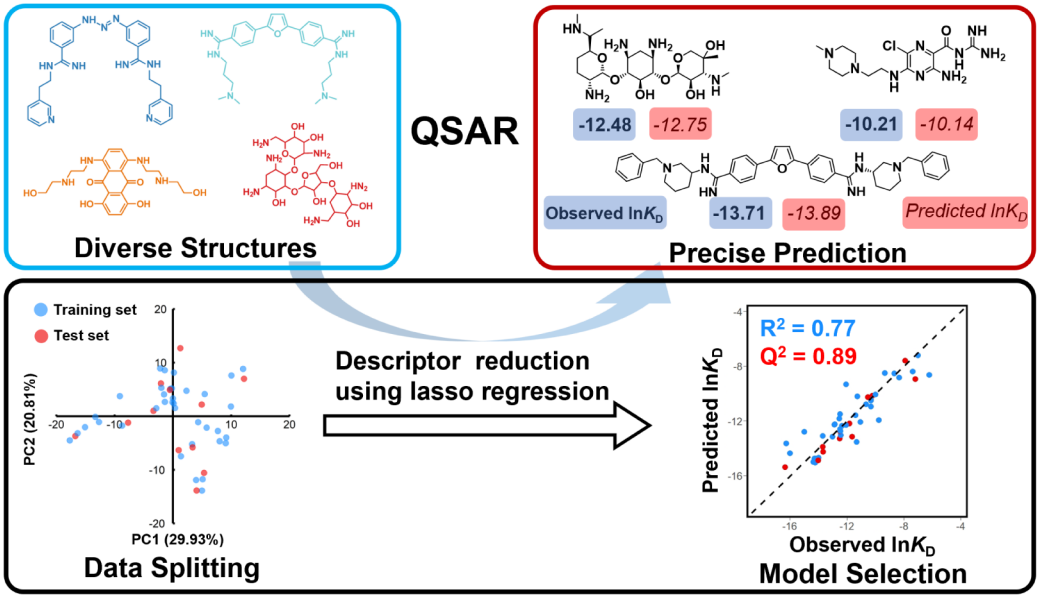
Zhengguo Cai; Martina Zafferani; Olanrewaju M. Akande; Amanda E. Hargrove*

Social Science Research Institute, 140 Science Drive, Durham, NC, 27708, USA

Department of Chemistry, Duke University, 124 Science Drive, Durham, NC 27708, USA

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

TOC



Quantitative structure-activity relationship models built on diverse scaffolds of RNA-targeted small molecules accurately predict binding affinities and kinetic rate constants.

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 challenges in the determination of high-resolution RNA structures and a limited understanding of the parameters that drive RNA recognition by small molecules, including a lack of validated quantitative structure-activity relationships (QSAR). Herein, we developed QSAR models that quantitatively predict both thermodynamic and kinetic-based binding parameters of small molecules and the

HIV-1 TAR model RNA system. A set of small molecules bearing diverse scaffolds was screened against the HIV-1-TAR construct using surface plasmon resonance, which provided the binding kinetics and affinities. The data was then analyzed using multiple linear regression (MLR) combined with feature selection to afford robust models for binding of diverse RNA-targeted scaffolds. The predictivity of the model was validated on untested small molecules. The QSAR models presented herein represent the first application of validated and predictive 2D-QSAR using multiple scaffolds against an RNA target. We expect the workflow to be generally applicable to other RNA structures, ultimately providing essential insight into the small molecule descriptors that drive selective binding interactions and, consequently, providing a platform that can exponentially increase the efficiency of ligand design and optimization without the need for high-resolution RNA structures.

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.[3-6] The implications of these RNAs in disease pathogenesis underscore their potential roles as drug targets. To date, small molecules have been used to target various ncRNAs from several different organisms, including mammals, viruses, bacteria, and fungi.[7-18]

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. 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.[19] Additionally, characterization of RNA-targeted small molecules often disregards binding kinetics, precluding a full understanding and optimization of binding behaviors of a compound. Many protein-targeted drugs are characterized by slow dissociation processes and prolonged target occupancy, supporting the significance of binding kinetics for *in vivo* activity.[20] 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,[21, 22] yet few studies have demonstrated how to intentionally optimize RNA binding kinetics.[23] Overall, there are clear unmet needs in identifying potential RNA-targeted chemical probes and to rationally design small molecules with desired binding behaviors, including appropriate binding kinetics.

To fully access the numerous potentially-druggable RNA targets, 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.[24-27] Among multiple computational tools, quantitative structure-activity relationship (QSAR) studies can pinpoint guiding principles for a specific target by correlating the experimentally observed binding properties with the molecular descriptors of the ligands.[28-30] 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.[31-33] While significant work has been done to explore key descriptors involved in RNA recognition,[34-36] this existing data cannot be used as input for a QSAR approach targeting a specific RNA structure, as these data are derived from disparate methods and RNA targets.

Herein, we build a general workflow utilizing QSAR as predictive platform 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 molecules bearing multiple scaffolds via surface plasmon resonance (SPR). Model building was accomplished by combining representative data splitting, descriptor selection, and linear regression. Post-modeling assessment validated the statistical assumption for linear regression and defined the specific applicable domain for the QSAR model in future use. To the best of our knowledge, this constitutes the first example of a systematic empirical QSAR study conducted on various scaffolds against a specific RNA target. We anticipate that this framework can be readily extended to different RNA targets 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

Selection of RNA target and small molecule training set: 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 process.[12, 37-39] In total, we selected 48 compounds in this study, including 29 reported TAR ligands and 19 compounds with known RNA-targeted scaffolds. These ligands could be classified into 5 categories, namely aminoglycosides (AGs), dimethyl amilorides[40, 41] (DMAs), diphenyl furans[42, 43] (DPFs), diminazenes[44] (DMZs) 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.

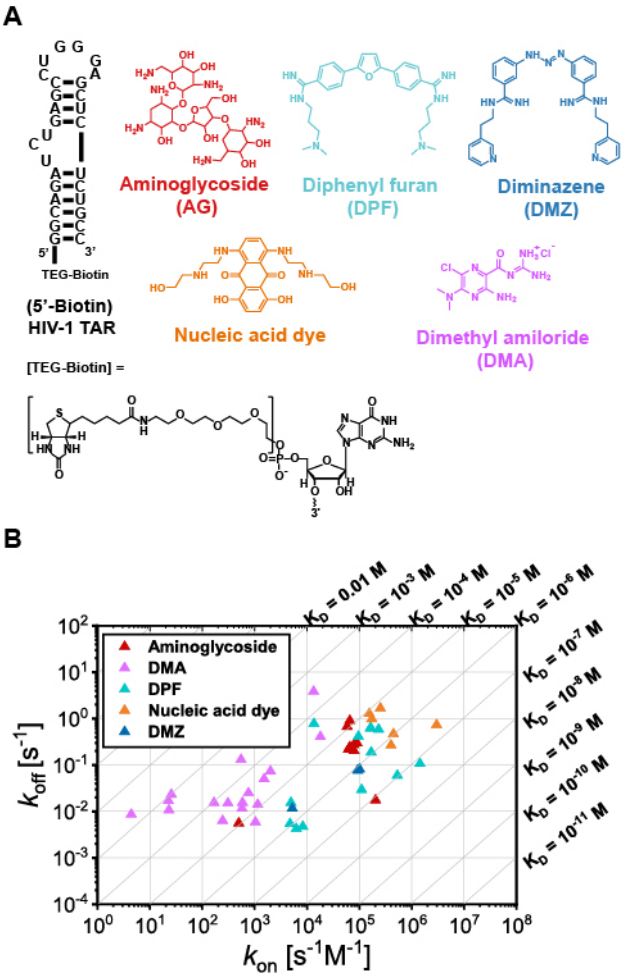
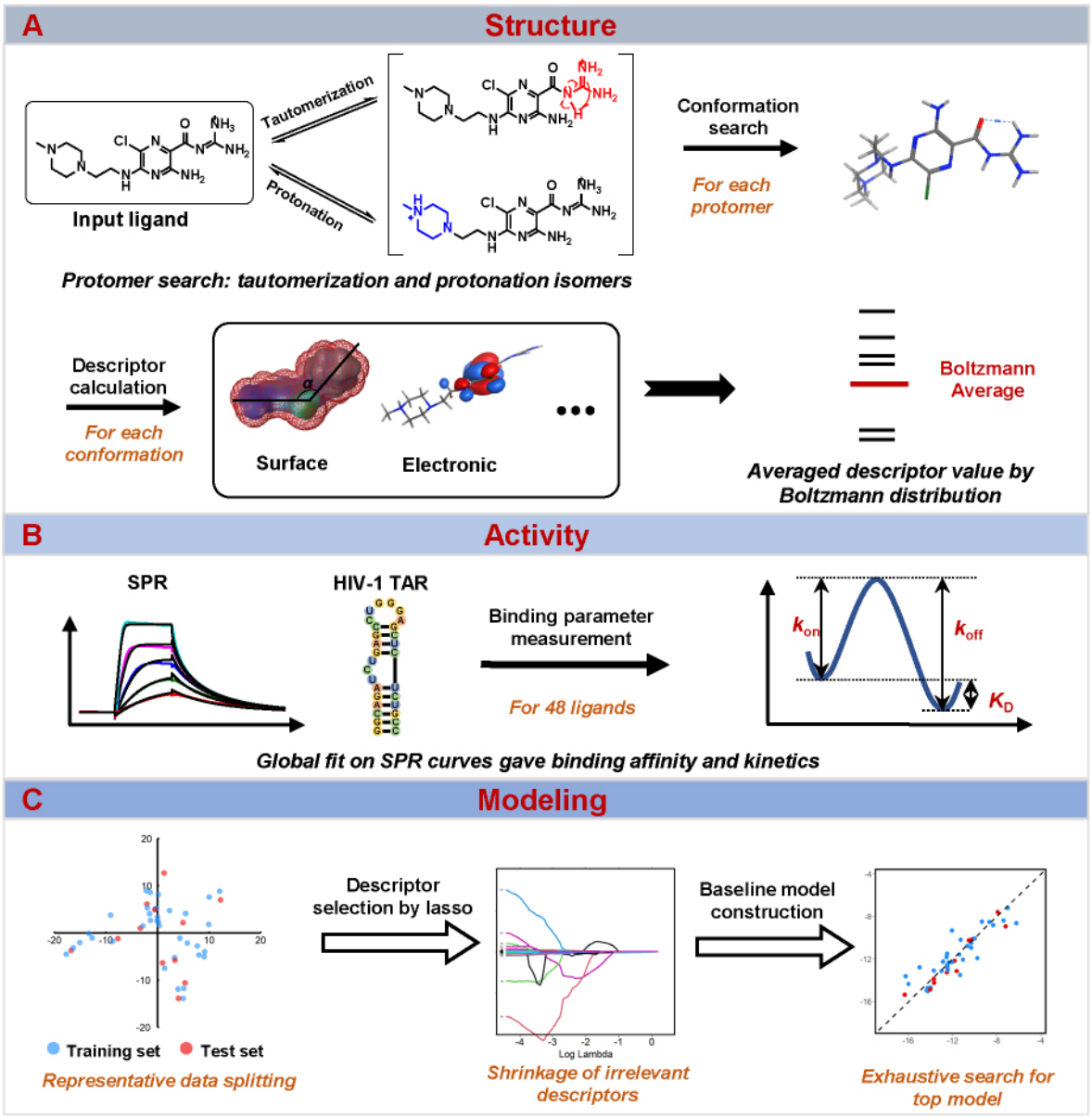


Figure 1 A. Sequence and structure of 5' biotinylated HIV-1 TAR and representative chemical structures of the scaffolds used in this work. **B**. Kinetics map of 48 tested ligands, represented on 10-based logarithmic coordinates. The diagonal lines represent K_D values calculated from k_{off}/k_{on}. Units of three parameters are shown. The rest of study used values based on these units.

Calculations of molecular descriptors: To begin, we obtained molecular information for 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, which represents topological connections within a molecule, was calculated using graph theory matrices, which lays the foundation of many other descriptors including related adjacency distance matrices as well as surface properties. In addition, many QSAR expressions in previous reports suggest that ligand binding preferences originate from non-covalent interactions exerted in the micro-space of the ligand.[45] Hence conformation-dependent 3D descriptors were included to account for the spatial environment of the ligands, such as partial charges and potential energy. In total, we calculated 435 descriptors of each ligand.

We also considered whether multiple species of a given molecule may exist at experimental conditions (**Panel A**, **Scheme 1**). Specifically, we evaluated protonation and tautomerization states for each ligand by distribution ratio as their population representation. For each state, potential conformations within 3 kcal/mol of the lowest energy conformation, as determined by 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 ratio mentioned before. 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 for "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.

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**). 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.[46]

We next compared our kinetics data to a previous survey that showed RNA ligand association was generally slower than that for protein.[47] 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**).[47] The overall association rate constant of an RNA-ligand pair for all three RNA-ligand sets listed in **Table 1** (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 $M^{-1}s^{-1}$).[47] 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 between sets 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 [47].

	k_{on} (M^{-1} s^{-1})	k_{off} (s^{-1})	K_d (M)
RNA (in vitro-selected) - ligand (N=13)[47]	**8.1×10^4**	**6.3×10^{-2}**	**4.3×10^{-7}**

RNA (naturally occurring) - ligand (N=24)[47]	**5.5×10^{4}**	**1.9×10^{-2}**	**3.0×10^{-7}**
HIV-1 TAR - ligand (N=48, used in this work)	**3.8×10^{4}**	**7.9×10^{-2}**	**5.0×10^{-6}**

QSAR modeling: baseline model construction

Data refinement

We used the log transformed versions of K_D, k_{on} and k_{off} as our response variables, as the transformed versions yielded residuals that better satisfy the normality assumption of linear regression models. To mitigate the redundancy of constant and intercorrelated descriptors, a descriptor pre-reduction was applied. First, constant descriptors that have more than 80% compounds sharing the same value were deleted.[48] Next, intercorrelation between every descriptor pair was calculated by Pearson correlation coefficient (ρ). High intercorrelation ($\rho > 0.95$ or $\rho < -0.95$) between descriptors can cause unstable estimation of regression coefficients, sign-change problems and insignificance of regression coefficients.[49] Therefore, multicollinearity (the occurrence of high intercorrelations among two or more descriptors) terms need to be deleted before multiple linear regression. Descriptors intercorrelated with multiple descriptors were deleted one-by-one based on the maximal number of multicollinearity terms. After several rounds, the maximal number of multicollinearity terms for any descriptor would be one, namely only pairwise intercorrelations left. In the remaining pairwise intercorrelations, the term with lower correlation to the response variable was deleted. The above procedure afforded 193 refined descriptors in the $\ln K_D$ and $\ln k_{on}$ datasets, and 191 in the $\ln k_{off}$ dataset.

Representative data splitting by Kennard-Stone algorithm

A key consideration for QSAR with diverse substrates is the continuity of the energy landscapes created by the ligands, i.e. whether gradual changes in ligand properties are smoothly plotted along the target activity function.[30, 50] 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.[51-53] 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.[54] For the model trained from the training set to be used to predict

unseen data in the test set, the distribution of the training set and test set molecules must be representative of the entire sample. To this purpose, we first applied principal component analysis (PCA) to reduce the dimension of the descriptor space. Then, the Kennard-Stone algorithm[55] was utilized to maximize the representativeness of the selected sample with the whole dataset, and the slightly different descriptor space between $\ln K_D$/$\ln k_{on}$ and $\ln k_{off}$ dataset did not alter the sampling results. This specific sampling method rather than random splitting was applied here due to the small sample size (48), which can guarantee that representative small molecules are chosen to achieve a uniform representation of the descriptor space, giving more confidence in future predictions of test set molecules that come from the same distribution of the training set (**Figure 2A**). The distribution of corresponding response variables ($\ln K_D$, $\ln k_{on}$ and $\ln k_{off}$) derived from SPR for training and test sets was visualized in a boxplot (**Figure 2B)**. Sampling of $\ln K_D$ dataset over descriptor space resulted in two subsets with the most representative distribution of the response variable, as seen by the similar range and median values. $\ln k_{on}$ has moderately consistent distribution while $\ln k_{off}$ poorly matched the distribution. This result indicated that the performance order of QSAR models might be $\ln K_D > \ln k_{on} > \ln k_{off}$, given the QSAR assumption that gradual changes in descriptor space lead to gradual changes in response variable. Importantly, the unique test set selected by Kennard-Stone algorithm contains diverse candidates from every scaffold (**Figure 2C**) and is thus a representative subset from the chemical structural perspective (**Supporting information, Section A**).

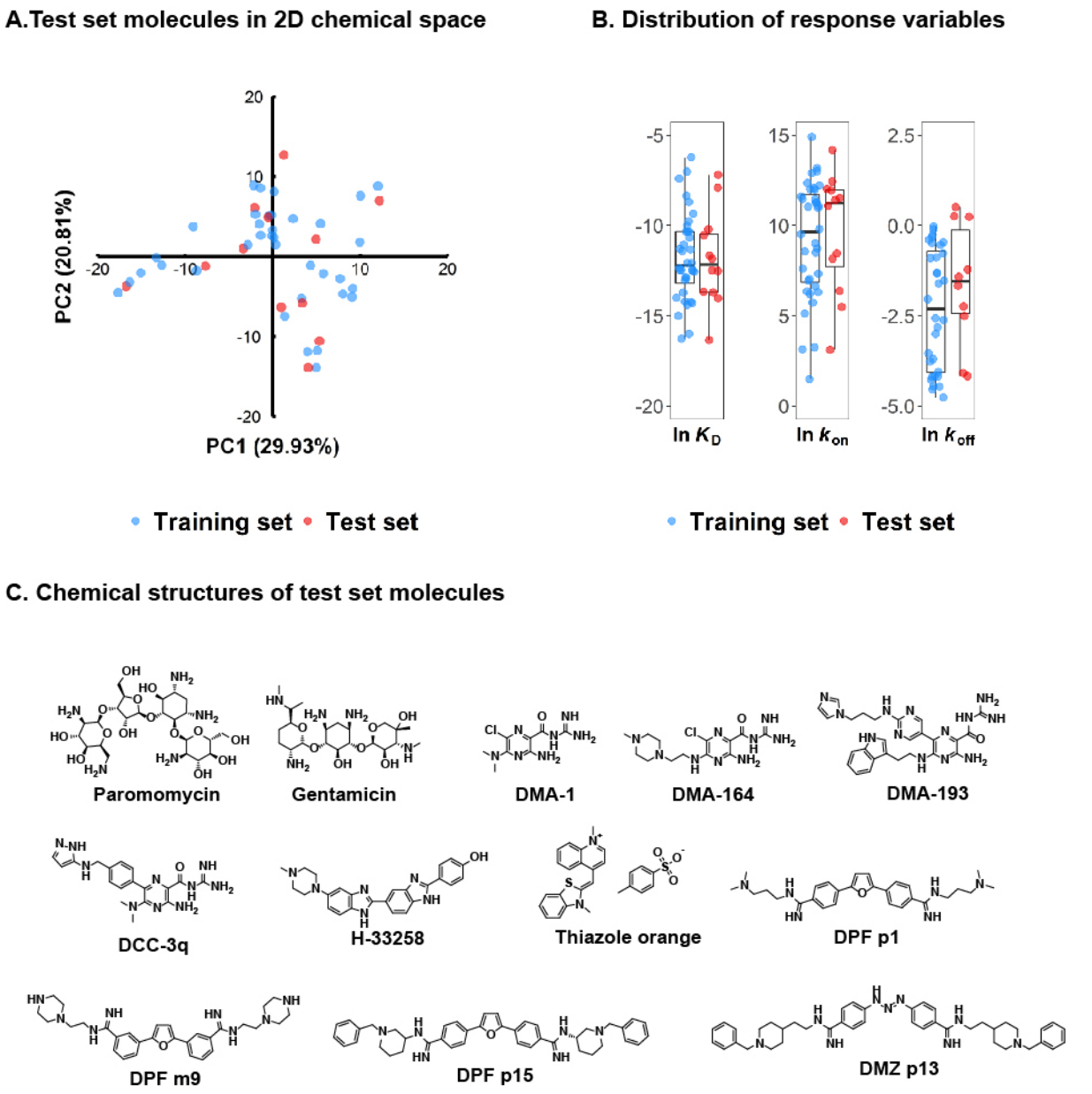


Figure 2 A. Locations of test set molecules in the 2D chemical space constructed from the first two principal components (29.9% and 20.8% of variance, respectively) of the whole dataset. **B.** Distribution of response variables for the test and training set molecules. **C.** Chemical structures of the test set molecules that selected with Kennard-Stone algorithm, the slightly different descriptor space between lnK_D/lnk_{on} and lnk_{off} dataset did not alter the sampling result.

QSAR model development and interpretation

To obtain a predictive and interpretable model, we used multiple linear regression (MLR) in this QSAR study, followed by an assumption evaluation. Due to the limited observations but large number of descriptors, classical MLR could not afford a unique close-form solution. To reduce dimension of the data and find the most relevant descriptors, we

applied least absolute shrinkage and selection operator (lasso) for descriptor selection prior to MLR.[56] Lasso has been widely used in QSAR studies to control the model complexity and increase the performance by applying a penalty constraint to the loss function that needs to be minimized during modeling.[57, 58] Specifically, a hyperparameter λ controls the model complexity as larger λ leads to more descriptor shrinkage. The operator can remove irrelevant descriptors by shrinking the regression coefficients to zero and keeping the most relevant ones. After descriptor selection by lasso, exhaustive searches for all combinations from selected descriptors using MLR was performed. The maximal number of descriptors in a MLR model was set as seven based on the Topliss rule,[59] namely that at least five compounds in the training set were required for adding an extra descriptor in the QSAR model. This exhaustive search afforded multiple model candidates, which were further screened by their performance on training and test sets, as well as statistical significance (*p*-value) of each descriptor involved. Additionally, the principle of "Occam's razor" was followed to choose the model with fewer descriptors if two have similar level of performance.[60]

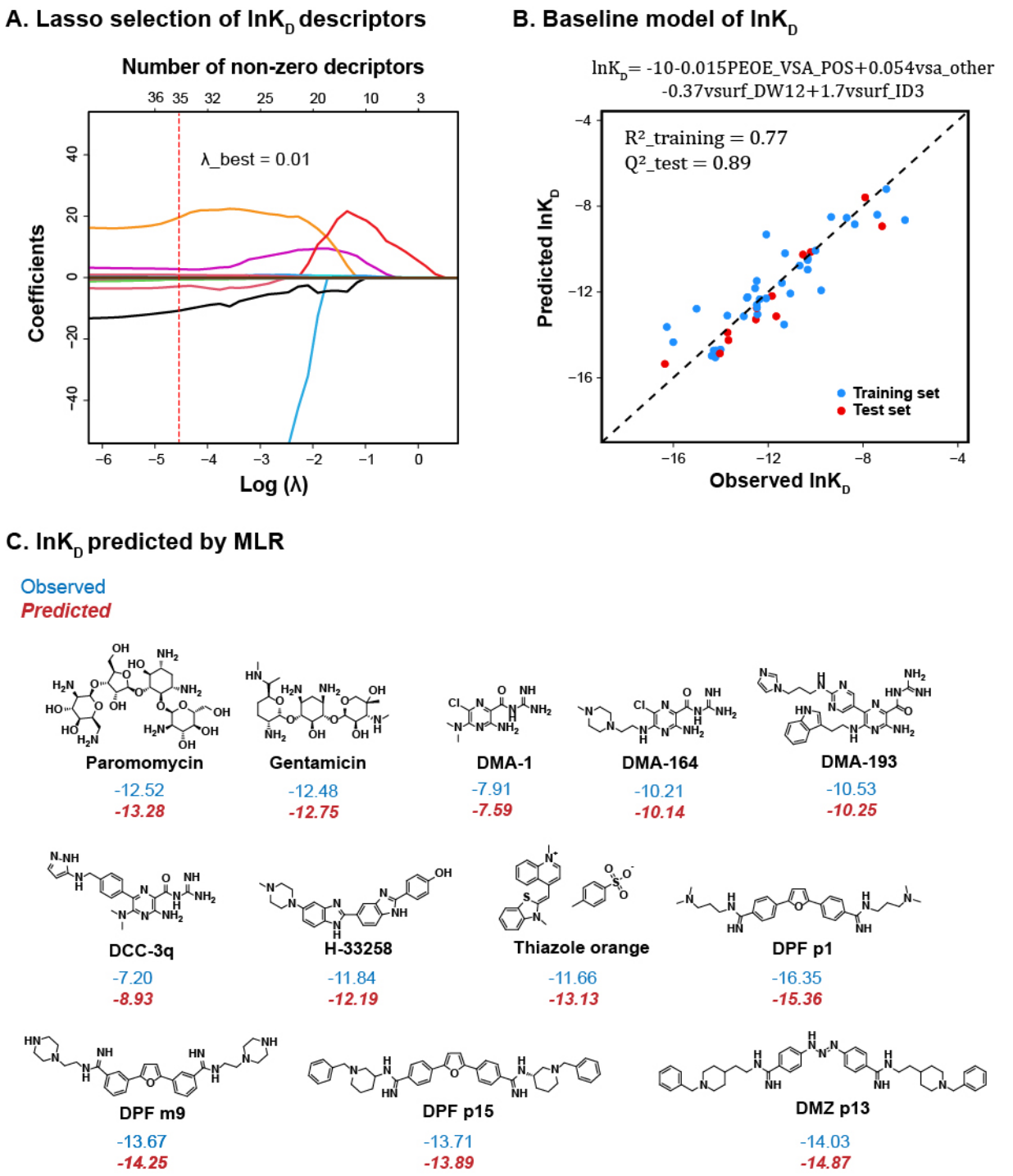


Figure 3 A. Coefficients of lnK_D descriptors were shrunk as λ increased using lasso regression, each curve with different color represented a descriptor coefficient shrinkage, the top x-axis showed the number of descriptors with non-zero coefficients at specific λ value that indicated by the bottom x-axis. The best λ value (0.01) was determined by the 5-fold cross validation. **B.** Observed lnK_D (both training and test set) was plotted with the value predicted by the MLR baseline model shown at top. **C.** Small molecules from the test set along with respective MLR-predicted lnK_D value (in red italics) versus the observed values (in blue).

In detail, for lnK_D modeling, lasso selection was applied to gradually shrink the size of the descriptor set, as hyperparameter λ increases (**Figure 3A**). The best λ was determined to be 0.01, as the result of 5-fold cross validation that aimed at minimizing the prediction biases or the mean cross-validated error. Using this λ value, the number of descriptors was shrunk to 35. These 35 descriptors formed the new descriptor space for exhaustive model search, from the simplest 2-parameter linear model to the most complex 7-parameter linear model. These model candidates were first screened by their performance on the training and test set (R^2 > 0.75, Q^2 > 0.75), then the statistical significance of each descriptor for explaining the model (*p*-value < 0.05).

The final model based on our selection process (**Figure 3B**) was found with below expression, which predicted lnK_D values of our structurally diverse test molecules with high accuracy (**Figure 3C**):

$$lnK_D = -10 - 0.015PEOE\_VSA\_POS + 0.054vsa\_other - 0.37vsurf\_DW12 + 1.7vsurf\_ID3 \ (R^2_{training} = 0.77, Q^2_{test} = 0.89)$$

The model included four physicochemical descriptors (PEOE_VSA_POS, vsa_other, vsurf_DW12 and vsurf_ID3) with their physical meaning shown in **Table 2**. The negative coefficient of PEOE_VSA_POS explicitly suggested that non-negative electrostatic properties of the molecule helped to improve lnK_D, which is consistent with the fact that RNA is overall negatively charged. Additionally, vsa_other describes the sum of van der Waals surface area of atoms typed as “other”. These “other” atoms are not H-bond acceptors, H-bond donors, acidic, basic, polar or hydrophobic residues, thus mostly referring to the surface area of carbon atoms near oxygen, nitrogen and halide atoms.[61] According to the model, decrease in vsa_other could favor tight binding for HIV-1 TAR. vsurf_DW12 is the contact distance between the physical location of first two hydrophilic energy interaction minima when a hydrophilic probe (OH2) interacts with the target molecule. The negative correlation of this descriptor indicated that high-affinity ligands have energy minima which are relatively distant from each other in 3D space, which is also consistent with a previous report.[62] Interaction energy (integy) moment is a type of descriptor that resembles dipole moment, but instead of describing separation of the partial charge, integy moments express the unbalance between the center of mass of a molecule and the barycenter of its hydrophilic or hydrophobic (vsurf_ID) regions.

Specifically for vsurf_ID3, it is the vector pointing from the center of mass to the center of the hydrophobic regions that is calculated at -0.6 kcal/mol energy level.[63] The positive correlation of this descriptor to lnK_D suggested tight binding could be achieved by small molecules that possess hydrophobic moieties that are either close to the center of mass or they balance at opposite ends of the molecule.

Table 2 Descriptors involved in 3 models and their physical meanings

Descriptor name	Physical meaning
PEOE_VSA_POS	Total positive van der Waals surface area.
vsa_other	van der Waals surface area (Å^2) of atoms typed as "other". Other: not H-bond acceptors, H-bond donors, acidic, basic, polar or hydrophobic residues.
vsurf_DW12	Contact distances of vsurf_EWmin1 and vsurf_EWmin2, vsurf_EWmin describes the lowest hydrophilic energy representing the distances, between the best three local minima of interaction energy when a water probe (OH2) interacts with the target molecule.
vsurf_ID3	Hydrophobic integy moment calculated at -0.6 kcal/mol energy level.
GCUT_PEOE_0	The GCUT descriptors are calculated from the eigenvalues of a modified graph distance adjacency matrix. Each (i,j) entry of the adjacency matrix takes the value 1/sqr(d_{ij}) where d_{ij} is the (modified) graph distance between atoms i and j. The diagonal takes the value of the PEOE partial charges. The resulting eigenvalues are sorted and the smallest (GCUT_PEOE_0), 1/3-ile, 2/3-ile and largest eigenvalues are reported.
vsurf_DD23	Contact distances of vsurf_EDmin2 and vsurf_EDmin3, vsurf_EDmin describes the lowest hydrophobic energy representing the distances, between the best three local minima of interaction energy when a hydrophobic probe (DRY) interacts with the target molecule.
a_base	Number of basic atoms.
a_nN	Number of nitrogen atoms.
vsurf_DD13	Contact distances of vsurf_EDmin1 and vsurf_EDmin3.

To investigate how molecular descriptors quantitatively impact the association process of HIV-1 TAR ligands, we performed lnk_{on} modeling. Similarly, lasso selection afforded 16 descriptors after regression coefficients shrinkage with optimized λ equaled to 0.22

(**Figure S2A**). Further model search led to the identification of the model below (**Figure S2B**):

$$lnk_{on} = -12 - 27GCUT\_PEOE\_0 - 0.093vsa\_other + 0.42vsurf\_DD23 + 0.59vsurf\_DW12 \ (R^2_{training} = 0.77, Q^2_{test} = 0.77)$$

This model included four physicochemical descriptors, namely GCUT_PEOE_0, vsa_other, vsurf_DD23 and vsurf_DW12 (**Table 2**). Two of them (vsa_other and vsurf_DW12) also appeared in the lnK_D model, consistent with the correlation between lnK_D and lnk_{on} ($\rho_{lnK_D,lnk_{on}} = -0.82$). GCUT_PEOE_0 encodes information of partial charge and atomic connectivity, supporting an important role for partial charge distribution on on-rate constants, though it is hard to directly deduce chemically intuitive information as it is the mathematical representation of atomic partial charge calculated from partial equalization of orbital electronegativities (PEOE) method combining atomic connectivity. The negative coefficient of this descriptor suggested decreased value of GCUT_PEOE_0 could accelerate the association process. The contribution of vsa_other and vsurf_DW12 followed the same trend identified in lnK_D model, namely lower van der Waals surface area for atoms typed as "other" and more distant distribution of hydrophilic interaction energy minima would benefit fast association, thus favoring tighter binding. Finally, vsurf_DD23 is another surface property descriptor, describing the physical distance between the location of the second-lowest and third-lowest hydrophobic energy interaction that measured by a specific hydrophobic probe (DRY).[64] The positive coefficient of this descriptor signified that by increasing the distance between these energy minima sites, the compounds were predicted to have faster association processes.

We next assessed whether the above workflow could afford a predictive lnk_{off} model. In this case, lasso selection refined the descriptor set to only four descriptors, using the cross-validated best λ value (λ = 0.50). This shrinkage appeared to be too stringent as lasso regression equally penalized all the descriptor coefficients and suffered with biased estimates at this situation, namely descriptors with large coefficients were over-penalized and descriptors with small coefficients were not detected.[65] Specifically, the combination of these four features poorly explained the data ($R^2_{training} = 0.43, Q^2_{test} = 0.38$). We

adjusted the λ value (λ = e^{-2} ~ e^{-4}) as a less stringent shrinkage to include more descriptors (**Figure S2C**) and found that when the descriptor vsurf_DD13 was included, the model performance could be greatly enhanced. The final model(**Figure S2D**) we found for explaining lnk_{off} is shown below:

$$lnk_{off} = 2.0 - 0.69a\_base - 0.42a\_nN + 0.27vsurf\_DD13 \ \ (R^2_{training} = 0.64, Q^2_{test} = 0.61)$$

This model matched that from an exhaustive search result using all 191 descriptors, suggesting that lasso was able to pick significant variables but sometimes needs fine tuning of the hyperparameter λ. In this model, the negative correlation between number of basic atoms (a_base) and the dissociation rate constants suggested that increased electrostatic interactions can slow ligand dissociation. Introduction of nitrogen-containing groups may also increase the retention time as a negative correlation was found between number of nitrogen atoms (a_nN) and the dissociation rate constants. The correlation between these two descriptors was low ($\rho_{a_{base},a\_nN} = 0.23$), indicating that they contribute to the rate constant differently, probably through electrostatic interactions (a_base) and π-π stacking from nitrogen-containing heterocyclic rings (a_nN), respectively. Additionally, vsurf_DD13 positively correlated with the off-rate constant, suggesting that decreasing the physical distance between the lowest and third-lowest hydrophobic energy interaction site will slow dissociation. Overall, however, regressions using lnk_{off} data could not afford a baseline model with comparable performance as above two models. This might be caused by a number of factors, including the poor representativeness of the selected subset in terms of the response variable distribution (**Figure 2B**) and the larger measurement variance as seen from different SPR replicates. Larger datasets are likely needed to precisely model the off-rate constants. Nonetheless, this data did show that QSAR can yield promising model for understanding dissociation process of HIV-1 TAR: small molecule recognition, assisting the design of ligands with prolonged retention time over the target. The success of training a predive and interpretable QSAR model for explaining different binding parameters of HIV-1 TAR ligands suggested that QSAR study could be a lens to investigate complicated macromolecular binding event and a guide for molecular design with specific response property.

Comparison with nonparametric ensemble tree methods

To further evaluate the performance of MLR baseline models, we compared them to models constructed by ensemble tree methods, such as bagging and boosting. Tree methods use a flow-chart like structure to make predictions (leaf) based on the outcomes (branch) of the tests (nodes).[66] By combining multiple decision tree models and making predictions from the averaged results, ensemble tree methods have been identified to improve the model performance and/or overcome the variance-bias tradeoff in prediction.[67] However, the ensemble process increases the difficulty of explicit model interpretation when compared to the single parametric model such as the one given by MLR due to its aggregated model complexity.

Table 3 Comparison of model performance built by different methods

	$\ln K_D$		$\ln k_{on}$		$\ln k_{off}$	
	Train	test	Train	test	Train	test
Decision tree	0.90	0.78	0.87	0.86	0.73	-0.1
Decision tree bagging	0.91	0.89	0.83	0.71	0.94	0.21
Random forest	0.90	0.87	0.90	0.70	0.89	0.39
Boosting	0.92	0.87	0.92	0.73	0.90	0.25
MLR	0.77	0.89	0.77	0.77	0.64	0.61

We started our comparison by building a single decision tree, which was the foundation of other ensemble-based models. Unlike MLR that needs a normality assumption to explain the randomness of the error (see Model assessment and applicable domain below), decision tree is a nonparametric method that can avoid the risk of mis-specifying these pre-assumptions and probability distributions. The complexity of the decision tree was controlled by the cross-validated error, which afforded us with the best number of terminal nodes in the pruned tree. Decision trees trained on $\ln K_D$ and $\ln k_{on}$ training set gave satisfactory predictions on the corresponding test set (**Table 3**). This result suggested that different scaffolds have distinguished binding affinities and association rate constants that can be revealed by the splitting nodes using existing descriptors. Meanwhile, the poor fitting on the dissociation rate constant indicated that more decisive

descriptors were needed to explain the observations. Parallel training of multiple decision trees over a subset of training data that was generated by bootstrapping (sampling with replacement) gave us bagging models. The optimized number of trees was determined based on the averaged error of samples that were not included in training or out-of-bag samples. Random forest is a special scenario of bagging that in addition of using bootstrapping samples, only a subset of descriptor space will be used for the training of each individual tree. **Figure 4A** shows that when training on $\ln K_D$ data, the out-of-bag error was gradually converged as number of trees increased. **Figure 4B** shows the random forest model trained for $\ln K_D$ using 400 trees. Boosting, however is a sequential training process that the current model trains on the residuals from last model by adding weight to the poorly predicted data point. Similarly, **Figure 4C** shows that loss function (squared error) decreased as the number of above sequential iterations increased, where the optimal iterations (990) could be found by looking at the cross validated error. Out-of-bag error was also plotted. The discrepancy between these two errors suggested the heterogeneity of the data set. **Figure 4D** represents the final boosting model trained for $\ln K_D$ using 990 iterations.

Overall, models trained by above methods with different response variables behaved with the same trend as in MLR, namely their performance order is: $\ln K_D$ models> $\ln k_{on}$ models> $\ln k_{off}$ models (**Table 3**). $\ln K_D$ models showed significant enhancement after the ensemble learning, namely aggregation of multiple weak learners led to a stronger learner, and the prediction accuracy on the test set was comparable to the MLR model. For $\ln k_{on}$, it was interesting that single decision tree with 6 nodes achieved both high training efficiency and prediction accuracy. Further application of the ensemble learning seemed to overfit the data as performance discrepancy between training set and test set data was observed. For this data set, ensemble learning failed to push the predictivity of the model to a higher level when compared to the MLR baseline model. For all $\ln k_{off}$ models the prediction on the test set was not satisfactory, probably due to the lack of decisive descriptors or the poor representativeness of the test set to the training set as seen from the $\ln k_{off}$ distribution (**Figure 2B**).

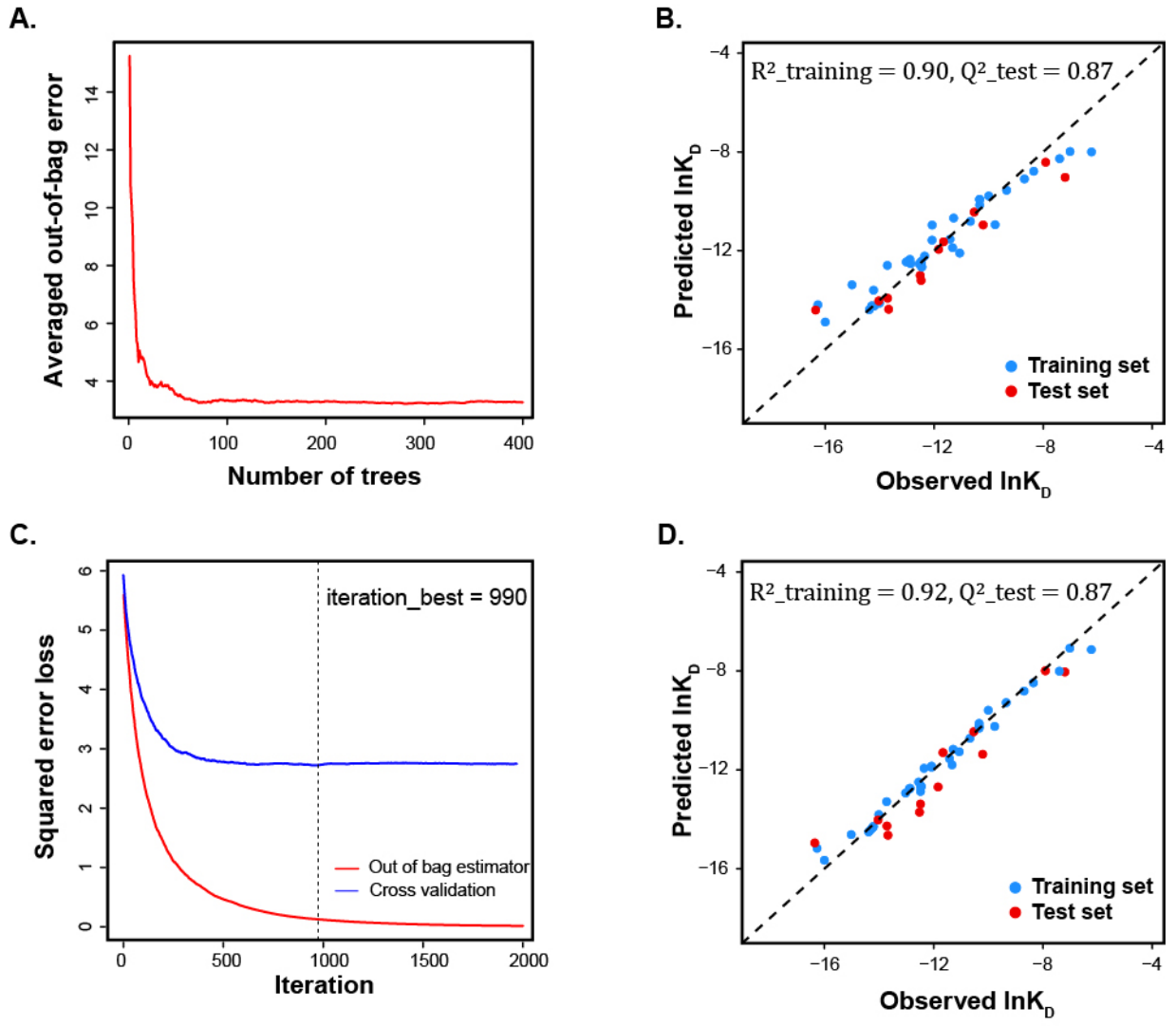


Figure 4 A. Out-of-bag error of random forest model vs. number of trees. **B.** Random forest model of lnK_D built with 400 decision trees. **C.** Squared error loss vs. number of iterations in boosting, two methods (out-of-bag method and cross validation method) were used to determine the best iteration number. D. Boosting model of lnK_D.

Model assessment and applicable domain

To validate the main regression assumption, namely that standardized residuals of MLR should follow a normal distribution, we plotted quantile-quantile (Q-Q) graphs. Q-Q plot is commonly used to compare distribution of two datasets. Herein standard quantiles of the normal distribution were plotted on the x-axis and the standardized residuals from MLR was plotted on the y-axis for comparison. Q-Q plots of all 3 MLR models (**Figure 5A**, **Figure S3A**) showed that residuals from linear regression lined around the 45-degree reference line, indicating the validity of the normality assumption. For the linearity

assumption check, we plotted residuals against each descriptor (**Figure S4**). In such plots, we found that residuals were randomly distributed around zero and no obvious trend could be observed, suggesting that no additional relationship with corresponding descriptor remained in residuals. For the independence and equal variance check, we plotted residuals against the fitted values (**Figure S5**). Similarly, the residuals were located randomly along zero with equal variance, suggesting the validity of the linear regression.

To further evaluate the MLR model for future predictions, we defined a proper range of small molecules that can be applied to the models or the applicable domain. Y-outliers represent data points that have significant deviations on response values that do not follow the general trend of the rest of the data, while influential compounds are those that have large impact on the regression and usually have extreme descriptor values or leverage values (a scoring metric between 0 and 1, large value represents far away the values of the predictor variables for the observation from those of other observations). We generated a Williams plot to identify outliers from the response variable perspective, as well as influential points from the descriptor perspective (**Figure 5B**, **Figure S3B**). In this plot, the leverage value of each compound was plotted against its standardized residuals and y-outliers could be detected if the standardized residuals were higher than the ±3 limit. Potential influential points that have extreme descriptor values could be found by checking leverage values whereas the threshold was set as $3(p+1)/n$ (p is the number of descriptors in MLR model, n is the number of data points). In these 3 Williams plots, we did not observe any outliers from the view of response variable. There is one compound, DMZ p8, that has high leverage values from the training set of lnk_{off} model. However, the fitting on this compound did not further support this is as an influential point. Meanwhile, by looking through the Williams plot, we could find potential inaccurately measured data points. For instance, the Williams plot of lnk_{off} model found that two compounds (DMA-1 and DCC-3k) have large fitting residuals but shared similar descriptor space as their leverage values were both low. In fact, both compounds were measured with much larger dissociation rate constants than other DMAs, indicating potential measurement error. Removal of DCC-3k in the training and DMA-1 in the test set would

increase the R^2(training) from 0.64 to 0.71 and prediction accuracy from 0.61 to 0.70 of the lnk_{off} MLR model.

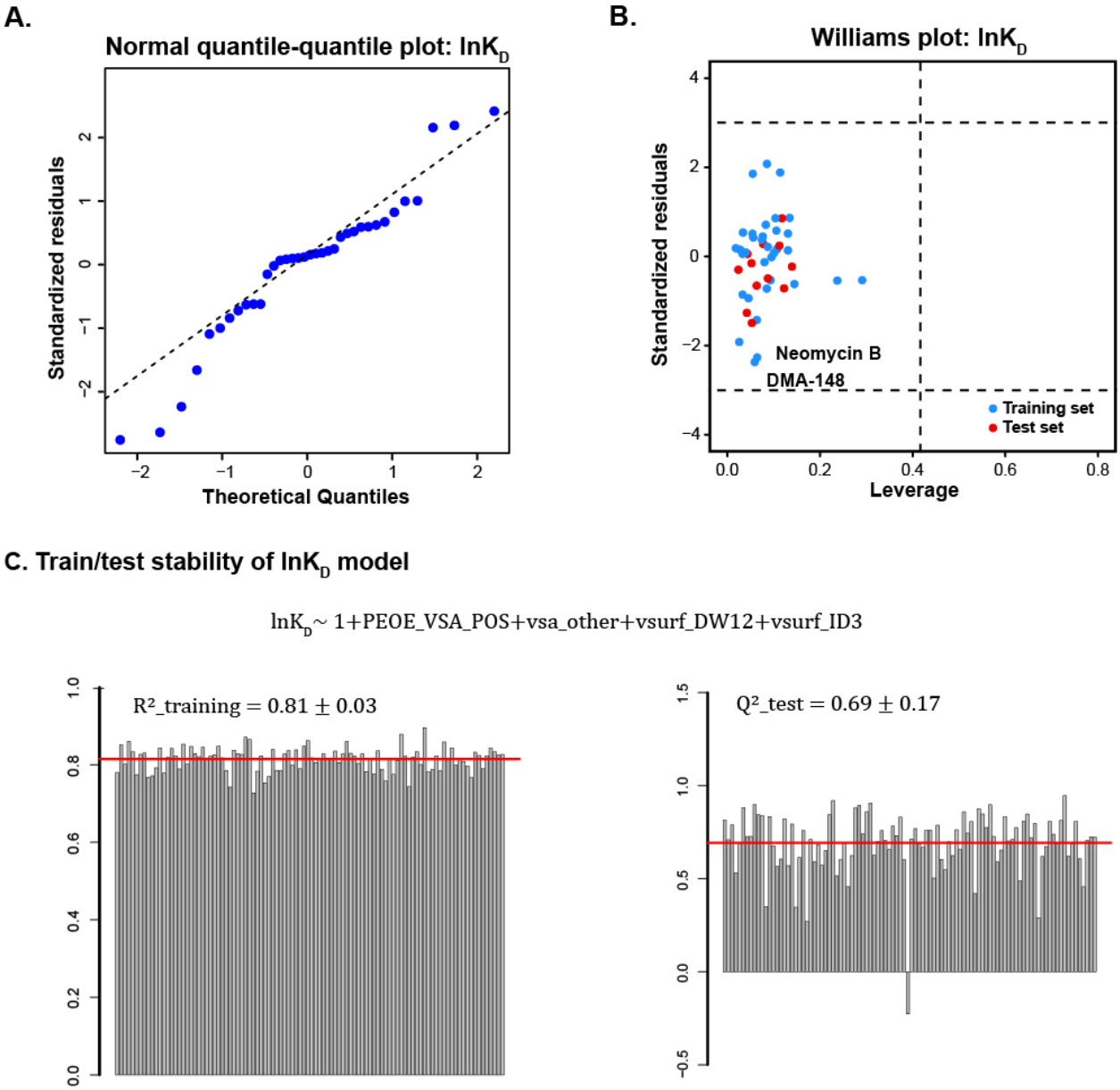


Figure 5 A. Normal quantile-quantile plots of lnK_D model. **B.** Williams plot showed applicable domain of lnK_D model with training and test sets. **C.** Model stability test on lnK_D data using below formula: lnK_D~ 1+ PEOE_VSA_POS + vsa_other + vsurf_DW12 + vsruf_ID3. The training and prediction stability were shown on the left and right, respectively. Each bar represented the result from a random sampling, totally 100 times.

To evaluate the robustness of the model constructed by above descriptors, a training/prediction stability test was performed for each MLR model. In this stability test, a set of 36 molecules were randomly selected as the training set, then a MLR model was

trained using the same descriptors found before on the training set. The prediction accuracy was calculated using the remaining 12 compounds in the test set. By repeating this process, we can test the robustness of identified descriptors for building a well-performed MLR model. In **Figure 5C** and **Figure S6**, the 100 random samplings gave distinctive training/test sets, but models trained with the same set of descriptors afforded high and stable training efficiency and were consistent to the original MLR model. In terms of the prediction accuracy on test sets, we still see overall high Q^2 scores for all of 3 datasets but with higher variance, which might be caused by the extremely unrepresentative data splitting.

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 have been discovered as potential drug targets 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. To this aim, 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 with diverse scaffolds.

By applying a representative data splitting, we trained models from 36 small molecules derived from structural classes (DMZ, DMA, DPF, AG, nucleic acid dyes) as the basis of

our understanding of RNA ligand chemical space. The trained models afforded satisfactory explanations for both binding affinities and kinetics data empirically gathered via SPR. The subsequent prediction of 12 previously untested compounds revealed similar or even higher precision as compared to the well-established ensemble learning-based methods, supporting the power of MLR models to inform compound design. Notably, the accurate prediction of the binding affinity and kinetics of 12 structurally-diverse small molecules not present in the training set underscored the breadth of application of the method to a general small molecule library. The detailed analysis of the descriptor space highlighted by the best models revealed important roles of ligand surface properties and potential charge in RNA recognition of small molecules. Moreover, the MLR model provided quantitative information on how the modification of these descriptors can better aid molecular design and lead optimization. Further evaluation of the applicable domain suggested the proper range of the future small molecules that can be appropriately predicted using these models.

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.

Data availability

Data for this paper, including the diminazene synthesis, structural characterization, surface plasmon resonance methods and sensorgrams, QSAR modeling methods and scripts, and supplementary figures/tables, are available at ESI.

Author contributions

AEH and ZC designed the research. ZC carried out SPR experiments. MZ synthesized the diminazene small molecules. ZC carried out the calculations. ZC, AEH and OMA carried out the analysis and modeling. AEH and OMA supervised the project. AEH and ZC co-wrote the paper. AEH and OMA guided and revised the paper. All authors read and commented on the manuscript.

Conflicts of interest

There are no conflicts to declare.

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. We thank Duke graduate student Jiayue Xu from interdisciplinary data science for constructive discussions and suggestions. 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. The ENCODE Project Consortium, *Nature*, 2007, **447**, 799-816.
2. Thomas R. Cech and Joan A. Steitz, *Cell*, 2014, **157**, 77-94.
3. Q. Ji, L. Zhang, X. Liu, L. Zhou, W. Wang, Z. Han, H. Sui, Y. Tang, Y. Wang, N. Liu, J. Ren, F. Hou and Q. Li, *British Journal of Cancer*, 2014, **111**, 736-748.
4. R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M.-C. Tsai, T. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. van de Vijver, S. Sukumar and H. Y. Chang, *Nature*, 2010, **464**, 1071-1076.
5. M. Esteller, *Nat Rev Genet*, 2011, **12**, 861-874.
6. J. A. Brown, D. Bulkley, J. Wang, M. L. Valenstein, T. A. Yario, T. A. Steitz and J. A. Steitz, *Nat Struct Mol Biol*, 2014, **21**, 633-640.
7. N. F. Rizvi and G. F. Smith, *Bioorg Med Chem Lett*, 2017, **27**, 5083-5088.
8. M. Matsui and D. R. Corey, *Nat Rev Drug Discov*, 2017, **16**, 167-179.
9. J. R. Thomas and P. J. Hergenrother, *Chemical Reviews*, 2008, **108**, 1171-1224.
10. H. S. Haniff, Y. Tong, X. Liu, J. L. Chen, B. M. Suresh, R. J. Andrews, J. M. Peterson, C. A. O'Leary, R. I. Benhamou, W. N. Moss and M. D. Disney, *ACS Central Science*, 2020, **6**, 1713-1721.
11. F. A. Abulwerdi, W. Xu, A. A. Ageeli, M. J. Yonkunas, G. Arun, H. Nam, J. S. Schneekloth, T. K. Dayie, D. Spector, N. Baird and S. F. J. Le Grice, *ACS Chemical Biology*, 2019, **14**, 223-235.
12. J. Sztuba-Solinska, S. R. Shenoy, P. Gareiss, L. R. H. Krumpe, S. F. J. Le Grice, B. R. O'Keefe and J. S. Schneekloth, *J Am Chem Soc*, 2014, **136**, 8402-8410.
13. M. G. Costales, B. Suresh, K. Vishnu and M. D. Disney, *Cell Chemical Biology*, 2019, **26**, 1180-1186.e1185.
14. A. C. Stelzer, A. T. Frank, J. D. Kratz, M. D. Swanson, M. J. Gonzalez-Hernandez, J. Lee, I. Andricioaei, D. M. Markovitz and H. M. Al-Hashimi, *Nat Chem Biol*, 2011, **7**, 553-559.
15. K. D. Warner, C. E. Hajdin and K. M. Weeks, *Nat Rev Drug Discov*, 2018, **17**, 547-558.
16. L. O. Ofori, J. Hoskins, M. Nakamori, C. A. Thornton and B. L. Miller, *Nucleic Acids Research*, 2012, **40**, 6380-6390.
17. O. Fedorova, G. E. Jagdmann, R. L. Adams, L. Yuan, M. C. Van Zandt and A. M. Pyle, *Nat Chem Biol*, 2018, **14**, 1073-1078.
18. J. A. Howe, H. Wang, T. O. Fischmann, C. J. Balibar, L. Xiao, A. M. Galgoci, J. C. Malinverni, T. Mayhood, A. Villafania, A. Nahvi, N. Murgolo, C. M. Barbieri, P. A. Mann, D. Carr, E. Xia, P. Zuck, D. Riley, R. E. Painter, S. S. Walker, B. Sherborne, R. de Jesus, W. Pan, M. A. Plotkin, J. Wu, D. Rindgen, J. Cummings, C. G. Garlisi, R. Zhang, P. R. Sheth, C. J. Gill, H. Tang and T. Roemer, *Nature*, 2015, **526**, 672-677.
19. B. S. Morgan, J. E. Forte and A. E. Hargrove, *Nucleic Acids Research*, 2018, **46**, 8025-8037.
20. G. K. Walkup, Z. You, P. L. Ross, E. K. H. Allen, F. Daryaee, M. R. Hale, J. O'Donnell, D. E. Ehmann, V. J. A. Schuck, E. T. Buurman, A. L. Choy, L. Hajec, K. Murphy-Benenato, V. Marone, S. A. Patey, L. A. Grosser, M. Johnstone, S. G. Walker, P. J. Tonge and S. L. Fisher, *Nat Chem Biol*, 2015, **11**, 416-423.
21. A. Schoop and F. Dey, *Drug Discovery Today: Technologies*, 2015, **17**, 9-15.

22. E. V. Schneider, J. Böttcher, R. Huber, K. Maskos and L. Neumann, *Proceedings of the National Academy of Sciences*, 2013, **110**, 8081.
23. R. N. Sengupta and D. Herschlag, *Biochemistry*, 2019, **58**, 2760-2768.
24. J.-Y. Guo, Y. Minko, C. B. Santiago and M. S. Sigman, *Acs Catal*, 2017, **7**, 4144-4151.
25. A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.
26. M. B. de Ávila, M. M. Xavier, V. O. Pintro and W. F. de Azevedo, *Biochemical and Biophysical Research Communications*, 2017, **494**, 305-310.
27. Y.-C. Lo, S. E. Rensi, W. Torng and R. B. Altman, *Drug Discovery Today*, 2018, **23**, 1538-1546.
28. P. A. Babu, D. J. Smiles, M. L. Narasu and K. Srinivas, *QSAR & Combinatorial Science*, 2008, **27**, 1362-1373.
29. G. Tugcu and M. Koksal, *Molecular Informatics*, 2019, **38**, 1800090.
30. E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chemical Society Reviews*, 2020, **49**, 3525-3564.
31. I. Maciagiewicz, S. Zhou, S. C. Bergmeier and J. V. Hines, *Bioorg Med Chem Lett*, 2011, **21**, 4524-4527.
32. Y. N. Sekhar, M. R. S. Nayana, N. Sivakumari, M. Ravikumar and S. K. Mahmood, *Journal of Molecular Graphics and Modelling*, 2008, **26**, 1338-1352.
33. P. Setny and J. Trylska, *Journal of Chemical Information and Modeling*, 2009, **49**, 390-400.
34. S. Jamal, V. Periwal, O. Consortium and V. Scaria, *Journal of Cheminformatics*, 2012, **4**, 16.
35. N. F. Rizvi, J. P. Santa Maria, A. Nahvi, J. Klappenbach, D. J. Klein, P. J. Curran, M. P. Richards, C. Chamberlin, P. Saradjian, J. Burchard, R. Aguilar, J. T. Lee, P. J. Dandliker, G. F. Smith, P. Kutchukian and E. B. Nickbarg, *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 2019, **25**, 384-396.
36. B. S. Morgan, J. E. Forte, R. N. Culver, Y. Zhang and A. E. Hargrove, *Angewandte Chemie International Edition*, 2017, **56**, 13498-13502.
37. H.-Y. Mei, M. Cui, A. Heldsinger, S. M. Lemrow, J. A. Loo, K. A. Sannes-Lowery, L. Sharmeen and A. W. Czarnik, *Biochemistry*, 1998, **37**, 14204-14212.
38. L. Zeng, J. Li, M. Muller, S. Yan, S. Mujtaba, C. Pan, Z. Wang and M.-M. Zhou, *J Am Chem Soc*, 2005, **127**, 2376-2377.
39. F. A. Abulwerdi, M. D. Shortridge, J. Sztuba-Solinska, R. Wilson, S. F. J. Le Grice, G. Varani and J. S. Schneekloth, *Journal of Medicinal Chemistry*, 2016, **59**, 11148-11160.
40. N. N. Patwardhan, L. R. Ganser, G. J. Kapral, C. S. Eubanks, J. Lee, B. Sathyamoorthy, H. M. Al-Hashimi and A. E. Hargrove, *MedChemComm*, 2017, **8**, 1022-1036.
41. N. N. Patwardhan, Z. Cai, A. Umuhire Juru and A. E. Hargrove, *Org Biomol Chem*, 2019, **17**, 9313-9320.
42. A. Donlic, B. S. Morgan, J. L. Xu, A. Liu, C. Roble Jr and A. E. Hargrove, *Angewandte Chemie*, 2018, **130**, 13426-13431.

43. A. Donlic, M. Zafferani, G. Padroni, M. Puri and Amanda E. Hargrove, *Nucleic Acids Research*, 2020, **48**, 7653-7664.
44. J. Zhou, V. Le, D. Kalia, S. Nakayama, C. Mikek, E. A. Lewis and H. O. Sintim, *Molecular BioSystems*, 2014, **10**, 2724-2734.
45. A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard and A. Tropsha, *Journal of Medicinal Chemistry*, 2014, **57**, 4977-5010.
46. P. Gedeck, B. Rohde and C. Bartels, *Journal of Chemical Information and Modeling*, 2006, **46**, 1924-1936.
47. K. R. Gleitsman, R. N. Sengupta and D. Herschlag, *RNA*, 2017, **23**, 1745-1753.
48. P. Gramatica, N. Chirico, E. Papa, S. Cassani and S. Kovarich, *Journal of Computational Chemistry*, 2013, **34**, 2121-2132.
49. *Journal*, 2007, DOI: 10.4135/9781412952644.
50. D. Stumpfe and J. Bajorath, *Journal of Medicinal Chemistry*, 2012, **55**, 2932-2942.
51. H. González-Díaz, I. Bonet, C. Terán, E. De Clercq, R. Bello, M. M. García, L. Santana and E. Uriarte, *European Journal of Medicinal Chemistry*, 2007, **42**, 580-585.
52. J. Devillers, *SAR and QSAR in Environmental Research*, 2001, **12**, 515-528.
53. A. A. Lagunin, A. Geronikaki, P. Eleftheriou, P. V. Pogodin and A. V. Zakharov, *Journal of Chemical Information and Modeling*, 2019, **59**, 713-730.
54. J. Bajorath, L. Peltason, M. Wawer, R. Guha, M. S. Lajiness and J. H. Van Drie, *Drug Discovery Today*, 2009, **14**, 698-705.
55. R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137-&.
56. R. Tibshirani, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, **58**, 267-288.
57. Z. Y. Algamal, M. H. Lee, A. M. Al-Fakih and M. Aziz, *Journal of Chemometrics*, 2015, **29**, 547-556.
58. A. Al-Fakih, M. Aziz, H. Abdallah, Z. Algamal, M. H. Lee and H. Maarof, *International Journal of Electrochemical Science*, 2015, **10**, 3568-3583.
59. J. G. Topliss and R. P. Edwards, *Journal of Medicinal Chemistry*, 1979, **22**, 1238-1244.
60. D. M. Hawkins, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 1-12.
61. B. Miryala, Z. Zhen, T. Potta, C. M. Breneman and K. Rege, *ACS Biomaterials Science & Engineering*, 2015, **1**, 656-668.
62. G. Cruciani, P. Crivori, P. A. Carrupt and B. Testa, *Journal of Molecular Structure: THEOCHEM*, 2000, **503**, 17-30.
63. F. A. Bernal and T. J. Schmidt, *Molecules*, 2019, **24**.
64. M. Chen, F. Yang, J. Kang, X. Yang, X. Lai and Y. Gao, *Molecules*, 2016, **21**.
65. H. Wang, G. Li and C.-L. Tsai, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2007, **69**, 63-78.
66. S. R. Safavian and D. Landgrebe, *IEEE Transactions on Systems, Man, and Cybernetics*, 1991, **21**, 660-674.
67. T. G. Dietterich, Berlin, Heidelberg, 2000.