bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Using QSARs for predictions in drug delivery

Edgardo Rivera-Delgado[1]¶, Alison Xin[2]¶, Horst A. von Recum[1]*

[1]Department of Biomedical Engineering, Case Western Reserve University, Cleveland, Ohio, United States of America

[2]Science Research and Engineering Program, Hathaway Brown School, Shaker Heights, Ohio

* Corresponding author

E-mail: horst.vonrecum@case.edu (HvR)

¶ These authors contributed equally to this work

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Drug delivery research is an inherently empirical process, however high-throughput approaches could take advantage of understanding drug/material interactions such as from electrostatic, hydrophobic, or other non-covalent interactions between therapeutic molecules and a drug delivery polymer. Cyclodextrin polymers have been investigated for drug delivery specifically due to their capacity to exploit this affinity interaction to change the rate of drug release. Testing drug candidates; however, for affinity is time-consuming, making computational predictions more effective. One option, molecular "docking" programs, provide predictions of affinity, but lack reliability, as their accuracy with cyclodextrin remains unverified experimentally. Alternatively, quantitative structure-activity relationship models (QSARs), which analyze statistical relationships between molecular properties, appear more promising. Previously constructed QSARs for cyclodextrin are not publicly available, necessitating an openly accessible model. Around 600 experimental affinities between cyclodextrin and guest molecules were cleaned and imported from published research. The software PaDEL-Descriptor calculated over 1000 chemical descriptors for each molecule, which were then analyzed in R to create several QSARs with different statistical methods. These QSARs proved highly time efficient, calculating in minutes what docking programs would take hours to accomplish. Additionally, on test sets, QSARs reached R^2 values of around 0.7-0.8. The speed, accuracy, and accessibility of these QSARs improve evaluation of individual drugs and facilitate screening of large datasets for potential candidates in cyclodextrin affinity-based delivery systems. An app was built to rapidly access model predictions for end users using the "shiny" library in R. To demonstrate the usability for drug release planning, the QSAR predictions were coupled with a mechanistic model of diffusion within the app. Integrating new modules should provide an accessible approach to use other cheminformatic tools in the field of drug delivery.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1. Introduction

Drug delivery is an inherently challenging process to predict, due to situational constraints (device size, polymer properties, diffusivity). Our lab and others have explored use of polymers containing specific chemistries allowing stronger polymer drug interactions to control the rate of drug release. As these affinities are increased the dependence on other factors such as size and diffusivity decrease. This field is termed "Affinity-based drug delivery", and relies on interactions between a drug delivery system and drug molecules, improves effectiveness of medication by extending the duration of drug release and thereby lengthening the duration of the treatment[1]. Mathematical modeling of these affinity systems has shown that the strength of the affinity interaction, the ratio of host binding sites to guest ligands, and the molecular path length of diffusion influence the transport of molecules out of the system. Of these physical forces, the affinity strength plays an important role in the classification of the system and the timescale of drug release[2]. Affinity interaction can be associated with a variety of physical properties, including charge, hydrophobicity, Van der Waals forces, etc. In the fields of biomaterials and drug delivery, affinity delivery has been used with small molecule drugs[3], proteins[1], cytokines, and antibodies[4].

Fig 1. Cyclodextrin complexation.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Our lab tests rings of glucose molecules as affinity hosts called cyclodextrins, which are particularly promising affinity drug delivery hosts due to their structural properties, biocompatibility and versatility. The most common cyclodextrin are composed of a ring 6, 7, or 8 glucose molecules (α, β, and γ-cyclodextrin, respectively), and the conformation of the hydroxyl groups of the ring create a basket-like structure with a hydrophobic interior and hydrophilic interior, allowing for complexation with drug molecules (Figure 1). Additionally, cyclodextrin can be polymerized into a variety of materials, including microparticles, viscous gels, and solid films. Unfortunately, experiments to confirm sustained release from the affinity guest-host system often takes weeks, making testing large numbers of potential candidates for cyclodextrin release systems impractical.

As an alternative to experimental testing, candidate molecules can be analyzed computationally. Predicting the binding affinity between cyclodextrin and drug molecules allows for the processing of molecules on the scale of minutes rather than weeks. There are two major methods for predicting molecular interaction: docking models and QSARs. Docking models use molecular force fields, which simulate interactions and potential energy between atoms. Force field parameters may be derived from experiments, calculations from quantum mechanics, or both[5]. In addition to providing a numeric estimate for binding affinity, docking programs produce visualizations of how molecules interact. QSARs, or Quantitative Structure-Activity Relationship models, statistically predict molecular interactions using molecular descriptors. Molecular descriptors are certain physical or chemical characteristics of molecules that can be evaluated numerically (for example, the number of hydrogen atoms or the length of the longest bond chain). Many different types of regression models and statistical learning methods can be used as QSARs, ranging in complexity from linear models to artificial neural networks[6].

Table 1. Results of previous cyclodextrin QSARs.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

QSAR	R2	Descriptors	Feature selection	Validation			
				Q2	*y-rand*	*AD*	*EV*
Cubist [9]	0.945	Pfizer*	**	**	**	Yes	Yes
Random forest [9]	0.912	Pfizer*	**	**	**	Yes	Yes
Partial least squares (PLS) [10]	0.68	ChemOffice, SYBYL, Pentacle*	Genetic algorithm	0.64	**	Yes	**
PLS [11]	0.74	SYBYL, Pentacle*	Fractional factorial design	0.75	Yes	Yes	**
Multiple linear regression (MLR) [12]	0.943	ISIS/Draw, CODESSA*	Forward selection	0.848			**
MLR [13]	0.841	ISIS/Draw, MOPAC, Web-DRAGON*	Genetic algorithm	0.821	Yes	Yes	**
MLR [14]	0.833	HyperChem, DRAGON*	Forward selection	0.826	Yes	Yes	**
MLR [15]	0.78	SYBYL, MOE, AutoDock Tools, BINANA*	Genetic algorithm	0.82		Yes	**
Artificial neural network [14]	0.957	HyperChem, DRAGON*	Forward selection	0.955	Yes	Yes	**
Support vector machine [16]	0.971	HyperChem, MOE*	Particle swarm	**	**	**	**

*Presence of a paywall, usually due to specialized software that requires a license

**Insufficient verification

None of the models investigated were both fully validated and openly accessible.

Previous investigations have been made on the accuracy of both docking and QSARs in predicting cyclodextrin affinity, but examining a sample of these papers reveals several concerns (Table 1). Notably, all of the investigated models used software hidden behind a paywall or only available with a license[7–14]. Additionally, many models lacked proper verification. Following Tropsha's publication detailing best practices for QSAR development, a

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

completely verified model should undergo leave-one-out cross-validation (LOO-CV) (reported as Q^2), y-randomization, pass a variety of internal accuracy tests, and be analyzed for applicability domain. Additionally, models should be evaluated on multiple test sets as well as a hold-out external validation set[15]. Of the papers investigated, none contained the full set of verification strategies.

In this study, the accuracy and usability of docking and QSARs were compared in order to establish the appropriate framework for the computational design of cyclodextrin based affinity delivery devices. Autodock VINA, an open-source docking program developed by Trott, was used to investigate docking methods[16]. A variety of statistical methods presented in previous cyclodextrin QSARs were also investigated. The performance of QSARs was evaluated on both a standard test set as well as an external validation set to confirm accuracy. Properly evaluating the use of docking and QSARs should improve selection of possible guests for cyclodextrin, reducing the rejection of good candidates (Type II error) and limiting experimental investigation of bad candidates (Type I error).

Finally, though the coded models could be made freely available, understanding the raw script remained a significant obstacle for new users. Additionally, users would have to download multiple files and programs to their own computers, creating potential issues with device compatibility, storage restrictions, processor limitations, etc. To overcome these obstacles and improve accessibility, the models were then integrated into a web application built with the R library “shiny” and then uploaded online. To demonstrate the ease of extendibility of the app and its value in planning drug delivery strategies the results from the QSAR studies were then integrated into a mechanistic model of drug release.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2. Materials and Methods

In order to be accessible, the models use only open-source software. Importing experimental data, cleaning data, and creating QSAR models were performed using R in RStudio. Both the coding language and the IDE are freely downloadable and easily accessible on Windows, Mac OS, and Linux. Descriptors were generated with PaDEL, also freely downloadable and open-source. Only the original observations of cyclodextrin complexation energies remain inaccessible to the public, but this does not have any effect on using the models for new predictions.

I. Dataset

Many of the models in Table 1 work from the same data source, a compilation of α- and β-CD affinities published by Suzuki in 2001[17] (additionally, the sources that cite a different paper by Katritzky ultimately use this same data, as the Katritzky paper cites Suzuki for observations[18]). In addition to Suzuki, we also compiled complexes of α- and β-CD Rekharsky and Inoue and Suzuki[19]. Complexes of γ-CD, missing from the Suzuki dataset and sparse in the Rekharsky and Inoue data, were collected from Connors[20]. Once compiled, the data were cleaned for reliable information, one-to-one cyclodextrin complexes, a temperature of 298 ± 2 K, and a solvent of water with pH 7. To obtain structure-data files (SDFs) of the ligands, the names of the guest molecules were passed through the Chemical Identifier Resolver, a web interface provided by the National Cancer Institute's Computer-Aided Drug Design Group (NCI/CADD). To handle the data, the R packages tidyverse, data.table, XML, RCurl, and Matrix were used[21–25].

Dataset splitting was performed using the R package caret[26]. First, the cleaned data was split between α-, β-, and γ-CD. Structural and activity outliers in each category were removed. Structural outliers were detected using a statistical method relying on standard deviations of molecular descriptors[27]. For activity outliers, molecules with reported ΔG values

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

greater than 2.5 standard deviations from the mean were removed. Though traditional practice advises classifies outliers as values more than only two standard deviations away, in this case, retaining data points remained a priority and a larger margin was allowed. There were 9, 21, and 11 α-, β-, and γ-CD outliers, respectively. After removal, around 200, 250, and 100 α-, β-, and γ-CD observations remained.

The data was then split into training, testing data and external validation. For each separate cyclodextrin, an external validation set was created from a random 15% subset of the data. To create multiple training and test sets, the remaining modeling data was split with representative resampling of ΔG values into ten different 75:25 train to test data partitions. Though not as advanced as maximum dissimilarity algorithms, this method proved more practical due to the large number of descriptors (over 1,000) generated for each guest molecule. Furthermore, maximum dissimilarity algorithms, when implemented in this instance, had the unfortunate tendency to select highly similar training and test sets, defeating the purpose of creating multiple sets in the first place.

II. Docking calculations

The process of docking is based on two processes: sampling and scoring[5]. Sampling refers to the capacity to search an active site on a protein, macromolecule or, in this case, affinity host. This can be performed with distance matrices, matching algorithms or incremental construction, multiple copy simultaneous searching, stochastic methods, or any combination of the aforementioned strategies. Scoring calculates the final binding affinity between the guest and host and can be dependent on force-field, empirical, or knowledge-based calculations. Docking generally involves the use of a host and a guest molecule which can be either rigid or flexible. Three types of conformation exist: rigid-rigid, rigid-flexible and flexible-flexible. In this paper we use AutoDock Vina, a version of AutoDock that uses Monte Carlo stochastic sampling coupled with a force field based scoring function from a resample of a drug like database to

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

derive its weighted parameters. Vina in particular uses a flexible drug guest and a rigid cyclodextrin host, although it allows side chain mobility when docking ligands onto proteins.

The PyRx Virtual Screening Tool provides a variety of services, including molecular energy minimization, docking calculation, and visualization of molecules. PyRx version 0.8 was used here [28]. To begin, all guest molecules went through energy minimization to determine the most likely atomic configurations. AutoDock Vina, integrated within PyRx, calculated the change in Gibbs free energy (kcal/mol). We tested the effect on the docking process of changes in the search space, search exhaustiveness, and scoring force field type.

III. Descriptor Generation

The open source software PaDEL-Descriptor calculated over 1000 descriptors for the remaining molecules, including fingerprints, structural details, and physical properties[29]. Additionally, PaDEL-Descriptor removed salts and minimized the energy of inputted files using an MM2 force field. To improve model interpretability, more abstract predictors, such as those related to eigenvalues for molecular matrices or autocorrelation, were excluded from calculation. The elimination of these descriptors did not produce any noticeable effect on final model accuracy and made feature selection less resource intensive.

IV. Feature selection

Recursive feature elimination (RFE), implemented with caret, was used to subset the predictors used for model-building (Kuhn 2018). Using this method, a random forest model is created using all available descriptors. Once trained, the relative importances of the predictors are calculated and differently sized subsets (defined by the user) of variables are selected to create and evaluate new models. The best combination of predictors is then returned by the model. RFE was performed on each of the ten train-test splits. The predictors determined to be useful for all folds were saved and used for tuning and training the models. This resulted in 13 variables for α-CD, 16 variables for β-CD, and 39 variables for γ-CD.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

V. QSAR development

We investigated the accuracy of several models that appeared in previous attempts at cyclodextrin QSARS (Table 1), including Cubist models, generalized linear models (GLM or GLMNet), random forests, partial least squares models, and support vector machines. Additionally, two QSAR methods not previously published for cyclodextrin – multivariate adaptive regression splines (MARS) and gradient-boosted models – were created and evaluated. Model building was accomplished with R-packages Cubist, glmnet, randomForest, pls, e1071, earth, and gbm, respectively [30–34].

Cross-validation was used to determine ideal tuning parameters for each QSAR. For faster QSARs – such as generalized linear models (GLM) and partial least squares (PLS) – tuning was performed using 10-fold cross validation. For more resource-intensive models or models with large parameter spaces – such as random forests, Cubist and support vector machines (SVM)– only 5 folds were used. Optimized models, QSARs built with the tuned parameters and trained on the entire training set, were used to predict the test for each combination of test and training set. Further fine tuning was also performed at this step. The model that produced the lowest root-mean square error (RMSE) and highest R^2 (or an otherwise most ideal combination) on all the test sets became the final model, i.e., the model saved for future use. Furthermore, the models were evaluated according to Tropsha and Golbraikh standards for QSARs[35]. Although R^2 and RMSE can be useful for generalizing predictive capacity, they may be misleading in certain cases, necessitating stricter additional standards of evaluation. As an additional test of reproducibility, the final models were used in ensemble to predict the values of the external validation set. Because this dataset was withheld from the entire model training process, the external validation set served to simulate model performance on new data.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

VI. Applicability domain

Applicability domain describes the range of molecules where the model can be expected to generate reliable predictions. A new molecule outside of the applicability domain is structurally quite different from the set of data the model was trained on, and thus a prediction will rely on extrapolation and may not be accurate. The applicability domain of the models was determined with the same method used to detect outliers when cleaning the dataset[27].

VII. Y-randomization

Y-randomization was used to further verify the significance of the results. Many advanced QSAR methods are powerful enough to model data off of noise, so y-randomization ensures that the modelling process produces results significantly more accurate than what could be obtained by chance. Randomization can be achieved by permutation (randomly changing the positions of observed values) or random number generation (replacing observed values with completely new data). Different combinations of permutation and/or random generation yields five different modes of y-randomization to investigate: 1) original ΔG values vs. randomly generated descriptors, 2) permuted ΔG vs. original descriptors, 3) random ΔG vs. original descriptors, 4) random ΔG vs. random descriptors, and 5) permuted ΔG vs. random descriptors. (Combinations including permutation of descriptors are not included because the large number of predictors in QSARs renders the effects of such a process virtually indistinguishable from random number generation.) However, because the y-randomization process is extremely resource-intensive (as each mode requires that several randomized iterations undergo the modeling process), only mode 1, the most common interpretation of y-randomization, was investigated[36]. The observed ΔG values were randomly assigned to guest molecules, and the entire model refitting process was re-done, from feature selection to external validation.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

VIII. Creating an App

Using R's "shiny" package, most of the process of running the QSAR could be implemented in a web app. The app was split into three main pages: Download, Upload, and Explore. "Download" accesses Chemical Identifier Resolver and obtain SDFs. The page also draws the obtained molecule using the package "ChemmineR", allowing the user to check that the SDF is accurate. "Upload" implements the QSARs after the user provides the app with a CSV of the descriptors from PaDEL-descriptor. After calculating the affinity and analyzing the applicability domain of the molecules, the user is provided with both a graph and a table of the results. The third page "Explore", stores the results of using the ensemble on FDA-approved drugs, as obtained from the annual publication "Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations." [40]

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

IX. Modeling release curves

Partial differential equations that model drug diffusion were solved using the R package deSolve[37] using the method of lines as previously done by Fu[2]. In the model, the release media was assumed to be water and the delivery system was assumed to be flat, thin circular cyclodextrin disc. The boundary condition between the polymer and the media was approached as described by Linge [41]. Diffusivities of drug molecules were calculated from molecular weight and viscosity using a modified Stokes-Einstein-Sutherland equation, as done by Vulic [38].

4. Results and discussion

I. Performance of docking

Fig 2. Results of PyRx docking.

When predicting on the entire cleaned dataset (all modeling data, which includes the training, testing, and external validation set), AutoDock Vina yielded an R^2 of 0.18 and a RMSE of 5.00 kJ/mol (Fig 2). Of the 547 cleaned complexes, docking provided calculations for 458, failing to provide data on 89 complexes. Adjusting settings in Vina, such as the minimization algorithm size of the steps in the calculation, did not yield significant differences in accuracy. In comparison, the affinity of only around 40 cleaned molecules could not be obtained by the ensemble QSAR. In these cases, the withheld molecules were determined to be outliers, and the actual QSAR model could still be used to predict a value.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

II. Performance of QSARs

Fig 3. Results of QSARs on test sets.

Table 2. Evaluation of QSARs on test sets.

QSAR	α-CD				β-CD				γ-CD			
	A	*B*	*C*	*D*	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*D*
Cubist	0.63	0.56	0.07	0.93	0.75	0.59	0.02	0.97	0.08*	-0.19*	0	0.98
GBM	0.78	0.5	0.05	0.95	0.83	0.73	0.02	0.98	0.35*	0.03*	0.68*	0.96
GLMNet	0.53*	0.54	0	0.94	0.52*	0.45	0.03	0.95	0.36*	-0.26*	0.17*	0.97
MARS	0.65	0.58	0.03	0.99	0.73	0.58	0	0.98	0.40*	-0.33*	0.33*	0.97
PLS	0.55*	0.47	0.05	0.93	0.55*	0.47	0.02	0.95	0.16*	-0.1*	0.28*	0.97
Polynomial SVM	0.65	0.55	0	0.98	0.74	0.56	0	0.98	0.45*	-0.28*	0.05	1
Random Forest	0.76	0.63	0.02	0.96	0.84	0.67	0.03	0.98	0.69	0.28*	0.24*	0.98
RBF SVM	0.74	0.64	0.02	0.96	0.85	0.61	0	0.98	0.35*	-0.19*	0.05	0.98
Sigmoid SVM	0.51*	0.52	0.05	0.93	0.50*	0.56	0.27	0.92	0.32*	-0.11*	0	0.98

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The columns labeled A-D indicate the four conditions outlined by Golbraikh and Tropsha. A: $R^2 > 0.6$; B: $q^2 > 0.5$, where q^2 is the result from leave one out cross-validation on the training set; C: $|R^2 – R'^2_0| / R^2 < 0.1$, indicating that the R^2 when the axes are flipped (R'^2_0) is close to the original R^2; D: $0.85 < k < 1.15$, where k, the slope of the regression line through the points is close to 1.

*Model failed condition

The results of predicting on the test data for each QSAR and cyclodextrin type are markedly higher than docking (with the exception of γ-CD), reaching an R^2 of around 0.5 to 0.7, as seen in Table 2 and Fig 3. The reported R^2 for each QSAR type is calculated from an average of the performance of the model on all test splits. Additionally, Table 2 contains information on verification of all the QSAR types (3-VII Validation methods). Of the types investigated, only PLS and GLMNet failed to pass the salvo of verification criteria, both falling short of attaining an R^2 of 0.6.

In terms of reliability, most models were able to handle the available data well, providing calculations for all provided molecules. Only the Random Forest and Cubist models failed to calculate the affinity of some molecules, possibly due to being based around decision-trees. The algorithm underlying both models attempts to draw predictions by categorizing entries based on their features. If they encounter a molecules entirely different from the data they trained on, the models may fail to create a prediction. Advantageously for our approach, the failure to calculate some values becomes less important where models are combined in an ensemble where the final prediction is averaged over many models.

Fig 4. QSAR ensemble prediction.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The results of ensemble prediction (averaging the results of many different QSARs) can be seen in Figure 4. While α- and β-CD models managed to reach moderately high predictive performance metrics, unfortunately, all γ-CD models lacked usable predictive power.

The models passing the verification in Table 2 were further verified using y-randomization. To ensure accuracy was not the result of the models building off of noise, 25 different permutations of ΔG values were created. All Q^2 values of the models created from the original data were calculated to lie well outside 3 standard deviations of the mean Q^2 of the randomized data. Additionally, the R^2 values of the ensemble QSARs were significantly greater than the R^2 values obtained from the ensemble models created from permuted data (means of 0.021 and 0.027 and standard deviations of 0.011 and 0.016 for α- and β-CD, respectively).

III. Variable importance

Interpretability of a model provides a rough check if a model is calculating off of random noise or if the model is drawing logical calculations from physical properties to molecular behavior. Each model, due to differences in statistical algorithms and approaches, has differing levels of interpretability. GLM, being similar to linear models, have easily accessible coefficients associated with each predictor, so the relative impact of each factor can be compared with reasonable confidence. Cubist models, on the other hand, tend to be difficult to interpret as variables are processed through multiple levels of decision trees.

Evaluation for the relative importance of variables are shown in Figure 5. Random forest was the only QSAR type with a pre-packaged importance function for variable analysis. PLS variables were analyzed using a function obtainable from Bjørn-Helge Mevik[39]. Max Kuhn's caret package was used to evaluate GLMNet, the two SVM kernels, and Cubist. Unfortunately, caret was unable to process the final models for GLMNet and SVM, and the reported variable importance values were actually derived from models created within caret's "train" function, and are thus slightly different from the models saved in the ensemble. To determine importance, the

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

"train" function removes a variable, rebuilds the model, and analyzes the effect on accuracy. The more important a variable, the larger the drop in accuracy. After each variable has been tested, the function can then rank the importance of the descriptors.

Fig 5. Variable importance.

For β-CD, XLogP, a measure of lipophilicity, appears to be important for all models, consistent with how the structure of cyclodextrin allows for easier complexation with small hydrophobic drugs. The same reasoning can be extended to LipoAffinityIndex and MLogP, additional approaches to quantifying lipophilicity. The number of carbons, nC, is also consistently important, possibly due to a relationship with molecule size. WTPT-2 is the PaDEL weighted path descriptor divided by the number of atoms, and also may be important due to encoding information on molecular size.

However, not all the variables can be linked to set chemical properties. SpMax and SpMin relates to eigenvalues of a modified connectivity matrix, a numerical representation of atomic and molecular bonds, and may not be associated with any interpretable physical property (the same analysis can also be used for GATS predictors). To aid interpretability, building a model with predictors easily attributed to physical or chemical properties may be advised. The extent to which interpretability should trade off with accuracy remains in question. Our findings go in accordance to those in the general literature from table 1 were lipophilicity tends to highly influence model output.

IV. Web Application and FDA Database

Fig 6. User interface of the Shiny App

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

After collecting a list of FDA-approved drugs and drug combinations from the Orange Book, an annual publication listing all approved pharmaceuticals [cite], the names were cleaned for individual active compounds. In total, 1401 unique molecules could be extracted. Of these, 1116 could be downloaded from Cactus and 1031 could be processed by PaDEL. Many of the molecules that could not be analyzed by PaDEL would have proven impractical for cyclodextrin delivery, such as simple ionic salts (e.g., potassium chloride), or large molecules made of more than 100 atoms. Running the remaining guests through applicability domain analysis yielded 638 molecules, 45.5% of the original set. While less than half of FDA-approved drugs could pass through the model, the 600 available guests spans a wide range of properties and uses, allowing the page to be useful for candidate selection.

Though the app could be uploaded online through shinyapps.io, server time limitations on the account hosting the app make it impractical for usage by a large number of individuals simultaneously. In order to run the app for more than a few hours, such as with screening a large dataset of molecules, the code would have to be downloaded through GitHub. In addition, the user would need to download the R libraries and the IDE RStudio, potentially negating the goal of creating an accessible, intuitive interface. The “Explore” page partially alleviates this obstacle, as it allows the user to perform a quick search of a pre-predicted affinity rather than spend time downloading the structure file, launching PaDEL, and running the QSAR.

V. Drug release module

Fig 7. Drug release curves

To demonstrate the extensibility of the shiny app and its value in the design of drug delivery strategies the results of the QSAR predictions can be fed into a mechanistic model of drug delivery. The results demonstrate the ranges of values expected from the strongest affinity binding predictions and from the weakest. As expected, strong predictions produce much slower

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

release profiles and weak predictions produce faster release profiles. Conservation of mass was verified as the sum of all mass within the system from the polymer and media compartment across all times as a test of the implementation. Notably, the implementation in R required a modification of the method of lines for appropriate modeling of the polymer to liquid media interface.[41] Future efforts in creating new modules could explore substructure searching to identify alternative strategies for weak binders or drugs that demonstrate unsuitable release profiles. It is expected that not all drugs will follow this simplistic model of drug release. For those cases our lab has built a whole suite of approaches to alter elution rates such as a wide range of formulations, supramolecular interactions, Schiff-base formation and multi-arm PEG substitutions.

5. Conclusions

In predicting the binding affinity of cyclodextrin with small drug molecules, QSARs such as Cubist, GBM, MARS, random forest, and SVM models can be created using accessible open-source software. These models outperform available molecular docking software in both accuracy and time consumption and pass statistical verification of reliability. The additional accuracy afforded by QSARs can be integrated into the previously published mechanistic model for predicting drug release curves for candidate molecules. This would both help narrow down candidates for affinity-based drug delivery using cyclodextrins, as well as help advise drug loading configurations to tailor release rate from a delivery system for a given biomedical application. Additionally, similar strategies could model other molecular interactions and allow application of computational approaches for higher-throughput research strategies in drug delivery, an inherently empirical field. Approaches like the one presented here will move drug delivery closer to a truly integrated approach in drug discovery.

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements: The authors would like to acknowledge the Hathaway Brown School and Dr. Crystal Miller for enabling Alison Xin's contributions through the Science Research & Engineering Program.

6. References

1. Rivera-Delgado E, Ward E, von Recum HA. Providing sustained transgene induction through affinity-based drug delivery. J Biomed Mater Res. 2016;104: 1135–1142. doi:10.1002/jbm.a.35643

2. Fu AS, Thatiparti TR, Saidel GM, von Recum HA. Experimental Studies and Modeling of Drug Release from a Tunable Affinity-Based Drug Delivery Platform. Ann Biomed Eng. 2011;39: 2466–2475. doi:10.1007/s10439-011-0336-z

3. Wang NX, von Recum HA. Affinity-Based Drug Delivery. Macromol Biosci. 2011;11: 321–332. doi:10.1002/mabi.201000206

4. Ortiz M, Fragoso A, O'Sullivan CK. Amperometric detection of antibodies in serum: performance of self-assembled cyclodextrin/cellulose polymer interfaces as antigen carriers. Org Biomol Chem. 2011;9: 4770–4773. doi:10.1039/C1OB05473B

5. Jacob RB, Andersen T, McDougal OM. Accessible High-Throughput Virtual Screening Molecular Docking Software for Students and Educators. PLOS Computational Biology. 2012;8: e1002499. doi:10.1371/journal.pcbi.1002499

6. Dehmer M, Varmuza K, Bonchev D. Statistical Modelling of Molecular Descriptors in QSAR/QSPR. John Wiley & Sons; 2012.

7. Merzlikine A, Abramov YA, Kowsz SJ, Thomas VH, Mano T. Development of machine learning models of β-cyclodextrin and sulfobutylether-β-cyclodextrin complexation free energies. International Journal of Pharmaceutics. 2011;418: 207–216. doi:10.1016/j.ijpharm.2011.03.065

8. Ahmadi P, Ghasemi JB. 3D-QSAR and docking studies of the stability constants of different guest molecules with beta-cyclodextrin. J Incl Phenom Macrocycl Chem. 2014;79: 401–413. doi:10.1007/s10847-013-0363-5

9. Ghasemi JB, Salahinejad M, Rofouei MK. An alignment independent 3D-QSAR study for predicting the stability constants of structurally diverse compounds with β-cyclodextrin. J Incl Phenom Macrocycl Chem. 2011;71: 195–206. doi:10.1007/s10847-011-9927-4

10. Veselinović AM, Veselinović JB, Toropov AA, Toropova AP, Nikolić GM. In silico prediction of the β-cyclodextrin complexation based on Monte Carlo method. Int J Pharm. 2015;495: 404–409. doi:10.1016/j.ijpharm.2015.08.078

11. Pérez-Garrido A, Helguera AM, Guillén AA, Cordeiro MNDS, Escudero AG. Convenient QSAR model for predicting the complexation of structurally diverse compounds with β-

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cyclodextrins. Bioorganic & Medicinal Chemistry. 2009;17: 896–904. doi:10.1016/j.bmc.2008.11.040

12. Xu Q, Wei C, Liu R, Gu S, Xu J. Quantitative structure–property relationship study of β-cyclodextrin complexation free energies of organic compounds. Chemometrics and Intelligent Laboratory Systems. 2015;Complete: 313–321. doi:10.1016/j.chemolab.2015.06.001

13. Mirrahimi F, Salahinejad M, Ghasemi JB. QSPR approaches to elucidate the stability constants between β-cyclodextrin and some organic compounds: Docking based 3D conformer. Journal of Molecular Liquids. 2016;219: 1036–1043. doi:10.1016/j.molliq.2016.04.037

14. Prakasvudhisarn C, Wolschann P, Lawtrakul L. Predicting Complexation Thermodynamic Parameters of β-Cyclodextrin with Chiral Guests by Using Swarm Intelligence and Support Vector Machines. Int J Mol Sci. 2009;10: 2107–2121. doi:10.3390/ijms10052107

15. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inf. 2010;29: 476–488. doi:10.1002/minf.201000061

16. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry. 2010;31: 455–461. doi:10.1002/jcc.21334

17. Suzuki T. A Nonlinear Group Contribution Method for Predicting the Free Energies of Inclusion Complexation of Organic Molecules with α- and β-Cyclodextrins. J Chem Inf Comput Sci. 2001;41: 1266–1273. doi:10.1021/ci010295f

18. Katritzky AR, Fara DC, Yang H, Karelson M, Suzuki T, Solov'ev VP, et al. Quantitative Structure−Property Relationship Modeling of β-Cyclodextrin Complexation Free Energies. J Chem Inf Comput Sci. 2004;44: 529–541. doi:10.1021/ci034190j

19. Rekharsky MV, Inoue Y. Complexation Thermodynamics of Cyclodextrins. Chem Rev. 1998;98: 1875–1918. doi:10.1021/cr970015o

20. Connors KA. Population Characteristics of Cyclodextrin Complex Stabilities in Aqueous Solution. Journal of Pharmaceutical Sciences. 1995;84: 843–848. doi:10.1002/jps.2600840712

21. Wickham H. Tidyverse: Easily install and load'tidyverse'packages. R package version. 2017;1.

22. Dowle M, Srinivasan A, Gorecki J, Short T, Lianoglou S, Antonyan E. data.table: Extension of "data.frame" [Internet]. 2017. Available: https://cran.r-project.org/web/packages/data.table/index.html

23. Duncan Temple Lang and the CRAN Team. XML: Tools for Parsing and Generating XML Within R and S-Plus [Internet]. 2018. Available: https://cran.r-project.org/web/packages/XML/index.html

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

24. Duncan Temple Lang and the CRAN team. RCurl: General Network (HTTP/FTP/...) Client Interface for R [Internet]. 2018. Available: https://cran.r-project.org/web/packages/RCurl/index.html

25. Bates D, Maechler M, Davis TA, AMD C, Oehlschlägel J, Riedy J, et al. Matrix: Sparse and Dense Matrix Classes and Methods [Internet]. 2017. Available: https://cran.r-project.org/web/packages/Matrix/index.html

26. Kuhn M, Quinlan R. caret: Classification and Regression Training [Internet]. 2018. Available: https://CRAN.R-project.org/package=caret

27. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems. 2015;145: 22–29. doi:10.1016/j.chemolab.2015.04.013

28. Dallakyan S, J Olson A. Small-Molecule Library Screening by Docking with PyRx. Methods in molecular biology (Clifton, NJ). 2015;1263: 243–250. doi:10.1007/978-1-4939-2269-7_19

29. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32: 1466–1474. doi:10.1002/jcc.21707

30. Kuhn M, Steve W, Chris K, Nathan C, code) RQ (Author of imported C, code) RRPL (Copyright holder of imported C. Cubist: Rule- And Instance-Based Regression Modeling [Internet]. 2017. Available: https://cran.r-project.org/web/packages/Cubist/index.html

31. Friedman J, Hastie T, Simon N, Qian J, Tibshirani R. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models [Internet]. 2017. Available: https://cran.r-project.org/web/packages/glmnet/index.html

32. Cutler F original by LB and A, Wiener R port by AL and M. randomForest: Breiman and Cutler's Random Forests for Classification and Regression [Internet]. 2015. Available: https://cran.r-project.org/web/packages/randomForest/index.html

33. Mevik B-H, Liland RW and KH. pls: Partial Least Squares and Principal Component Regression [Internet]. 2016. Available: https://cran.r-project.org/web/packages/pls/index.html

34. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, C++-code) C-CC (libsvm, et al. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [Internet]. 2017. Available: https://cran.r-project.org/web/packages/e1071/index.html

35. Golbraikh A, Tropsha A. Beware of q2! Journal of Molecular Graphics and Modelling. 2002;20: 269–276. doi:10.1016/S1093-3263(01)00123-1

36. Rücker C, Rücker G, Meringer M. y-Randomization and Its Variants in QSPR/QSAR. J Chem Inf Model. 2007;47: 2345–2357. doi:10.1021/ci700157b

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

37. Soetaert K, Petzoldt T, Setzer RW. deSolve: Solvers for Initial Value Problems of Differential Equations (“ODE”, “DAE”, ’DDE’) [Internet]. 2018. Available: https://CRAN.R-project.org/package=deSolve

38. Vulic K, Pakulska MM, Sonthalia R, Ramachandran A, Shoichet MS. Mathematical model accurately predicts protein release from an affinity-based delivery system. Journal of Controlled Release. 2015;197: 69–77. doi:10.1016/j.jconrel.2014.10.032

39. Bjørn-Helge Mevik. VIP.R: Implementation of VIP (variable importance in projection) (*) for the “pls” package [Internet]. Bjørn-Helge Mevik; 2007. Available: http://mevik.net/work/software/VIP.R

40. [Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations]

[Internet]. [cited 31 May 2019]. Available: https://www.fda.gov/media/71474/download

41. Linge S, Langtangen HP. Programming for Computations - Python [Internet]. Texts in Computational Science and Engineering. 2016. doi:10.1007/978-3-319-32428-9

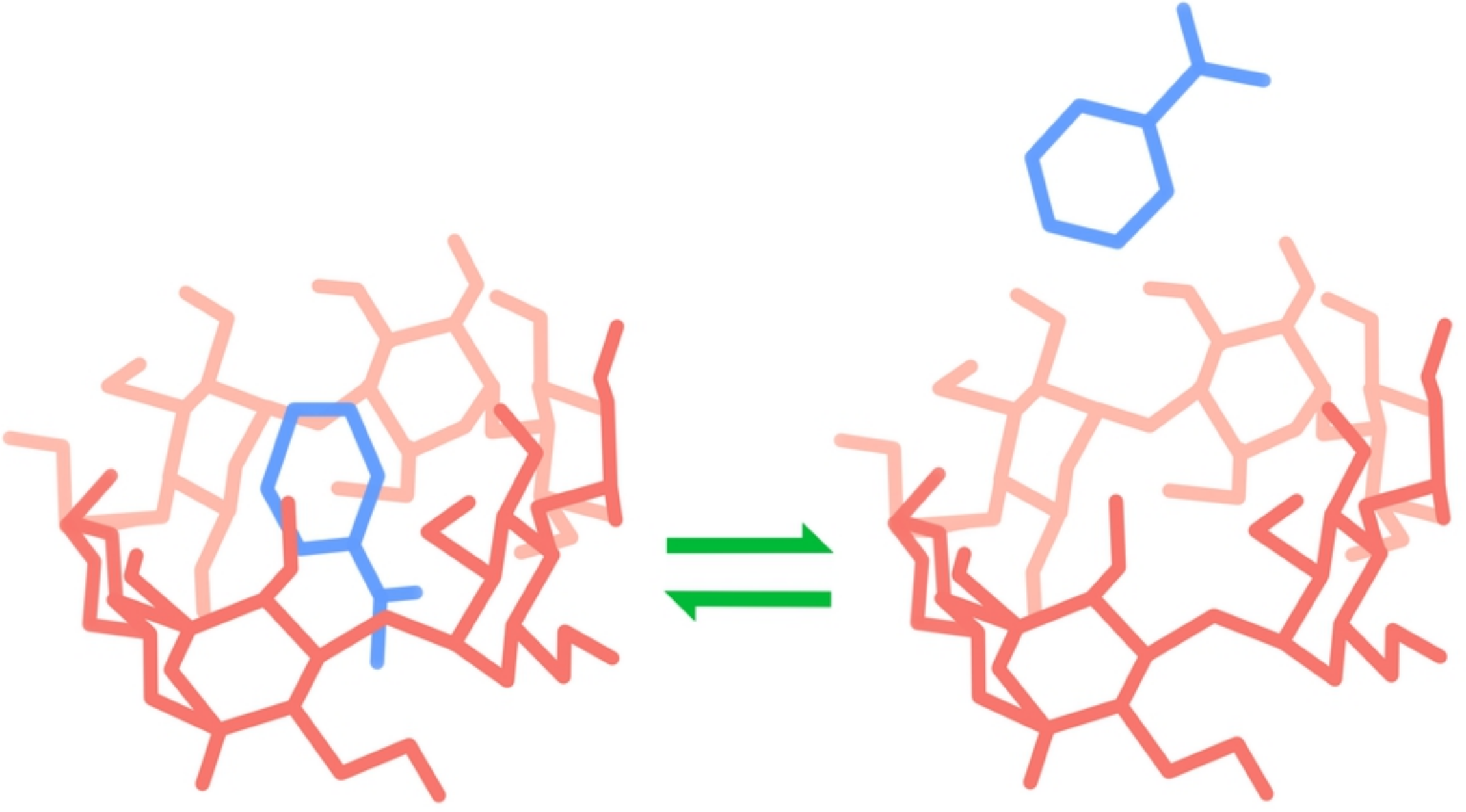

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

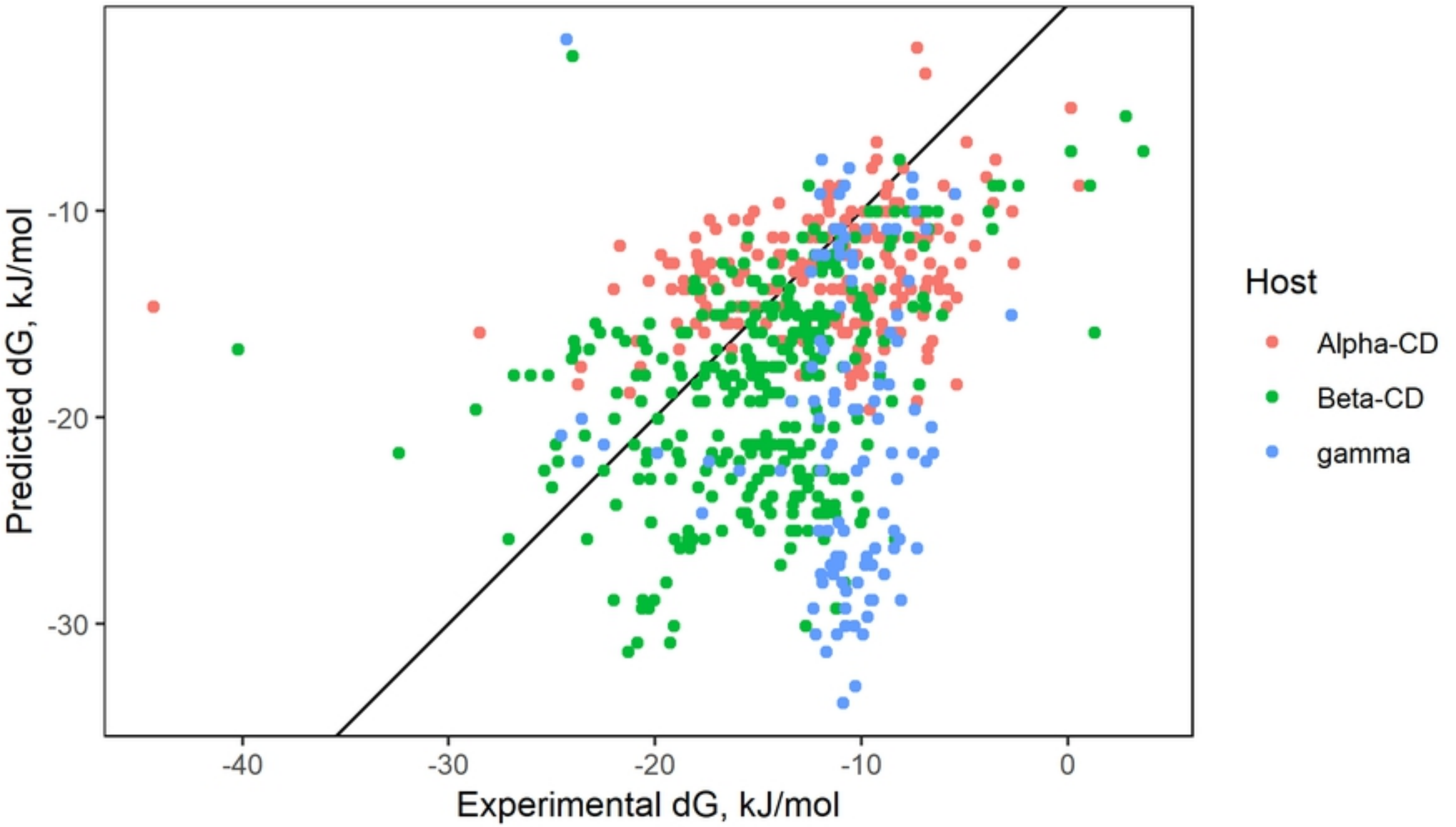

Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

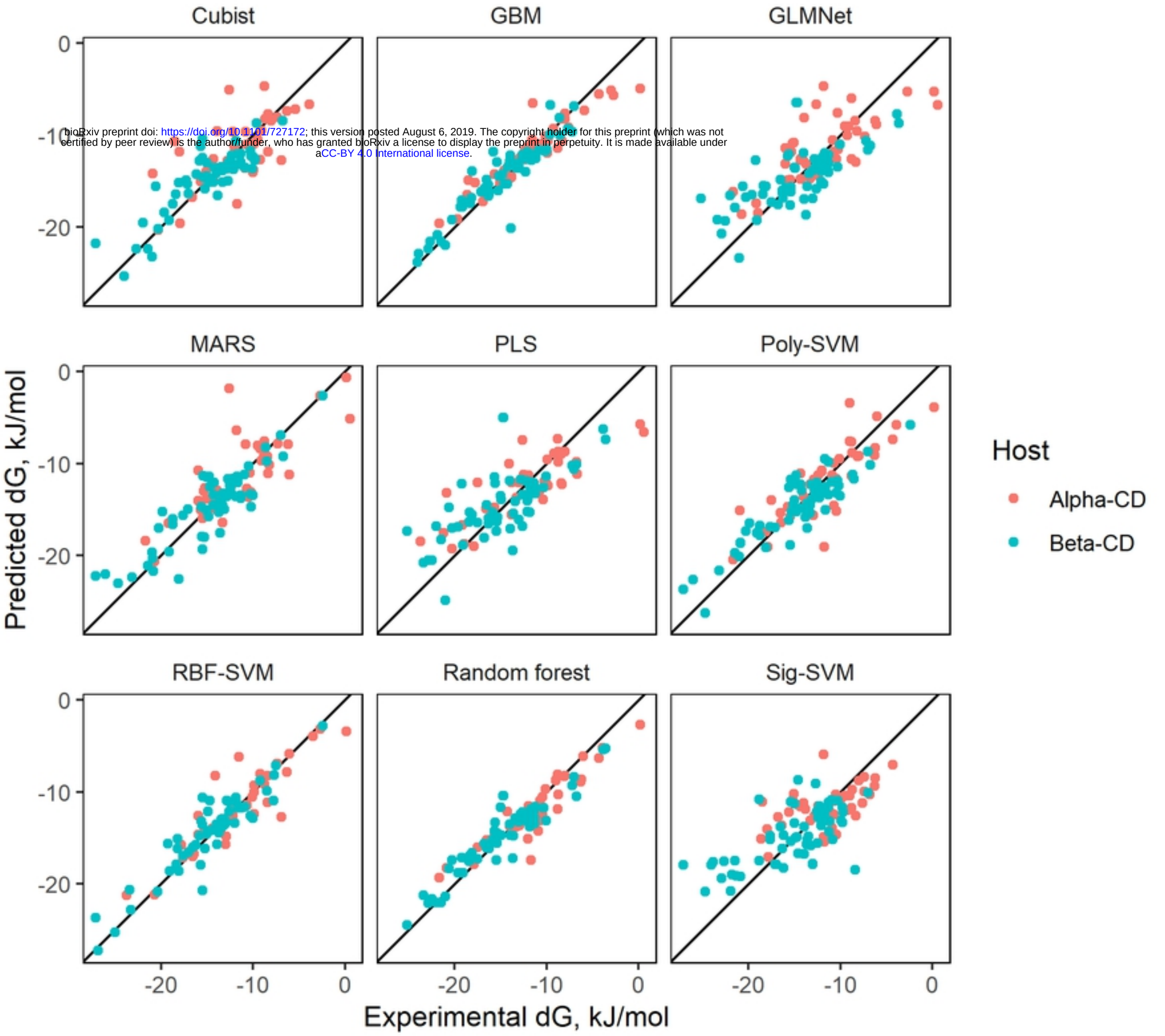

Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/727172; this version posted August 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

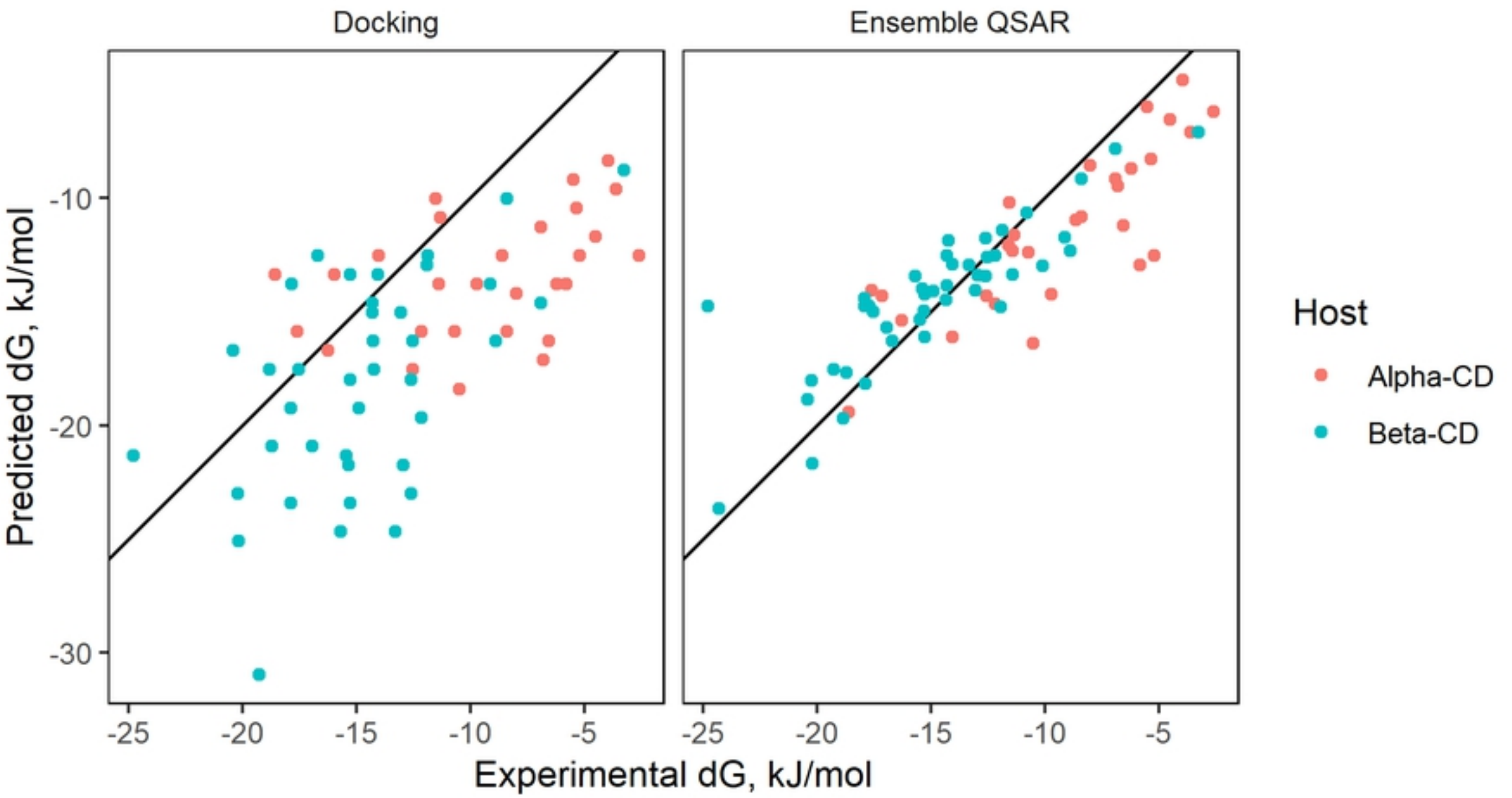

Figure 4

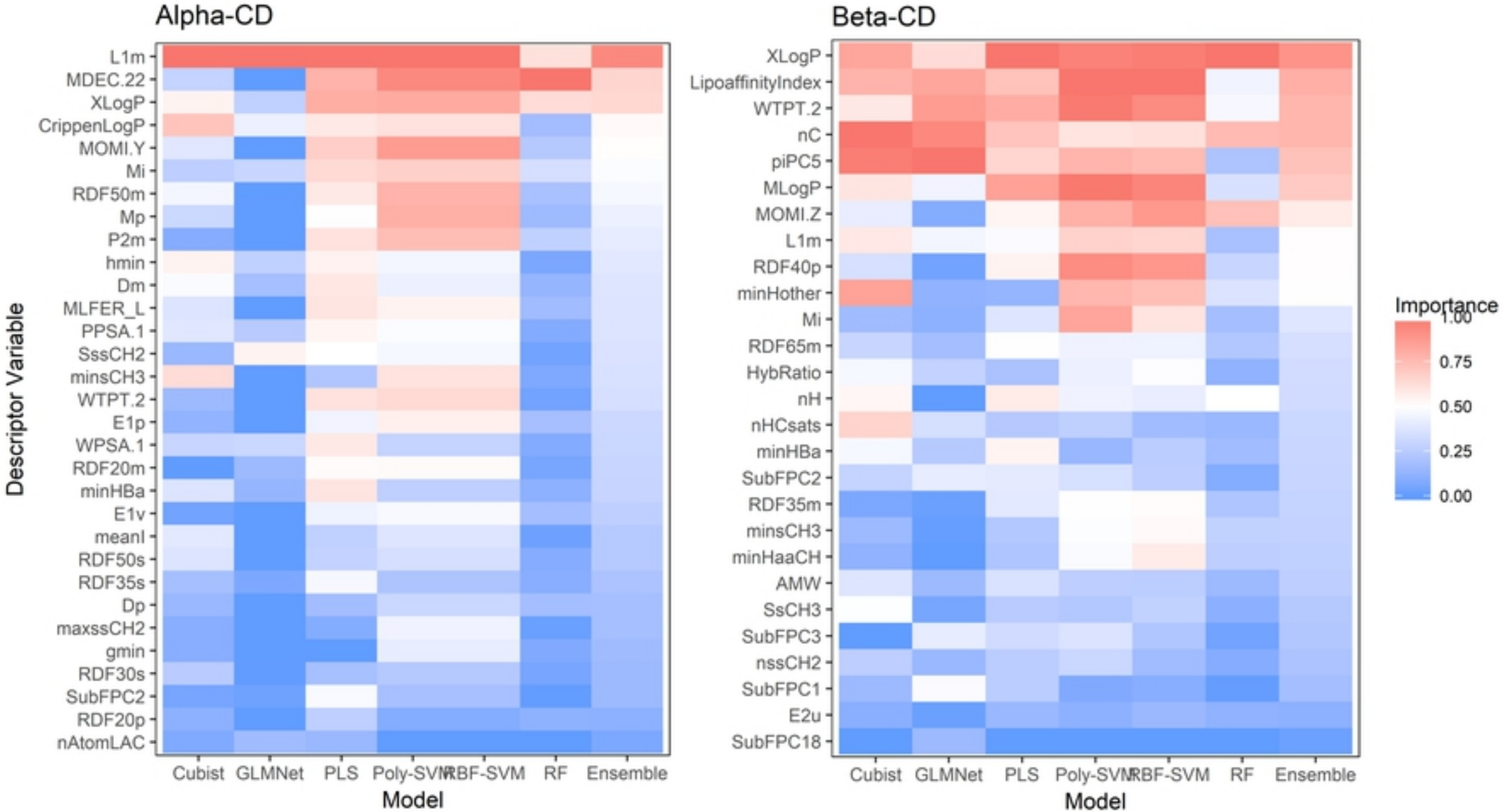
Alpha-CD
Beta-CD
Descriptor Variable
L1m
MDEC.22
XLogP
CrippenLogP
MOMI.Y
Mi
RDF50m
Mp
P2m
hmin
Dm
MLFER_L
PPSA.1
SssCH2
minsCH3
WTPT.2
E1p
WPSA.1
RDF20m
minHBa
E1v
meanI
RDF50s
RDF35s
Dp
maxssCH2
gmin
RDF30s
SubFPC2
RDF20p
nAtomLAC
XLogP
LipoaffinityIndex
WTPT.2
nC
piPC5
MLogP
MOMI.Z
L1m
RDF40p
minHother
Mi
RDF65m
HybRatio
nH
nHCsats
minHBa
SubFPC2
RDF35m
minsCH3
minHaaCH
AMW
SsCH3
SubFPC3
nssCH2
SubFPC1
E2u
SubFPC18
Cubist
GLMNet
PLS
Poly-SVM
RBF-SVM
RF
Ensemble
Model
Importance
1.00
0.75
0.50
0.25
0.00

Figure 5

Cyclodextrin Affinity Prediction
INSTRUCTIONS
DOWNLOAD
UPLOAD
EXPLORE
RELEASE
Search for specific molecules in the data table or search by category
CYCLODEXTRIN TYPE
Alpha-CD
Beta-CD
RANGE OF PREDICTIONS, KJ/MOL
-50
-30
10
50
APPLICABILITY DOMAIN, NEWSK
0
3
6
TABLE
PLOT
Predicted dG, kJ/mol
-10
-20
1
2
3
Applicability, newSk
Guest: Lacosamide
Affinity: -14.2kJ/mol
Host: Beta
Applicability: Inside
Domain
inside
Host
Alpha-CD
Beta-CD

Figure 6

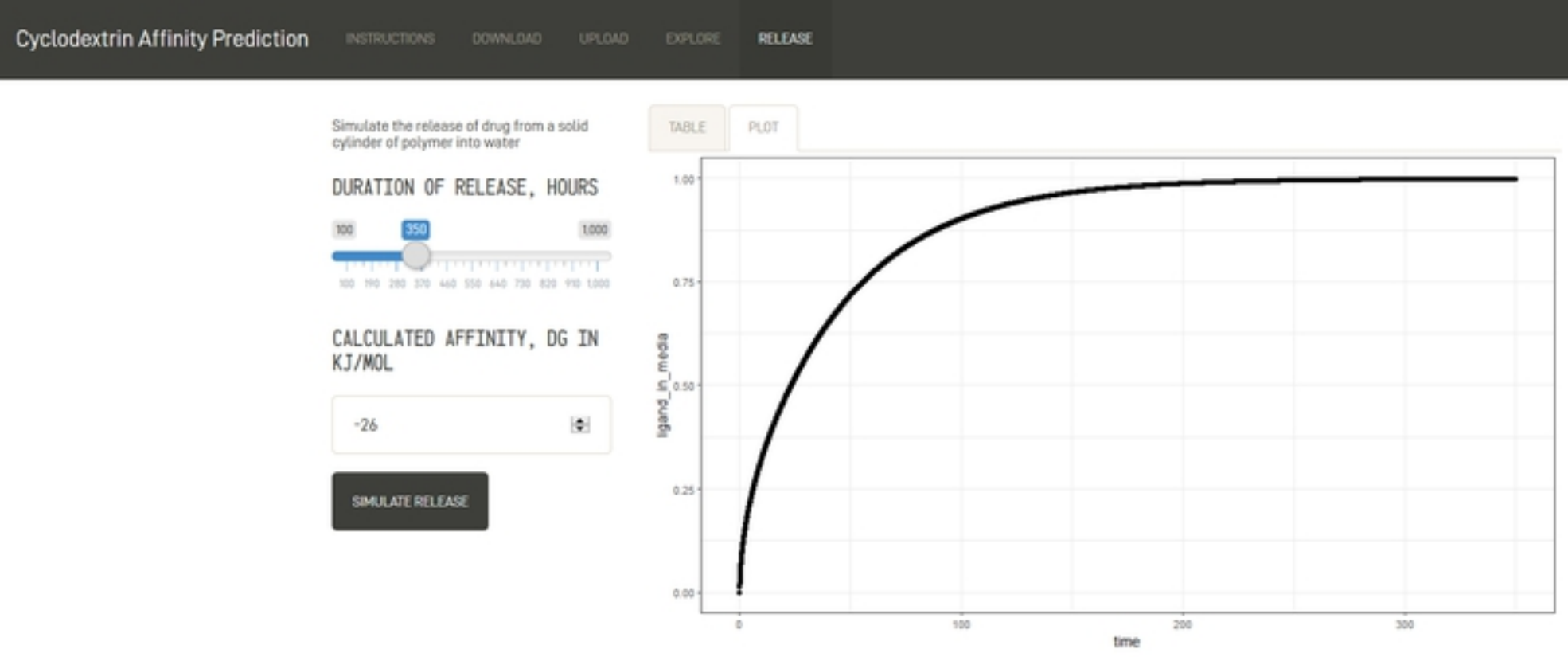
Cyclodextrin Affinity Prediction
INSTRUCTIONS
DOWNLOAD
UPLOAD
EXPLORE
RELEASE
Simulate the release of drug from a solid cylinder of polymer into water
DURATION OF RELEASE, HOURS
100
350
1,000
CALCULATED AFFINITY, DG IN KJ/MOL
-26
SIMULATE RELEASE
TABLE
PLOT
ligand_in_media
1.00
0.75
0.50
0.25
0.00
0
100
200
300
time

Figure 7