Gaussian Processes Regression for Cyclodextrin Host-Guest Binding Prediction

Ruan M Carvalho
UFJF: Universidade Federal de Juiz de Fora
Iago G. L. Rosa
UFJF: Universidade Federal de Juiz de Fora
Diego E. B. Gomes
UFJF: Universidade Federal de Juiz de Fora
Priscila V. Z. C. Goliatt
UFJF: Universidade Federal de Juiz de Fora
Leonardo Goliatt (goliatt@gmail.com)
UFJF: Universidade Federal de Juiz de Fora https://orcid.org/0000-0002-2844-9470

Short Report

Keywords: Molecular Interaction, Machine Learning, Cyclodextrin

Posted Date: July 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-372834/v1

Version of Record: A version of this preprint was published at Journal of Inclusion Phenomena and Macrocyclic Chemistry on July 12th, 2021. See the published version at https://doi.org/10.1007/s10847-021-01092-4.

Noname manuscript No.
(will be inserted by the editor)

Gaussian Processes Regression for Cyclodextrin Host-Guest Binding Prediction

Ruan M. Carvalho · Iago G. L. Rosa · Diego E. B. Gomes · Priscila V. Z. C. Goliatt · Leonardo Goliatt*

Received: date / Accepted: date

Abstract Machine Learning (ML) techniques are becoming an integral part of rational drug design and discovery. Data-driven modeling regularly outperforms physics-based models for predicting molecular binding affinities, placing ML as a promising tool. Cyclodextrins are nano-cages used to improve the delivery of insoluble or toxic drugs. Due to chemical similarity to proteins, ML approaches could vastly profit to improve affinity prediction and enhance their carriable drug portfolio. Here we evaluate the performance of three well-known ML methods - Support Vector Regression (SVR), Gaussian Process Regression (GPR), and eXtreme Gradient Boosting (XGB) - to predict the binding affinity of cyclodextrin and known ligands. We perform hyperparameter tuning through Random Search. The results were compatible with the presented literature. We were able to increase our previous prediction performance and present a GPR model to adjust to the data ($R^2 = 0.803$) with low prediction errors (RMSE = 1.811 kJ/mol and MAE = 1.201 kJ/mol).

Keywords Molecular Interaction · Machine Learning · Cyclodextrin

Ruan M. Carvalho
Computational Modeling, Federal University of Juiz de Fora, Brazil
E-mail: ruan.medina@engenharia.ufjf.br

Iago G. L. Rosa
Computational Modeling, Federal University of Juiz de Fora, Brazil
E-mail: iago.rosa@engenharia.ufjf.br

Diego E. B. Gomes
Computational Modeling, Federal University of Juiz de Fora, Brazil
E-mail: diego.gomes@ufjf.edu.br

Priscila V. Z. C. Goliatt
Computational Modeling, Federal University of Juiz de Fora, Brazil
E-mail: priscila.capriles@ice.ufjf.br

*L. Goliatt (corresponding author)
Computational Modeling, Federal University of Juiz de Fora, Brazil
E-mail: leonardo.goliatt@ufjf.br

1 Introduction

Cyclodextrins (CD) are widely used as drug nano-carriers, mainly for enhancing solubility, stability, and bioavailability of a variety of drugs [17,22, 21]. They are extensively applied in the pharmaceutical field, have relatively controllable variables, and specific evaluation indexes [22,12,29]. Drug-CD complexation and electrostatic interactions can significantly alter the drug's pharmacokinetic profile [30]. The CD-drug complexes work as a type of molecular encapsulation [7], CD wraps the drug and partly shields it from the external environment, preventing drug degradation [25](Figure 1). Also, due to their simplicity and chemical similarity to proteins, CD complexes serve to develop predictive models. [33].

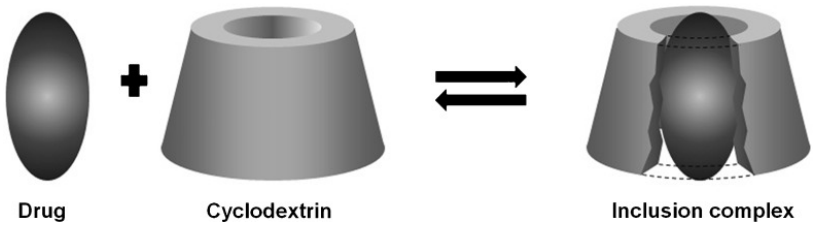


Fig. 1 The interaction of a drug and a cyclodextrin to form an inclusion complex [25].

To enable using CD nano-carrier system for a novel class of drugs, one should perform a screening and/or biochemical characterization of potential complexes, which can be a quite laborious and expensive process [28]. Computational methods aim to automatize and reduce the time and cost to discover new drugs [18,15], however, traditional scoring functions (SF), empirical or physics-based, still struggle to rank potential binders, unless rigorous free energy calculations could be performed, this limitation has been addressed in several community challenges [24], highlighting the potential of Machine Learning Scoring Functions (MLSF) to improve predictions. Therefore, it is crucial to develop models for pre-screening to direct rigorous tests.

MLSF consistently outperforms classical scoring functions to predict binding properties, leading to more accurate hit rates and potencies, placing ML techniques as a promising tool [33]. Zhao *et al.* [33] presented a complete study when applying different ML methods (LightGBM, Random forest, and Deep learning) for predicting interaction energy between a vast number of cyclodextrin classes and ligand molecules.

Solov'ev *et al.* [27] presented a study over 3D molecular fragment descriptors for structure-property modeling. The authors apply a multiple linear regression model to predict the complexation free energy between antipodal guests and β-cyclodextrins. Xu *et al.* [31] presented a quantitative structure-property relationship study of β-cyclodextrin complexation free

energies of organic compounds applying multiple linear regression and artificial neural network.

Di *et al.* [6] perform an *in silico* prediction of binding capacity and interaction forces of organic compounds with α- and β-cyclodextrins applying a consensus method coupling five ML methods (Radial Basis Function Regression, Gaussian Processes, Random Forest, Gradient Boosted Trees, and Tree Ensemble). Kerner *et al.* [16] propose a computational strategy to predict drug interactions to unassociated biomedical implants and present tests over cyclodextrin host-guest systems. The authors apply Artificial Neural Networks coupled with an autoencoder structure to reduce the large dimensionality input data.

In this paper, we assess the performance of three well-known ML methods to predict binding affinity between cyclodextrin and ligand molecules in host-guest systems: (i) Support Vector Regression (SVR), (ii) Gaussian Process Regression (GPR), and (iii) eXtreme Gradient Boosting (XGB). We perform hyperparameters tuning through Random Search (RS). We focused on building a generalized model capable of pre-screening systems of different classes of cyclodextrin and a variety of ligand molecules.

2 Material and Methods

2.1 Data Collection

Data was curated and made available by the `BindingDB` community (`https://www.bindingdb.org`). Each record covers a host (large) and a guest (small) molecule. The database provided the molecule structural information through *SMILES*, and experimental conditions, including pH and temperature (°C), and the binding free energy, measured as ΔG (kJ/mol). Here we focused on α, β, and γ cyclodextrin (-CD) [11] classes, that differs in the number of glucose units: 6, 7, and 8, respectively. For sake of consistency, we only considered the experiments with available pH and temperature measurements within the following range: $6.9 \leq \text{pH} \leq 7.4$ and $14.5 \leq \text{Temp} \leq 30.1$; resulting in 280 unique observations of α-CD (73), β-CD (164), and γ-CD (43) experiments.

Physical-chemical molecular properties were calculated using the module RDKit descriptor calculation [20] from `KNIME` (`https://www.knime.com`). Table 1 shows the calculated descriptors for the three CD hosts. The guest's representation follows with the same descriptors listed plus the formal charge (FC). Figure 2 shows some of the distributions for the guest molecules descriptors, plus the distribution for the complexation energy values (ΔG).

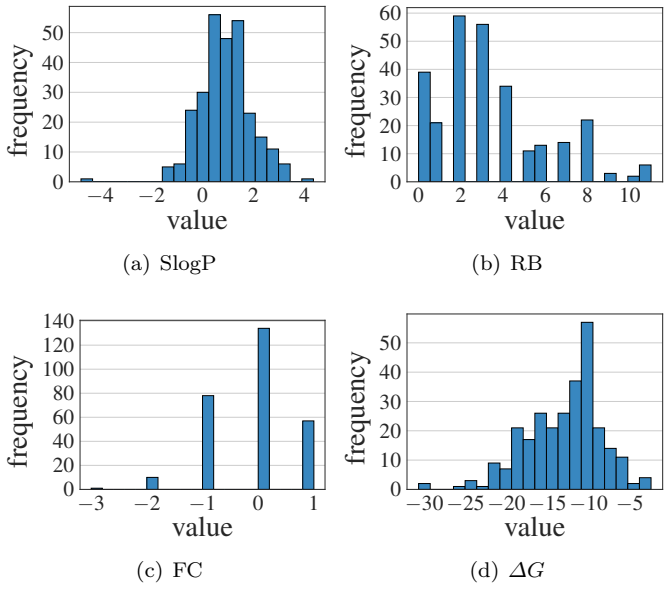
2.2 Machine Learning Approach

The database comprises 280 unique observations, 145 guests, 3 hosts, and 25 descriptors (9 for hosts, 10 for guests, 2 for the environment, 3 identifiers,

Table 1 Descriptors for each Cyclodextrin.

	α-CD	β-CD	γ-CD
SlogP	-13.06	-15.23	-17.41
SMR[a]	195.80	228.43	261.07
ASA[b]	372.13	433.92	495.72
TPSA[c]	474.90	554.05	633.20
AMW[d]	972.84	1134.99	1297.13
HBA[e]	30.0	35.0	40.0
HBD[f]	18	21	24
RB[d]	6	7	8
Atoms	126	147	168

[a]SMR: Molecular Refractivity. [b]ASA: Approximate Surface Area. [c]TPSA: Topological Polar Surface Area. [d]AMW: Average Molecular Weight. [e]HBA: HB Acceptor. [f]HBD: HB Donor. [g]RB: number of Rotatable Bonds.



(a) SlogP (b) RB

(c) FC (d) ΔG

Fig. 2 Distributions for guest descriptors in (a), (b), and (c), and ΔG in (d)

and 1 objective variable). The dataset was randomly divided into a unique set of training (224) and testing (56) using the Stratified K-fold method to maintain the same proportion of instances of each CD class in both sets [8] coupled with a KullbackLeibler divergence analysis between the sets. The data and code are available from the authors upon request. Additionally, Table 7 shows more descriptive information over the considered input and output variables and Table 8 and Table 9 show statistics of the training and testing sets, respectively.

In the present paper, we compare the performance of three ML methods to predict the binding affinity of numerous ligands to cyclodextrins: (i) ε-Support

Vector Regression, (ii) Gaussian Process Regressor, and (iii) eXtreme Gradient Boosting.

2.2.1 Gaussian Process Regressor (GPR)

Gaussian Process Regressor (GPR) is a non-parametric method that, instead of inferring a distribution over parametric function parameters, infers a distribution over functions directly. As described for Xiaoling Ou *et al.* [23], for all regression approaches, a Gaussian process model can be defined as:

$$y = f(X) + \epsilon$$

where $f(.)$ is a non-linear function mapping the input vector $\mathbf{X}$ to a scalar output y. ϵ is Gaussian noise with zero mean (i.e. $\epsilon \sim G(O, \sigma_v^2)$). The prior for $f(.)$ is assumed to be a Gaussian process, $f(.) \sim G(\mu, \mathbf{C})$, where μ is usually set to 0 and $\mathbf{C}$ is the covariance matrix.

The priors covariance is specified by passing a kernel object. Here, we applied both `RBF` kernel (as described above), and `Matern` kernel, given by

$$K\left(\mathbf{x}, \mathbf{x}_i\right) = \frac{1}{\Gamma_f(\nu) 2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{\sigma} \left\| \mathbf{x} - \mathbf{x}_i \right\|^2 \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\sigma} \left\| \mathbf{x} - \mathbf{x}_i \right\|^2 \right)$$

where σ is the length scale, ν controls the smoothness of the resulting function, $\Gamma_f(.)$ is the gamma function, and K_ν is a modified Bessel function [19].

2.2.2 ε-Support Vector Regression (ε-SVR)

The ε-Support Vector Regression (ε-SVR) is a version of the classical Support Vector Machines [3] applied in several fields and a potential tool for drug discovery studies [14]. Previous studies performed for our research group indicate good results with this method, although there is room for improvement. The ε-SVR is a linear regression model, as described by

$$f(\mathbf{x}) = \sum_{j=1}^{N} w_j K\left(\mathbf{x}, \mathbf{x}_j\right) + b$$

where $K(\cdot, \cdot)$ is a kernel function or a nonlinear transformation, $\mathbf{w} = [w_1, \cdots, w_N]$, is the vector of weights, b is a bias and N is the number of samples. In this paper, we use the radial basis kernel function (`RBF`) to the nonlinear transformation, given by

$$K\left(\mathbf{x}, \mathbf{x}_i\right) = \sum_{i=1}^{N} \exp\left(-\gamma \left\| \mathbf{x} - \mathbf{x}_i \right\|^2 \right)$$

where γ is the length scale.

In ε-SVR the optimal $\mathbf{w}$ and b are computed by minimizing the Eq. (1) [13]:

$$J = \sum_{i=1}^{N} w_i^2 + \frac{C}{N} \sum_{i=1}^{N} L_\varepsilon \left(y_i - f\left(\mathbf{x}_i \right) \right) \tag{1}$$

where

$$L_\varepsilon(y - f(\mathbf{x})) = \begin{cases} 0 \text{ if } |y - f(\mathbf{x})| \leq \varepsilon \\ |y - f(\mathbf{x})| \text{ otherwise} \end{cases}$$

and y_i is output data associated with $\mathbf{x}_i$. L_ε is the ε-insensitive loss function [10], C is a regularization parameter and ε is a ε-SVR parameter.

2.2.3 The eXtreme Gradient Boosting (XGB)

The eXtreme Gradient Boosting XGB [4] is an ensemble method for combining several low-complexity and high-error estimators, called weak learners, to generate a robust learner. In the case of the XGB, weak learners are regularized decision trees [5]. The XGB prediction for a instance i is

$$\hat{y_i} = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i)$$

where `M` is the number of estimators, and h_m are the weak learners built based on parameters of `max_depth` and a minimum loss factor for partitioning a new tree leaf (Γ_{tree}). Since it is a integrative boosting method, it is built in a greedy fashion of the form

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

where η is a learning rate and the newly added tree $h_m(x)$ is fitted in order to minimize a sum of losses L_m and is given by Eq. (2).

$$h_m = \arg\min_h L_m = \arg\min_h \sum_{i=1}^{N} l(y_i, F_{m-1}(x_i) + h(x_i)) + \Omega(F_{m-1}) \tag{2}$$

In Eq. (2), we take

$$l(y_i, F_m(x)) = \left[y_i - F_m\left(x_i \right) \right]^2$$

where N is the number of samples and

$$\Omega(F_m) = \alpha_{reg} T + \frac{1}{2} \lambda_{reg} ||\mathbf{w}||^2$$

is the a regularization term, T is the number of leaves in the decision trees (weak learners), $\mathbf{w}$ are the leaf weights, and α_{reg} and λ_{reg} are a L_1 and L_2 regularization term on weights, respectively.

Many researchers have recently obtained excellent results with the application of the XGB method to predict molecular interaction in Host-Guest systems [33,8]. Thus, we test this ML technique for our predictions.

2.2.4 Randomized Search (RS) strategy

In addition to the internal parameters adjusted in the training step, ML methods are generally sensitive to hyperparameter definitions [26]. We propose a hybrid approach coupling the ML methods with hyperparameters tuning through a Randomized Search (RS) strategy in this work. The RS performs a random choice of values, where each setting is sampled from a distribution over possible parameter values [1]. The process allows a budget to be chosen independently of the number of parameters; besides, adding parameters that do not influence the performance does not decrease efficiency.

Each RS run considers 3-fold cross-validation and 1000 machine samples. All optimized parameters follow uniform distributions. For ε-SVR, $\texttt{C} \in [0, 10^4]$, $\varepsilon \in [0, 10]$, and $\Gamma \in [0, 10]$. Each machine is internally adjusted in 10,000 iterations (maximum), using the RBF kernel. For XGB, $\texttt{M} \in [1, 300]$, $\texttt{max\_depth} \in [1, 10]$, $\eta \in [10^{-5}, 1]$, $\Gamma_{tree} \in [10^{-5}, 5]$, $\alpha_{reg} \in [10^{-5}, 2]$, $\lambda_{reg} \in [10^{-5}, 2]$. For GPR, $\texttt{n\_restarts} \in [0, 10]$ and the kernel were selected between `RBF` ($\sigma \in [0.1, 30]$) or `Matern` ($\sigma \in [0.1, 30]$ and $\nu \in [0.1, 3]$). All runs have $\alpha_{reg} = 50^{-5}$.

2.2.5 Model performance criteria

We apply three metrics to calculate the prediction errors

$$\mathbf{R^2} = 1 - \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \right) \tag{3}$$

$$\mathbf{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{4}$$

$$\mathbf{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{5}$$

where y is the measured value, $\hat{y}$ is the predicted value and $\bar{y}$ is the average of the measured values given by $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ considering the n instances i.

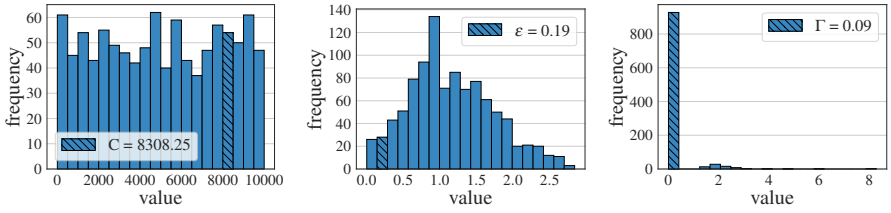
We analyzed the application domains of the best regression through a Williams plot. Williams plot is a method widely applied for evaluating application domains, which provides leverage values plotted against prediction errors [9]. The leverage value (h) measures how far a new instance i is from the centroid of the training set, and we can calculate it by

$$h_i = x_i^T (X^T X)^{-1} x_i$$

where X is the training set descriptor matrix [9]. Here, we considered the warning leverage $h^* = 3\frac{(p+1)}{n}$, where p is the number of molecular descriptors and n is the number of training samples. If a new instance of CD-ligand has leverage higher than the threshold h^*, its predictive value is considered unreliable. We may hold these molecules outside the descriptor space, either the application domain.

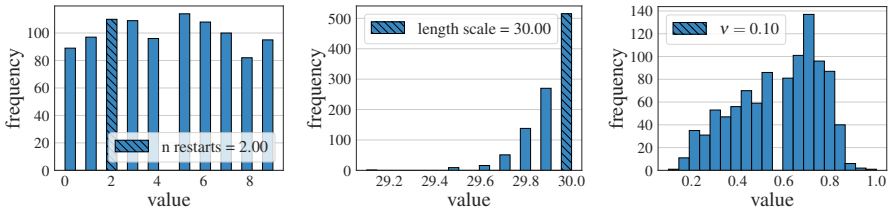
3 Results and Discussion

First, we present the results related to the hyperparameter search through a Randomized Search (RS). Each run of RS finds an adjusted machine (model) with an associated training RMSE. We considered the best-optimized model the machine associated with the lowest training RMSE. Here we present the distribution of the optimized parameters selected over 1000 runs using Random Search for each of the ML strategies. The SVR parameters are shown in Figure 3, the GPR parameters appears in Figure 4, while Figure 5 display the XGB parameters. The hatched bar indicates the range that involves the best-optimized model among the RS runs. Note that for SVR, Figure 3, even though ε search space was defined in a larger range, the best models were consistent, having $\varepsilon \leq 3$. A similar result is observed for Γ, once it mostly converges to values closer to zero. For GPR, none of the best results were based on `RBF` kernel selection. In Figure 4, we notice the sigma increases for GPR with the Matern kernel while nu decreases in the best results.



(a) `C` (b) ε (c) Γ

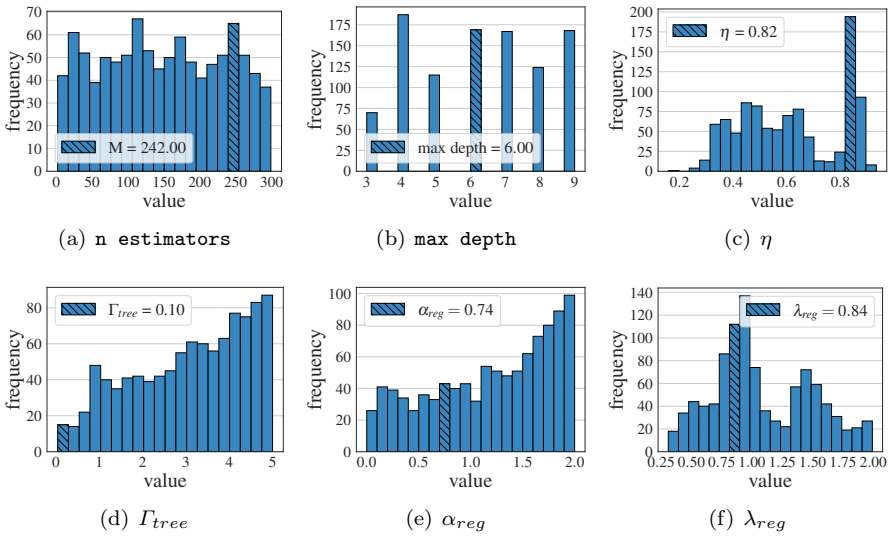
Fig. 3 Best SVR hyperparameters tuning distribution during 1000 runs of RS.



(a) `n restarts` (b) `length scale` (c) ν

Fig. 4 Best GPR hyperparameters tuning distribution during 1000 runs of Random Search. `length scale` and ν are `Matern` kernel parameters.

Table 2 shows the overall average metrics for training and testing sets for each method. The results demonstrate the consistency of the methods. The GPR achieved better results compared with the other methods. Table 3 shows the metrics for the best-optimized models of each tested method achieved



(a) `n estimators` (b) `max depth` (c) η

(d) Γ_{tree} (e) α_{reg} (f) λ_{reg}

Fig. 5 Best XGB hyperparameters tuning distribution during 1000 runs of RS.

Table 2 Average metrics (mean ± std) over 1000 runs of Random Search.

	Method	R^2 **score**	**RMSE (kJ/mol)**	MAE (kJ/mol)
Training	**SVR**	0.881 ± 0.040	1.621 ± 0.256	1.188 ± 0.399
	XGB	0.894 ± 0.027	1.537 ± 0.189	1.002 ± 0.170
	GPR	**0.943 ± 0.000**	**1.132 ± 0.004**	**0.421 ± 0.005**
Testing	**SVR**	0.727 ± 0.041	2.127 ± 0.154	1.692 ± 0.220
	XGB	0.719 ± 0.078	2.142 ± 0.285	1.567 ± 0.156
	GPR	**0.755 ± 0.020**	**2.015 ± 0.082**	**1.270 ± 0.028**

Table 3 Best metrics over 1000 runs of Random Search. RMSE and MAE in kJ/mol.

Method	**Training**			**Testing**		
	R^2	**RMSE**	**MAE**	R^2	**RMSE**	**MAE**
SVR	0.922	1.331	0.505	0.776	1.932	1.351
XGB	0.940	1.163	0.558	0.688	2.277	1.468
GPR	**0.953**	**1.034**	**0.396**	**0.803**	**1.811**	**1.201**

by the RS strategy. Table 4 presents the set of optimized hyperparameters associated with each best-optimized model. Again, we notice better results from GPR in every metric. Thus we follow the analysis focusing on this approach.

Figure 6 shows the comparison of experimental and predicted binding affinity obtained by the best GPR model according to Ligand Formal Charge (FC), Ligand Ratatable Bonds (RB), and the kind of cyclodextrin (CD). Figure 6(a)-(c) shows the high adjustability of the best-optimized model to the training data ($R^2 = 0.953$) along with the other evaluation metrics.

Table 4 Optimized hyperparameters set for each ML found through Random Search.

Method	Hyperparameters
SVR	`C`: 8308.25; ε:0.19; Γ:0.09
XGB	`M`: 242; `max_depth`: 6; η: 0.82; Γ_{tree}: 0.10; α_{reg}: 0.74; λ_{reg}: 0.84
GPR	`n_restarts`: 2; `Matern`(σ: 30; ν: 0.10)

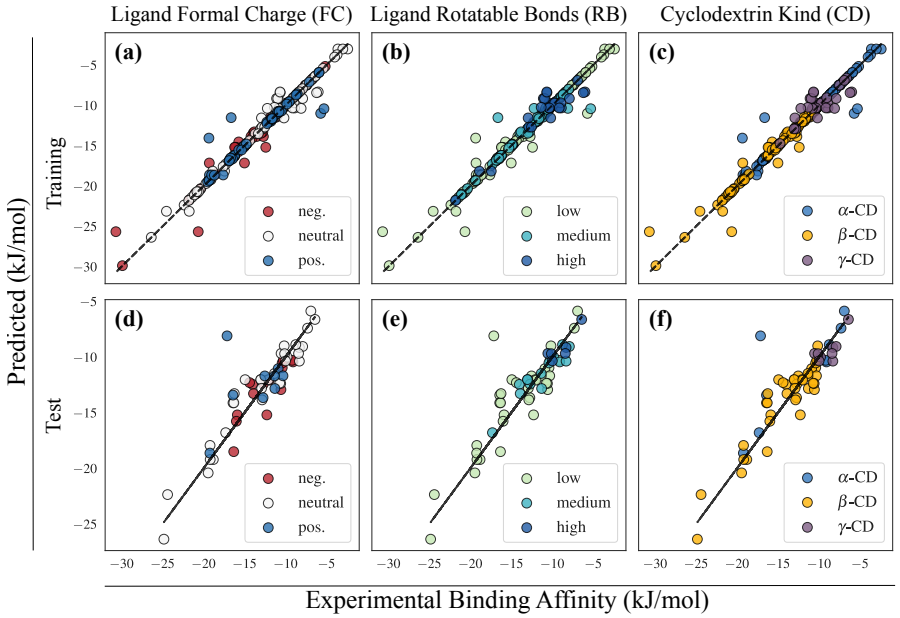


Fig. 6 Prediction results for the best GPR model. Subfigures (a)-(c) show the model ajustability in the training set ($R^2 = 0.953$, RMSE = 1.034 and MAE = 0.396). Subfigures (d)-(f) show the model prediction capability in the test set ($R^2 = 0.803$, RMSE = 1.811 and MAE = 1.201). Each consireded instance class (FC, RB, and CD) are discribed in Table 5.

We indicate instance attributes that frequently mislead the predictions. The definition of each class is described in Table 5.

Figure 6(d)-(f) shows the model prediction ability on the testing set. The generalization ability of the GPR model is evident according to the values of the performance metrics presented in Table 2. Compared to the model's performance obtained in the training set, the increase in the prediction errors for the test set was expected. However, it is observed that there was no overfitting since the worsening of performance in the test set is not significant, demonstrating the good performance of the strategy on a set of unseen samples. Figure 6 explicit the occurrence of an outlier prediction in the test set (sample with greatest error). The instance related to this prediction is the interaction between the α-CD host and the guest 1-butylamine – CCCC[NH3+]. The $\Delta G_{real} = -17.124$ and the $\Delta G_{pred} = -8.095$. In this particular instance, SVR and XGB models predicted $\Delta G_{pred} \sim -13.5$, which still characterizes a high error but closer to the actual value than GPR.

Table 5 RMSE values in kJ/mol (mean ± std) for each classification of ligand Formal Charge (FC), ligand Rotatable Bonds (RB), and host type. We considered a low ligand RB the range $[0, 3]$, medium RB the range $[4, 7]$, and high RB the range $[8, 11]$. We disregard the instance that generated the outlier prediction for this RMSE calculation.

Dataset Group		**RMSE Metric**		
FC	**Classes**	**Neg.**	**Neutral**	**Pos.**
	Training	1.017 ± 0.001	0.708 ± 0.001	1.821 ± 0.011
	Testing	1.564 ± 0.026	1.623 ± 0.114	1.196 ± 0.025
RB	**Classes**	**Low**	**Medium**	**High**
	Training	1.122 ± 0.005	1.132 ± 0.003	1.180 ± 0.002
	Testing	1.760 ± 0.096	1.135 ± 0.024	0.999 ± 0.025
CD	**Classes**	**α-CD**	**β-CD**	**γ-CD**
	Training	1.585 ± 0.009	0.801 ± 0.000	1.154 ± 0.001
	Testing	1.495 ± 0.115	1.646 ± 0.078	1.136 ± 0.015

Table 5 show the overall average RMSE (1000 runs) for each instance separation (charged ligand, ligand flexibility, and CD classes). There are indeed oscillations in the predicted measurements between the classes. Although the difference in RMSE does not overcome $\sim$ 0.8 kJ/mol, it should take into account before applying the ML method over new datasets.

Given the previous results, we need to verify if the error levels are compatible with the model's molecular interactions application domain. In this context, interactions of electrostatic potential, van der Waals, salt bridges, and hydrogen bonds (HB) are prevalent forms of interaction. Among them, hydrogen bonds and electrostatic interactions are determinants in receptor-ligand complexes [32].

HB is weaker than electrostatic interactions but extremely more frequent. The literature highlights that HB connections have an interaction potential always lower than 10 kJ/mol; in the vast majority of cases, they are lower than 3 kJ/mol [32], [2]. In protein complexes, for example, the contributions from the formation of HB are 5 ± 2.5 kJ/mol [32]. Thus, the error level in the predictions can be related to the number of HB inserted in the non-predicted range. For this work, an error of up to one non-predicted hydronium bond is considered acceptable.

The RMSE and the MAE metrics indicate an average error while keeping the same measure as our objective variable (kJ/mol). Observing our GPR best model RMSE and MAE in training (RMSE = 1.034 and MAE = 0.396) and testing (RMSE = 1.811 and MAE = 1.201), we see that the values remain safely below the threshold of 5 kJ/mol. The same is valid for our overall average analysis presented in Table 2. It guarantees an average error of only one hydrogen bond between the real and the predicted ΔG. We can also verify the best GPR model's application domain through the Williams plot in Figure 7. The majority of compounds in the training and testing sets are placed within the reliable area, indicating that these compounds are most likely to be well predicted by the model. There are four training compounds

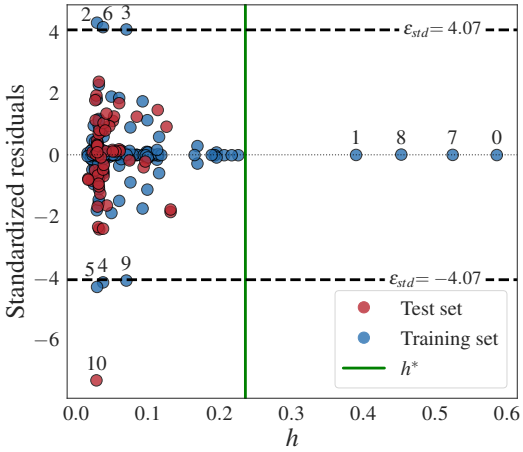


Fig. 7 Williams plot for the best GPR model.

Table 6 Outliers and border instances according to Williams plot for the best GPR model.

#	Instance ID	Host ID	Guest ID	h	Standardized Residual
0	**6_22**	BDBM11	BDBM5	0.586	-0.008
1	**3532_56**	BDBM11	BDBM36131	0.390	-0.005
2	**3529_16**	BDBM4	BDBM36019	0.029	4.303
3	**3535_4**	BDBM11	BDBM50029034	0.070	4.083
4	**3529_7**	BDBM11	BDBM36022	0.037	-4.157
5	**3529_14**	BDBM4	BDBM36018	0.029	-4.305
6	**3529_5**	BDBM4	BDBM36012	0.037	4.158
7	**3532_55**	BDBM11	BDBM36130	0.524	-0.003
8	**6_2**	BDBM4	BDBM5	0.453	0.006
9	**3535_5**	BDBM11	BDBM36146	0.070	-4.097
10	**3540_35**	BDBM4	BDBM36198	0.028	-7.349

with leverage values greater than the threshold of h^*. They differ from the other instances of ligand SMR, TPSA, or RB equal to zero (more info in the Complementary Results[1]). These instances may influence the model's performance in the training set but are not necessarily outliers to be deleted from the training data set once their standard residual values are shallow and within the established limit. Overall, the standardized residuals are smaller or in the limit of our safe region given by $\varepsilon_{std} = \pm\frac{5KJ/mol}{\sigma_{resid}} \sim \pm 4.07$, where σ_{resid} is the residuals standard deviation, except one instance of the test set (instance #10). This instance has already been identified in Figure 6(d)-(f). Once its descriptors leverage does not outrage the value h^*, it may behold as a worst-case scenario.

The labeled data in Figure 7 are instances arranged at the limit or outside the Williams plot security region. Table 6 relates host and guest BindingDB identifies with the relative h and standardized residuals obtained for the best GPR model.

Finally, we briefly compare the results with some papers presented in the literature. Note that this comparison is not over results based on the same

datasets. We only seek to demonstrate that, for the current database, the error levels are compatible, *a priori*, with it, is frequently presented in the literature. Dimas *et al.* [28] selected a set of complexes between β-CD and 57 small organic molecules that have been previously studied with the binding energy distribution analysis method in combination with an implicit solvent model. Even being a study focus only on the β-CD and applying a physics-based method, the errors levels were $R^2 = 0.66$ and RMSE = 9.330 kJ/mol, values that are worst than we achieved with our average metrics in Table 2 and best model in Figure 6(d)-(f).

Zhao *et al.* [33], on the other hand, also applied ML methods for predicting interaction energy between a vast number of cyclodextrin classes and ligand molecules (see Sect. 1). The best result obtained for Zhao was based on Light Gradient Boosting Machine (LightGBM) approach, similar to XGB approach, with metrics $R^2 = 0.86$, RMSE = 1.83 kJ/mol, and MAE = 1.38 kJ/mol. Comparing with our best XGB model presented in Table 3, we achieved competitive results considering our smaller database size. However, we may be on the edge of an overfitting situation as our training data were adjusted considerably better than our testing data. Since the XGB hyperparameter space has many degrees of freedom, it demands reasonable dataset sizes to achieve its optimal generability levels. Future analysis may address XGB search hyperparameters space, reducing possible values driving the solution for a local optimum of an overfitting occurrence. We may indicate the learning rate (η) hyperparameter (Figure 5(c)) as a possible overfitting originator. Our best machine indicates a high value in this parameter ($\eta = 0.82$), indicating the possibility of a biased solution and a high tendency of accumulating specific information of the training data. In this scenario, the ML does not get generalized solutions for the testing data, as we desire. Through our results in Figure 5(c), we may find a better solution for XGB when $\eta \leq 0.6$, or with a even smaller cutoff value.

Comparing the Zhao *et al.* [33] results with our best GPR model presented in Table 3 and Figure 6(d)-(f), we see that we achieved better MAE and RMSE values. Moreover, GPR is a method of less computational complexity compared to XGB. Disregarded the outlier prediction, the metrics are even better for the mean error metrics. However, the R^2 could not be adjusted with the same values as Zhao *et al.* [33].

4 Conclusion

This paper proposes an approach combining machine learning methods (SVR, XGB, and GPR) coupled with a Random Search hyperparameter tuning strategy to perform to predict the interaction between cyclodextrins host-guest systems. GPR achieved a better average and best results than the other methods. The proposed approach was enough to define a GPR model with great generality ($R^2 = 0.803$) and low associated errors (RMSE = 1.811 and MAE = 1.201) modest prediction on the application domain. The results

were compatible with the presented literature, even though using a less computational complexity method. As further research, we will apply other ML methods to this task and investigate using an evolutionary algorithm for optimizing the hyperparameters.

Acknowledgements The authors thank the financial support from FAPEMIG and CAPES.

Conflict of interest

The authors declare that they have no conflict of interest.

Apendix

Table 7 Input and output variables and their respective descriptions.

	Variable	Description
X_1	**pH**	Solution pH
X_2	**Temp**	Environment temperature
X_3	**SlogP – Guest**	Guest SlogP (log of the octanol/water partition coefficient)
X_4	**SMR – Guest**	Guest refractivity
X_5	**LabuteASA – Guest**	Guest Labutes approximate surface area
X_6	**TPSA – Guest**	Guest topological polar surface area
X_7	**AMW – Guest**	Guest average molecular weight
X_8	**NumLipinskiHBA – Guest**	Guest Lipinski number of receptors on hydrogen bonds
X_9	**NumLipinskiHBD – Guest**	Guest Lipinski number of donors in hydrogen bonds
X_{10}	**NumRotatableBonds – Guest**	Guest number of bonds with free rotational freedom
X_{11}	**NumAtoms – Guest**	Guest number of atoms
X_{12}	**Formal Charge – Guest**	Guest formal charge
X_{13}	**SlogP – Host**	Host SlogP (log of the octanol/water partition coefficient)
X_{14}	**SMR – Host**	Host refractivity
X_{15}	**LabuteASA – Host**	Host Labutes approximate surface area
X_{16}	**TPSA – Host**	Host topological polar surface area
X_{17}	**AMW – Host**	Host average molecular weight
X_{18}	**NumLipinskiHBA – Host**	Host Lipinski number of receptors on hydrogen bonds
X_{19}	**NumLipinskiHBD – Host**	Host Lipinski number of donors in hydrogen bonds
X_{20}	**NumRotatableBonds – Host**	Host number of bonds with free rotational freedom
X_{21}	**NumAtoms – Host**	Host number of atoms
y	ΔG	Binding free energy

Table 8 Descriptive statistics of the training set.

Variable	mean	std	min	25%	50%	75%	max
X_1	6.929	0.089	6.900	6.900	6.900	6.900	7.200
X_2	24.980	0.445	20.700	25.000	25.000	25.000	30.000
X_3	0.864	1.035	-4.756	0.298	0.825	1.557	4.342
X_4	52.243	26.540	0.000	35.376	44.481	57.051	129.975
X_5	83.718	43.094	23.783	52.684	71.665	92.797	209.562
X_6	58.354	39.443	0.000	27.640	43.280	81.248	171.600
X_7	199.471	105.300	58.085	125.920	165.168	231.761	514.705
X_8	3.326	2.500	0.000	1.000	2.000	5.000	10.000
X_9	1.781	1.591	0.000	0.000	1.000	3.000	5.000
X_{10}	3.473	2.693	0.000	2.000	3.000	5.000	11.000
X_{11}	28.393	12.766	3.000	21.000	25.000	30.000	79.000
X_{12}	-0.165	0.823	-3.000	-1.000	0.000	0.000	1.000
X_{13}	-14.939	1.413	-17.406	-15.231	-15.231	-13.055	-13.055
X_{14}	224.063	21.185	195.800	195.800	228.434	228.434	261.067
X_{15}	425.648	40.116	372.130	372.130	433.924	433.924	495.717
X_{16}	543.450	51.383	474.900	474.900	554.050	554.050	633.200
X_{17}	1113.272	105.260	972.846	972.846	1134.987	1134.987	1297.128
X_{18}	34.330	3.246	30.000	30.000	35.000	35.000	40.000
X_{19}	20.598	1.948	18.000	18.000	21.000	21.000	24.000
X_{20}	6.866	0.649	6.000	6.000	7.000	7.000	8.000
X_{21}	144.188	13.633	126.000	126.000	147.000	147.000	168.000
y	-13.084	4.762	-30.690	-16.320	-12.050	-10.278	-2.400

Table 9 Descriptive statistics of the test set.

Variable	mean	std	min	25%	50%	75%	max
X_1	6.911	0.056	6.900	6.900	6.900	6.900	7.200
X_2	24.981	0.050	24.850	25.000	25.000	25.000	25.000
X_3	0.776	0.924	-1.436	0.258	0.771	1.236	3.090
X_4	51.394	24.636	17.355	33.752	45.431	54.937	108.101
X_5	82.264	40.071	26.263	51.472	71.847	92.717	172.976
X_6	57.919	34.578	17.070	28.735	40.460	78.460	135.960
X_7	193.604	96.244	60.096	125.935	166.244	222.472	413.430
X_8	3.375	2.340	1.000	1.750	2.000	5.000	9.000
X_9	1.857	1.420	0.000	1.000	1.000	3.000	5.000
X_{10}	3.304	2.628	0.000	2.000	3.000	4.000	10.000
X_{11}	28.018	10.914	12.000	21.000	25.500	30.250	55.000
X_{12}	-0.125	0.689	-2.000	-1.000	0.000	0.000	1.000
X_{13}	-15.231	1.245	-17.406	-15.231	-15.231	-15.231	-13.055
X_{14}	228.434	18.669	195.800	228.434	228.434	228.434	261.067
X_{15}	433.924	35.351	372.130	433.924	433.924	433.924	495.717
X_{16}	554.050	45.280	474.900	554.050	554.050	554.050	633.200
X_{17}	1134.987	92.757	972.846	1134.987	1134.987	1134.987	1297.128
X_{18}	35.000	2.860	30.000	35.000	35.000	35.000	40.000
X_{19}	21.000	1.716	18.000	21.000	21.000	21.000	24.000
X_{20}	7.000	0.572	6.000	7.000	7.000	7.000	8.000
X_{21}	147.000	12.014	126.000	147.000	147.000	147.000	168.000
y	-12.951	4.115	-24.830	-16.052	-12.170	-10.232	-6.400

References

1. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. The Journal of Machine Learning Research **13**(1), 281–305 (2012)
2. Blundell, C.D., Nowak, T., Watson, M.J.: Measurement, interpretation and use of free ligand solution conformations in drug discovery. In: Progress in medicinal chemistry, vol. 55, pp. 45–147. Elsevier (2016)
3. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) **2**(3), 1–27 (2011)
4. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pp. 785–794. ACM, New York, NY, USA (2016)
5. Chen, T., He, T.: Higgs boson discovery with boosted trees. In: NIPS 2014 workshop on high-energy physics and machine learning, pp. 69–80 (2015)
6. Di, P., Chen, J., Liu, L., Li, W., Tang, Y., Liu, G.: In silico prediction of binding capacity and interaction forces of organic compounds with α-and β-cyclodextrins. Journal of Molecular Liquids **302**, 112585 (2020)
7. Gadade, D.D., Pekamwar, S.S.: Cyclodextrin based nanoparticles for drug delivery and theranostics. Advanced Pharmaceutical Bulletin **10**(2), 166 (2020)
8. Gao, H., Ye, Z., et al: Predicting drug/phospholipid complexation by the lightgbm method. Chemical Physics Letters p. 137354 (2020)
9. Golbraikh, A., Wang, X.S., Zhu, H., Tropsha, A.: Predictive qsar modeling: methods and applications in drug discovery and chemical risk assessment. In: Handbook of computational chemistry, pp. 1309–1342. Springer Netherlands (2012)
10. Gunn, S.R., et al.: Support vector machines for classification and regression. ISIS technical report **14**(1), 5–16 (1998)
11. Hu, Q.D., Tang, G.P., Chu, P.K.: Cyclodextrin-based host–guest supramolecular nanoparticles for delivery. Accounts of chemical research **47**(7), 2017–2025 (2014)
12. Jansook, P., Ogawa, N., Loftsson, T.: Cyclodextrins: structure, physicochemical properties and pharmaceutical applications. International journal of pharmaceutics **535**(1-2), 272–284 (2018)
13. Kargar, K., et al.: Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. Engineering Applications of Computational Fluid Mechanics **14**(1), 311–322 (2020)
14. Karthikeyan, M., Vyas, R.: Machine learning methods in chemoinformatics for drug discovery. In: Practical chemoinformatics, pp. 133–194. Springer (2014)
15. Katsila, T., et al: Computational approaches in target identification and drug discovery. Computational and structural biotechnology **14**, 177–184 (2016)
16. Kerner, J.J., von Recum, H.: Predicting drug interactions to unassociated biomedical implants using machine learning techniques and model polymers. bioRxiv (2020)
17. Kim, D.H., Lee, S.E., Pyo, Y.C., Tran, P., Park, J.S.: Solubility enhancement and application of cyclodextrins in local drug delivery. Journal of Pharmaceutical Investigation pp. 1–11 (2020)
18. Kumar, N., Hendriks, B.S., et al: Applying computational modeling to drug discovery and development. Drug discovery today **11**(17-18), 806–811 (2006)
19. Kumar, R.: The generalized modified bessel function and its connection with voigt line profile and humbert functions. Ad. in Applied Mathematics **114**, 101986 (2020)
20. Landrum, G.: Rdkit: Open-source cheminformatics software. GitHub and SourceForge **10**, 3592822 (2016)
21. Muankaew, C., et al.: Cyclodextrin-based formulations: a non-invasive platform for targeted drug delivery. Clinical Pharmacology & Toxicology **122**(1), 46–55 (2018)
22. Mura, P.: Advantages of the combined use of cyclodextrins and nanocarriers in drug delivery: A review. International Journal of Pharmaceutics p. 119181 (2020)
23. Ou, X., Morris, J., Martin, E.: Gaussian process regression for batch process modelling. IFAC Proceedings Volumes **37**(9), 817–822 (2004)
24. Rizzi, A., et al.: The SAMPL6 SAMPLing challenge: assessing the reliability and efficiency of binding free energy calculations. Journal of Computer-Aided Molecular Design **34**(5), 601–633 (2020)

25. Salústio, P.J., et al.: Advanced technologies for oral controlled release: cyclodextrins for oral controlled release. Aaps Pharmscitech **12**(4), 1276–1292 (2011)
26. Schmidt, M., et al.: On the performance of differential evolution for hyperparameter tuning. In: International Joint Conference on Neural Networks, pp. 1–8 (2019)
27. Solov'ev, V., Solovev, A.: 3d molecular fragment descriptors for structure-property modeling. In: 3rd Kazan Summer School on Chemoinformatics, pp. 70–70 (2017)
28. Suárez, D., Díaz, N.: Affinity calculations of cyclodextrin host–guest complexes: assessment of strengths and weaknesses of end-point free energy methods. Journal of chemical information and modeling **59**(1), 421–440 (2018)
29. Tang, P., Ma, X., et al.: Posaconazole/hydroxypropyl-β-cyclodextrin host–guest system: improving dissolution while maintaining antifungal activity. Carbohydrate polymers **142**, 16–23 (2016)
30. Tian, B., Hua, S., Liu, J.: Cyclodextrin-based delivery systems for chemotherapeutic anticancer drugs: A review. Carbohydrate Polymers **232**, 115805 (2020)
31. 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 **146**, 313–321 (2015)
32. Zerbe, O., Jurt, S.: Applied NMR spectroscopy for chemists and life scientists. John Wiley & Sons (2013)
33. Zhao, Q., Ye, Z., Su, Y., Ouyang, D.: Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques. Acta Pharmaceutica Sinica B **9**(6), 1241–1252 (2019)