A machine learning approach to discriminate MR1 binders: the importance of the phenol and carbonyl fragments

Running title: A model to discriminate MR1 binders

Jamal Shamsara[1,2]* (ORCID ID: 0000-0002-6162-6037), Gerrit Schüürmann[2,3] (ORCID ID: 0000-0002-3789-1703)

1 Pharmaceutical Research Center, Pharmaceutical Technology Institute, Mashhad University of Medical Sciences, Mashhad, Iran.

2 UFZ Department of Ecological Chemistry, Helmholtz Centre for Environmental Research, Leipzig, Germany.

3 Institute of Organic Chemistry, Technical University Bergakademie Freiberg, Freiberg, Germany

*Corresponding Author

Email: shamsaraj@mums.ac.ir

Abstract

Aims: In this study, we attempted to discriminate between MR1 binders and non-binders using machine learning (ML) approach and emphasized the important descriptors. Background: The major histocompatibility complex (MHC) class I-related molecule, MR1, is a component of the Immune system and interacts with T cell receptor (TCR) to modulate the immune response against various antigens. MR1 has raised many interests in recent years due to the potential of presenting a broader range of small molecules. MR1 has a small ligand-binding pocket interacting with agonistic or antagonistic ligands to stimulate or inhibit the immune response, respectively. Objective: There are limited studies on designing small molecules for the MR1 binding site, and the available raw data for MR1 binders is insufficient to exploit them for prioritizing chemicals. Therefore, the objective of this study was to provide validated and precise outcomes to expand the knowledge of critical structural features of MR1 binders. Method: We developed QSAR classifier models using Decision Tree (DT), Artificial Neural Network (ANN), Random Forest (RF), Extra Tree (ET), Linear Support Vector Machine (LSVM), Logistic Regression (LR), Naïve Bayesian classification (NB), and K-nearest-neighbors (KN). Result: The total accuracies for the best Machine Learning (ML) models were over 85%. The developed Decision Tree (DT) using suggested descriptors (fr_C_O_noCOO, fr_phenol, PEOE_VSA2) was able to classify the binders and non-binders with the accuracy of 85% for the train set and 100% for the test set. However, the 100% accuracy might be achieved by chance (due to simple random split of train/test set). DT models are easily interpretable. Therefore, a set of simple association rules was provided based on the provided DT model. Moreover, a LR equation was provided. Conclusion: The developed DT and LR models and rules could be used directly for ligand optimization, virtual screening, or re-scoring structure-based virtual screening results after consideration of the domain applicability. In general, the most important descriptors were found to be

fr_C_O_noCOO, fr_phenol, PEOE_VSA2 and to lesser extents, NumHDonors and VSA_Estate8 that were consistent with available crystallographic structures.

Keywords Classification; Decision Tree; Machine learning; Neural Network, MR1; QSAR

Introduction

MR1 is a receptor on the antigen processing cells. Mucosal-associated invariant T (MAIT) cells are an abundant population of innate-like T cells in humans that are activated by an antigen(s) bound to the MHC class I-like molecule MR1. It presents fragments of an antigen to the T cell receptors to activate MAIT cells immune response against a specific antigen [1]. MR1 is ideally suited to bind ligands originating from vitamin B metabolites. The well-known agonists that bind and interact with MR1 and TCR are riboflavin (vitamin B2) metabolites.

On the other hand, the folic acid (vitamin B9) metabolites are considered non-stimulatory to MAIT cells after interacting with MR1 binding pocket [2-5]. A recent study [6] demonstrated that despite the smallness of the MR1 ligand-binding pocket it could bind a variety of small molecules with either inhibitory or excitatory effects. That study paved the way for finding new potential therapeutic compounds that could interact with MR1.

The Quantitative Structure-Activity Relationship (QSAR) is a regression or classification method to predict the activity or class of new compounds and explores the possible mechanism of the interactions of ligands with the targets. It is routinely applied alongside experimental and other computational methods in drug development procedure [7-9]. QSAR modeling by Machine Learning (ML) methods has been helping the researchers to explore new possibilities. The various ML methods have been implemented in python to solve regression, clustering, classification and optimization problems [10]. ML methods are a collection of mathematical algorithms that aim to build predictive models in either a

supervised or a non-supervised manner. The supervised algorithms for classification include Neural Network, Decision Tree, etc. [11]

There are some standard metrics that if are applied wisely they will guide us to find a robust predictive model. This study compares different classification ML models alongside different metrics to ensure the validity of the proposed final models. In this study, we developed QSAR classifier models using Decision Tree (DT), Artificial Neural Network (ANN), Random Forest (RF), Extra Tree (ET), Linear Support Vector Machine (LSVM), Logistic Regression (LR), Naïve Bayesian classification (NB), and K-nearest-neighbors (KN). The ML methods are different in terms of performance and interpretability. DT is the most interpretable and transparent method among them, but it is prone to over-fitting. Therefore, variable selection is strongly suggested beforehand. That is also true in the case of LR, and variable selection is required before modeling. ANN, RF, ET, and SVM perform well on complex data. However, the developed model using these methods are less interpretable [12-14].

In addition, the quality of the data set has a direct and substantial impact on the developed QSAR model [15]. The availability of non-binders alongside the binders is beneficial. If a data set consists of a decoy set rather than a non-binder set, it will contain imprecise data, and the resulting model probably will be less robust. Decoy set is a library of often randomly selected molecules assumed to be inactive. Therefore, the decoy set may contain some unknown binders.

In this study, we attempted to discriminate between MR1 binders and non-binders using ML approach and emphasize the important descriptors. The workflow of the study is shown in Figure 1. To process the chemical data and applying ML methods, RDKit[16] and Scikit-learn[10] python libraries were employed, respectively.

Methods

Data set

A data set containing known binders and non-binders of MR1 was retrieved from recent studies [6, 17]. In a recent study [6], a set of non-binders has been provided alongside the new set of binders that is very helpful to construct a data set without the inclusion of compounds with unknown activity (Supplementary Material). Some other binders have been retrieved from other studies [2, 17]. Therefore, the data set containing all compounds (experimentally determined as binders = 30 or non-binders = 60) was constructed in SDF format with rationalized 3D structures. Most of the 3D structures were retrieved from PubChem database, and they have CID code within their names (Supplementary Material). For the other compounds, minimization was carried out using MMFF99 force field in Chem3D (PerkinElmer Informatics). Python implementation of the RDKit[16] package was used to calculate molecular descriptors. Gasteiger Partial charges have been computed using RDKit package. A set of 196 descriptors including 2D, 3D, and fragmental ones were calculated for each compound (Supplementary Material). The prepared descriptor table was used to build a series of statistically valid classifiers.

Machine learning algorithms

Pre-processing was performed to omit descriptors that all of the values were zero. Highly correlated descriptors were also omitted with a threshold of 0.9 (correlation coefficient). The descriptor scaling was carried out prior to modeling, except in the case of the Decision Tree (DT) modeling. A method called standardization was used to scale each of the individual descriptors. Standardization scales and centers in a parallel way.

To provide models that could classify the data set, a variety of ML methods have been employed. Because of the importance of model interpretability, it was attempted to employ the algorithms for which Scikit-learn has provided the feature importance extraction function. In the case of ANN, the employed

model had one hidden layer that makes it possible to extract feature importance by simple calculations [18-19]. In all modeling attempts 20% of the data set (n=18) was dedicated to the test set. The data was apportioned by train/test split function of Scikit-learn, and that was randomized. Besides, the different parameters for each method were explored (by grid search) to find the best hyperparameters for an individual model. The following classification methods were applied to the data set:

Support vector machine with linear kernel (LSVM) [12, 20]: SVM is a state-of-the-art statistical learning and a maximum margin classifier to perform classification, regression and outlier detection. It performs well even with a large-dimensional feature vector. We used a linear kernel in our study, and the penalty parameter of the error term was set to 1.0. In our study, The LSVM was applied to the range of descriptor sets with 5 to 100 descriptors. LSVM was the right choice for that purpose because it is a linear model with a low number of parameters to be optimized. The best results were obtained with 13 descriptors. "f_classif" function of Scikit-learn that computes ANOVA F-values for the descriptors was employed as the feature selection method.

Decision tree (DT) [12, 20]: DTs are tree-like graphs that suggest some rules, which are usually learned by splitting the set of training data into subsets based on the values for all variables. The top node presents the essential features. DTs are widely used as a predictive model or decision support tool because they are easily interpreted. Among the eight selected methods, DT was not able to develop a reliable model by 13 descriptors. The maximum depth was selected as a range of 2 to 10 for each set of descriptors. The best train and CV accuracy were achieved by 3 descriptors and a maximum depth of 3.

Artificial Neural Network (ANN) [13, 21]: ANN is originally based on the structure and the function of the network of neurons. ANN is a well established ML algorithm for developing QSAR models. The ANN used here was a Multi-Layer Perceptron (MLP) and trained with backpropagation. MLP consists of layers of neurons that each node in one layer connects with a certain weight to every node in the following layer.

MLP at least consists of three layers, input, hidden, and output layers. The hidden layer might be more than one. The L2 penalty parameter, number of hidden layers, and number of units in the hidden layer were set to 0.0001, 1, and 10, respectively. The activation function was set as the rectified linear unit function. Adam was used as the weight optimizer, and it is a stochastic gradient-based optimizer [22].

Random forest (RF) [13, 21]: RF consists of an ensemble of several weighted DTs. Therefore, it is not readily interpretable, and each DT generated using a different selected subset of the descriptors. If enough data points are available, it will perform well with high-dimensional data. RF is developed to improve the predictive power of the DT for high dimensional problems. To have better control of the level of fitting, the number of trees in the RF model was limited. The number of trees in the forest was set to 10 after a grid search.

Extra Trees (ET) [12, 23]: It is another algorithm based on the average of the randomized DTs, but the randomness is higher than the one in RF. For example, the ET algorithm splits nodes using entirely random cut points and grows the trees using the entire sample. Similar to the RF, the number of trees in the forest was set to 10.

Logistic regression (LR) [12, 23]: LR is a transformation of multiple linear regression to model the probabilities for classification problems with two possible outcomes. It uses the same formula as the logistic regression. However, instead of continuous output it deals with a two states output, 1 or 0. It can give the probability of each sample to either be in class 1 or 0 (here, binder, or non-binder). The LR function returns the probability of success (here to be binder). It is given by $p(x) = 1/(1 + \exp(-(B_0 + B_1X_1 + ... B_nX_n))$. B_0 is the intercept. B_1 through B_n are the coefficients. X_1 through X_n are the features.

Naïve Bayesian (NB): Gaussian naïve Bayes classification is a probabilistic supervised method that could classify the different classes based on Bayes' theorem. As implemented here, it does not have critical parameters to be tuned.

K-nearest-neighbors: K-nearest-neighbors (KN) classification is one of the most basic classification algorithms in ML. KN depends on a Euclidean distance metric between data points in order to predict the class labels. The K is a hyperparameter defined as the number of the nearest neighbors. It was selected from the set: 1, 3, 5 (default), 10, 30. Best CV results were obtained with K = 3.

The Scikit-learn package has a function that extracts feature importance of some models, including LSVM, RF, LR, and DT. For the ANN, we used a method called connection weights [18-19]. In this method, the product of input-hidden and hidden-output connection weights was calculated. In general, the absolute value of the provided importance value shows the extent of their effects on the predicted result, whereas the sign of the values shows how a descriptor affects the compounds to be in either of the classes, binders, or non-binders.

Model assessments

The results of a classification job can be classified into four categories:

True positive (TP): The active (here is binder) compounds, which have been predicted as active.

True negative (TN): The inactive (here is non-binder) compounds which have been predicted as inactive.

False positive (FP): The inactive compounds which have been predicted as active.

False negative (FN): The active compounds which have been predicted as inactive.

Based on these classes, the following metrics were used to select the best models. Each metric can be calculated for the train, cross-validation, test or total set:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

All the metrics are between 0 (the worst prediction) and 1 (the best prediction) except Matthews Correlation Coefficient (MCC). The MCC is regarded as a balanced measure, which can be employed even if the classes are of diverse sizes. The MCC is merely a correlation coefficient between the observed and predicted binary classifications, and it returns a value between −1 and +1. A coefficient of +1 signifies a perfect prediction, 0 an average random prediction, and −1 an inverse prediction [24].

All the metrics were calculated for the train, cross validation and test sets and according to the QSAR guideline of OECD (Organization for Economic Cooperation and Development) they demonstrates goodness-of-fit, robustness and predictivity of a model, respectively. The test set is only used once the model is developed. Therefore, the test set data are considered as unseen to the developed model. Another method to assess the predictive ability of the model is cross validation. During cross-validation, each time a part of the train set was removed from the input data and then used for the model evaluation. Leave one out (LOO), and random sampling (validation set = 20%) with 2000 repetitions were employed as cross-validation (CV) methods. This method is also called Monte Carlo cross-validation. The purpose of CV is to detect a possible over-fitted model with high internal accuracy whereas low external prediction power. LOO is performed by repeatedly removing one data point at a time form the train set whereas, in random sampling, a group of data points (here, 20%) were removed each time. The modeling process is shown in Figure 2.

Results and discussion

Model development and validation

In this study, we have introduced ML models discriminating between MR1 binders and non-binders with good internal and external accuracy, sensitivity, specificity, and precision. DT has an interpretable and straightforward output that makes it favorable for classification QSAR tasks. The DT and LR models presented here had both robust internal and external predictive power. The ANN, RF, and ET usually require higher observation to descriptors ratio to demonstrate its modeling potentials. Therefore, those complex methods were not good choices for the data set. Moreover, the equation of the LR model is provided and explained.

Table 1 demonstrates the statistical parameters for the best models achieved by several ML algorithms. The models were selected based on both train set accuracy and Cross-validated (CV) accuracy. The model building without feature selection resulted in the over-fitted models with low CV and external set accuracy. The graph of DT is shown in Figure 3, and the extracted rules are shown in the Supplementary Material. It should be noted that the 100% accuracy for DT test might be achieved by chance (due to simple random split of train/test set). The ANN and RF models seem to suffer from the over-fitting problem. In the next section, it has been shown that the critical descriptors suggested by these two methods were slightly different from the other methods. The individual prediction for each compound is shown in Supplementary Material.

The significant difference between train set accuracy and LOO CV accuracy (> 0.25-0.30) could be a sign of an over-fitting problem, and it is the case for the ANN and RF models in our study (Figure 4). The over-fitting issue could be responsible for the different sets of descriptors proposed by these models.

There is a reverse correlation between the accuracy of external prediction (test set) for the models and the difference between train set accuracy and LOO CV (Leave One Out Cross-validated) accuracy (Figure 4, Pearson's correlation coefficient = -0.64). It emphasizes the reliability of CV accuracy in the

selection of the most predictive models. The negative correlation was because of the over-fitting issue of the models with a great difference between the train and LOO CV accuracy. In those cases despite an increase in internal accuracy, external predictive power decreased.

Descriptor Importance

One of the flaws in QSAR studies, especially with a high ratio of descriptors to compounds (high dimensional data) is the chance correlation. Using different algorithms with a consistent set of proposed important descriptors could decrease the occurrence of chance correlation in the models. In addition, to understand which characteristics make a compound an MR1 binder we attempted to analyze the importance of the descriptors involved the best models.

Table 2 shows the most important descriptors proposed by each model for classification of the set. They have grouped based on either increasing or decreasing the effect on the probability of MR1 binding capability, generally. The consistent descriptors are shown in bold faces. Among the most significant descriptors fr_C_O_noCOO, fr_phenol, PEOE_VSA2, NumHDonors, and VSA_Estate8 were shared among the various models. fr_C_O_noCOO and fr_phenol indicate the number of carbonyls (except carboxyl) and phenol groups, respectively. , NumHDonors is equal to the number of hydrogen bond donor groups in the molecule. The *van der Waals surface* area (VSA) is a volume surface area that can be calculated either for the whole molecule or for parts of the molecule with specified attributes. PEOE_VSA2 corresponds to the partition of the molecular surface area conditioned by the partial charges between -0.25 and -0.30, whereas and VSA_Estate8 corresponds to the electrotopological state (Estate) indices values between 6.45 and 7.0.

The calculated LR equation shows a general trend of the effect of descriptors on the probability of a compound to be active p (x):

p (x) = 1/ (1+ exp (0.92 - 0.74 (fr_phenol) - 0.63 (fr_C_O) – 0.59 (fr_C_O_noCOO) – 0.56 (PEOE_VSA2) – 0.52 (NumHDonors) – 0.28 (PEOE_VSA11) – 0.22 (EState_VSA7) – 0.06 (SMR_VSA1) – 0.05 (EState_VSA10) + 0.34 (PEOE_VSA3) + 0.43 (VSA_EState9) + 0.48 (PEOE_VSA7) + 0.75 (VSA_EState8)))

For details on the equation, see the Supplementary Materials.

The average values and distribution of the crucial descriptors for both groups are shown in Figure 5. Figure 5 demonstrates that the most of the suggested descriptors by the ML algorithms were differently distributed between Binders and non-binders. For example, number of carbonyl fragments are generally higher in the binders group than non-binders group. Similarly, in the case of PEOE_VSA2 the difference between two groups is obvious. A 3D scatter plot for three important descriptors that were used by DT is shown in Figure 6. It represents the value of fr_phenol, fr_C_O_noCOO and PEOE_VSA2 for the data set. It has decreased the variable space to three descriptors. In addition, the value of fr_phenol and fr_C_O_noCOO descriptors are discrete. Therefore, the data points are overlapping. This may cause loose of some information. However, the limitation of the variable space could be an effective way to achieve a generalized model with a simple algorithms like DT. Locally, their impact also depends on the values of other descriptors; however, the extracted feature importance shows a general trend of a descriptor on the probability of being either MR1 binder or non-binder across the entire train set.

Analysis of the interactions of MR1 ligands with MR1 active site

The available MR1 X-ray structures with a unique co-crystallized ligand are listed in Table 3. Some of the corresponding interactions are shown. As it is shown, most of the compounds make a covalent bond with the Lys43 residue of the MR1 binding site. The Schiff base formation after reaction between the ligand carbonyl group and the amino group of the Lys43 is shown in Figure 7. The carbonyl group of the OP-RU, OP-RE, RL-6-Me-7-OH, and Ac-6-FP ligands makes hydrogen bond as a hydrogen bond acceptor

with the amino group of the Arg9 and Arg94 residues of the MR1 active site. π- π interactions are seen in the interactions list for all ligands. This finding was consistent with one of the important descriptors, fr_phenol, which was suggested by the developed ML models and this is discussed in the next section.

The experimental binding pose of MR1 inhibitors showed that most of them make a covalent bond with Lys43 [25]. On the other hand, it was found that the fr_C_O_noCOO descriptor had a significant effect on several ML classifiers performance. The number of the carbonyl group could be attributed to the possibility of Schiff base formation with the amino group of Lys43. Moreover, carbonyl acted as a hydrogen bond acceptor for the amino group of the Arg9 and Arg94.

The number of phenyl fragments could be attributed to the ability of a ligand to make π- π interaction with Tyr7, Tyr62 or Tyr169 of the active site [2, 6]. In addition, in some cases, the phenolic hydroxyl group made hydrogen bonds with His58 (Table 3).

The importance of NumHDonors descriptor could be explained by the fact that most of the MR1 co-crystallized ligands made several hydrogen bonds as hydrogen-bond donors with MR1 active site residues [2, 6].

VSA_Estate and to a lesser extent, PEO_VSA descriptor families are considered hard to interpret. Generally, they are correspondent to the charge and electron distribution and topological state of the molecule.

Conclusion

The MR1 has been suggested as a therapeutic target, especially since the analyses of the ligand interactions indicated the ability of MR1 binding pocket to bind a variety of small molecules in addition to the processed antigens like vitamin B2 metabolites [6]. Our study was conducted to pave the path to find

new small ligands as lead structures for future drug development. We plan to use the developed DT and LR models and extracted rules for rescoring of structure-based virtual screening results on MR1 for possible enrichment of the true binders. To accomplish that, one should keep in mind the importance of the applicability domain. A QSAR model could be trustfully applicable only if the tested compound falls within the applicability domain of it.

Availability of Data and Materials

The data that supports the findings of this study are available in the Supplementary Material of this article.

Supplementary Material: An excel file includes descriptor table, the 2D images of the entire dataset, name of the descriptors, descriptor importance suggested by the models, details of the LR model, and model performance metrics are provided as Supplementary Material.

Acknowledgments

This work was supported in part by Mashhad University of Medical Sciences (MUMS). A part of the study was carried out at the Helmholtz Center for Environmental Research (UFZ).

Conflict of interest

The authors declare no conflict of interest, financial or otherwise.

References

1. Lopez-Sagaseta, J.; Dulberger, C. L.; McFedries, A.; Cushman, M.; Saghatelian, A.; Adams, E. J. MAIT recognition of a stimulatory bacterial antigen bound to MR1. *Journal of immunology (Baltimore, Md. : 1950)*, **2013,** *191* (10), 5268-77.
2. McWilliam, H. E.; Birkinshaw, R. W.; Villadangos, J. A.; McCluskey, J.; Rossjohn, J. MR1 presentation of vitamin B-based metabolite ligands. *Current opinion in immunology*, **2015,** *34*, 28-34.

3. Eckle, S. B.; Corbett, A. J.; Keller, A. N.; Chen, Z.; Godfrey, D. I.; Liu, L.; Mak, J. Y.; Fairlie, D. P.; Rossjohn, J.; McCluskey, J. Recognition of Vitamin B Precursors and Byproducts by Mucosal Associated Invariant T Cells. *The Journal of biological chemistry*, **2015,** *290* (51), 30204-11.
4. Birkinshaw, R. W.; Kjer-Nielsen, L.; Eckle, S. B.; McCluskey, J.; Rossjohn, J. MAITs, MR1 and vitamin B metabolites. *Current opinion in immunology*, **2014,** *26*, 7-13.
5. Patel, O.; Kjer-Nielsen, L.; Le Nours, J.; Eckle, S. B.; Birkinshaw, R.; Beddoe, T.; Corbett, A. J.; Liu, L.; Miles, J. J.; Meehan, B.; Reantragoon, R.; Sandoval-Romero, M. L.; Sullivan, L. C.; Brooks, A. G.; Chen, Z.; Fairlie, D. P.; McCluskey, J.; Rossjohn, J. Recognition of vitamin B metabolites by mucosal-associated invariant T cells. *Nat Commun*, **2013,** *4*, 2142.
6. Keller, A. N.; Eckle, S. B.; Xu, W.; Liu, L.; Hughes, V. A.; Mak, J. Y.; Meehan, B. S.; Pediongco, T.; Birkinshaw, R. W.; Chen, Z.; Wang, H.; D'Souza, C.; Kjer-Nielsen, L.; Gherardin, N. A.; Godfrey, D. I.; Kostenko, L.; Corbett, A. J.; Purcell, A. W.; Fairlie, D. P.; McCluskey, J.; Rossjohn, J. Drugs and drug-like molecules can modulate the function of mucosal-associated invariant T cells. *Nature immunology*, **2017,** *18* (4), 402-411.
7. Wold, S.; Eriksson, L.; Clementi, S. Statistical Validation of QSAR Results. In *Chemometric Methods in Molecular Design*, **2008**; Vol. 2, pp 309-338.
8. Hadizadeh, F.; Shamsara, J. Receptor-based 3D-QSAR approach to find selectivity features of flexible similar binding sites: case study on MMP-12/MMP-13. *Int. J. Bioinform. Res. Appl.*, **2015,** *11* (4), 326-346.
9. Shamsara, J. A random forest model to predict the activity of a large set of soluble epoxide hydrolase inhibitors solely based on a set of simple fragmental descriptors. *Combinatorial chemistry & high throughput screening*, **2019**.
10. Pedregosa, F.; Ga; Varoquaux, l.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, d. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **2011,** *12*, 2825-2830.
11. Kotsiantis, S. B.; Zaharakis, I. D.; Pintelas, P. E. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, **2006,** *26* (3), 159-190.
12. Choudhary, R.; Gianey, H. K. In *Comprehensive Review On Supervised Machine Learning Algorithms*, 2017 International Conference on Machine Learning and Data Science (MLDS), 14-15 Dec. 2017; 2017; pp 37-43.
13. Somvanshi, M.; Chavan, P. In *A review of machine learning techniques using decision tree and support vector machine*, 2016 International Conference on Computing Communication Control and automation (ICCUBEA), 12-13 Aug. 2016; 2016; pp 1-7.
14. Mollazadeh, S.; Shamsara, J.; Iman, M.; Hadizadeh, F. Docking and QSAR studies of 1,4-dihydropyridine derivatives as anti-cancer agent. *Recent patents on anti-cancer drug discovery*, **2017,** *12* (2), 174-185.
15. Wenlock, M. C.; Carlsson, L. A. How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models. *Journal of chemical information and modeling*, **2015,** *55* (1), 125-134.
16. Landrum, G. RDKit: open-source cheminformatics. http://www.rdkit.org. 2006.
17. Keller, A. N.; Corbett, A. J.; Wubben, J. M.; McCluskey, J.; Rossjohn, J. MAIT cells and MR1-antigen recognition. *Current opinion in immunology*, **2017,** *46*, 66-74.
18. Olden, J. D.; Joy, M. K.; Death, R. G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, **2004,** *178* (3), 389-397.
19. Olden, J. D.; Jackson, D. A. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, **2002,** *154* (1), 135-150.

20. Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. *Drug discovery today*, **2015,** *20* (3), 318-331.
21. Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **2014,** *4* (5), 468-481.
22. Zhong, H.; Chen, Z.; Qin, C.; Huang, Z.; Zheng, V. W.; Xu, T.; Chen, E. Adam revisited: a weighted past gradients perspective. *Frontiers of Computer Science*, **2020,** *14* (5), 145309.
23. Lima, A. N.; Philot, E. A.; Trossini, G. H.; Scott, L. P.; Maltarollo, V. G.; Honorio, K. M.; Saez, Y.; Baldominos, A.; Isasi, P. Use of machine learning approaches for novel drug discovery

A Comparison Study of Classifier Algorithms for Cross-Person Physical Activity Recognition. *Expert opinion on drug discovery*, **2016,** *11* (3), 225-39.
24. Roy, K.; Kar, S.; Das, R. Statistical Methods in QSAR/QSPR. In *A Primer on QSAR/QSPR Modeling*, Springer International Publishing: **2015**; pp 37-59.
25. Mak, J. Y.; Xu, W.; Reid, R. C.; Corbett, A. J.; Meehan, B. S.; Wang, H.; Chen, Z.; Rossjohn, J.; McCluskey, J.; Liu, L.; Fairlie, D. P. Stabilizing short-lived Schiff base derivatives of 5-aminouracils that activate mucosal-associated invariant T cells. *Nat Commun*, **2017,** *8*, 14599.
26. Stierand, K.; Rarey, M. Drawing the PDB: Protein−Ligand Complexes in Two Dimensions. *ACS medicinal chemistry letters*, **2010,** *1* (9), 540-545.

Table 1. The accuracy of the best models.

*The confusion matrixes are presented as $\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$

**Acc.: Accuracy, Spec.: Specificity, Prec.: Precision, CV: Cross-validation with 2000 times Resampling.

^ The 100% accuracy might be achieved by chance (due to simple random split of train/test set)

Algorithm	Confusion matrix (train)*	Confusion matrix (test)*	Acc.**						Spec.** (Total)	Sens.** (Total)	Prec. **(Total)	MCC (Total)
			LOO	CV**	Train	Test	train - CV	Total				
DT	$\begin{bmatrix} 17 & 4 \\ 7 & 44 \end{bmatrix}$	$\begin{bmatrix} 6 & 0 \\ 0 & 12 \end{bmatrix}$	**0.85**	**0.83**	0.85	1.00^	0.00	0.88	0.93	0.77	0.85	0.72
ANN	$\begin{bmatrix} 24 & 0 \\ 0 & 48 \end{bmatrix}$	$\begin{bmatrix} 5 & 3 \\ 1 & 9 \end{bmatrix}$	0.63	0.68	1.00	0.78	0.32	0.96	0.95	0.97	0.91	0.90
RF	$\begin{bmatrix} 23 & 0 \\ 1 & 48 \end{bmatrix}$	$\begin{bmatrix} 5 & 1 \\ 1 & 11 \end{bmatrix}$	0.68	0.69	0.99	0.89	0.25	0.97	0.98	0.93	0.97	0.92
ET	$\begin{bmatrix} 24 & 0 \\ 0 & 48 \end{bmatrix}$	$\begin{bmatrix} 5 & 1 \\ 1 & 11 \end{bmatrix}$	0.75	0.71	1.00	0.89	0.27	0.98	0.98	0.97	0.97	0.95
LSVM	$\begin{bmatrix} 16 & 3 \\ 8 & 45 \end{bmatrix}$	$\begin{bmatrix} 6 & 2 \\ 0 & 10 \end{bmatrix}$	0.71	0.71	0.85	0.89	0.14	0.86	0.92	0.73	0.81	0.67
LR	$\begin{bmatrix} 16 & 3 \\ 8 & 45 \end{bmatrix}$	$\begin{bmatrix} 6 & 1 \\ 0 & 11 \end{bmatrix}$	**0.80**	**0.75**	0.84	0.94	0.04	0.86	0.93	0.73	0.85	0.69
NB	$\begin{bmatrix} 14 & 6 \\ 10 & 42 \end{bmatrix}$	$\begin{bmatrix} 5 & 1 \\ 1 & 11 \end{bmatrix}$	0.74	0.72	0.78	0.89	0.04	0.80	0.82	0.73	0.63	0.54
KN (K=3)	$\begin{bmatrix} 17 & 3 \\ 7 & 45 \end{bmatrix}$	$\begin{bmatrix} 3 & 1 \\ 3 & 11 \end{bmatrix}$	0.72	0.69	0.86	0.78	0.14	0.84	0.85	0.83	0.67	0.64

Table 2. Important descriptors suggested by the ML models after feature selection.

*>= 3 occurrence at the top of the descriptor list within the 6 models

Algorithm	Important descriptors with the positive effect on binding	Important descriptors with the negative effect on binding
Consistent *	**fr_C_O_noCOO, fr_phenol, PEOE_VSA2, NumHDonors**	**VSA_Estate8**
DT	fr_C_O_noCOO, fr_phenol, PEOE_VSA2	
ANN	PEOE_VSA2, NumHDonors	VSA_EState8, PEOE_VSA3, SMR_VSA1
RF	PEOE_VSA11	VSA_EState9 VSA_EState8 PEOE_VSA7
ET	fr_C_O_noCOO fr_phenol	PEOE_VSA7
LSVM	fr_phenol, PEOE_VSA2, fr_C_O_noCOO, NumHDonors, fr_C_O	EState_VSA10, VSA_EState8, VSA_Estate9, VSA_Estate7
LR	fr_phenol, fr_C_O_noCOO, PEOE_VSA2, NumHDonors	VSA_EState8

Table 3. The summary of the interactions between MR1 binders with available co-crystallized X-ray structures. Only interactions that could be related to the critical descriptors suggested by the ML models have been mentioned. The interactions were identified by PoseView.

* Hydrogen Bond Donor

PDB ID	Ligand	hydrogen bond with a ligand carbonyl	A covalent bond with Lys43	π- π interaction	Number of HBDs*	Phenolic hydroxyl group (HBD)
4nqc	5-OP-RU	With Arg9	Yes	TYR7	3	-
4nqe	5-OE-RU	With Arg9	Yes	TYR7	1	-
4l4v	RL-6-Me-7-OH	With Arg9	No	TYR7	1	-
5u1r	Diclofenac	-	No	TYR7, TYR62	2	-
5u72	5OH-Diclofenac	-	No	TYR7, TYR62, TRP164	1	With a TCR residue
4gup	6FP	-	Yes	TYR7	1	-
4pj5	Ac-6-FP	With Arg9 and Arg94	Yes	TYR7	0	-
5u17	DA-6-FP	-	Yes	TYR7	0	-

5u16	2-OH-1-NA	-	Yes	TYR7, TYR62	1	With His58
5u2v	HMB	-	Yes	TYR7, TYR62	1	With His58
5u6q	3-F-SA	-	Yes	TYR7	0	No

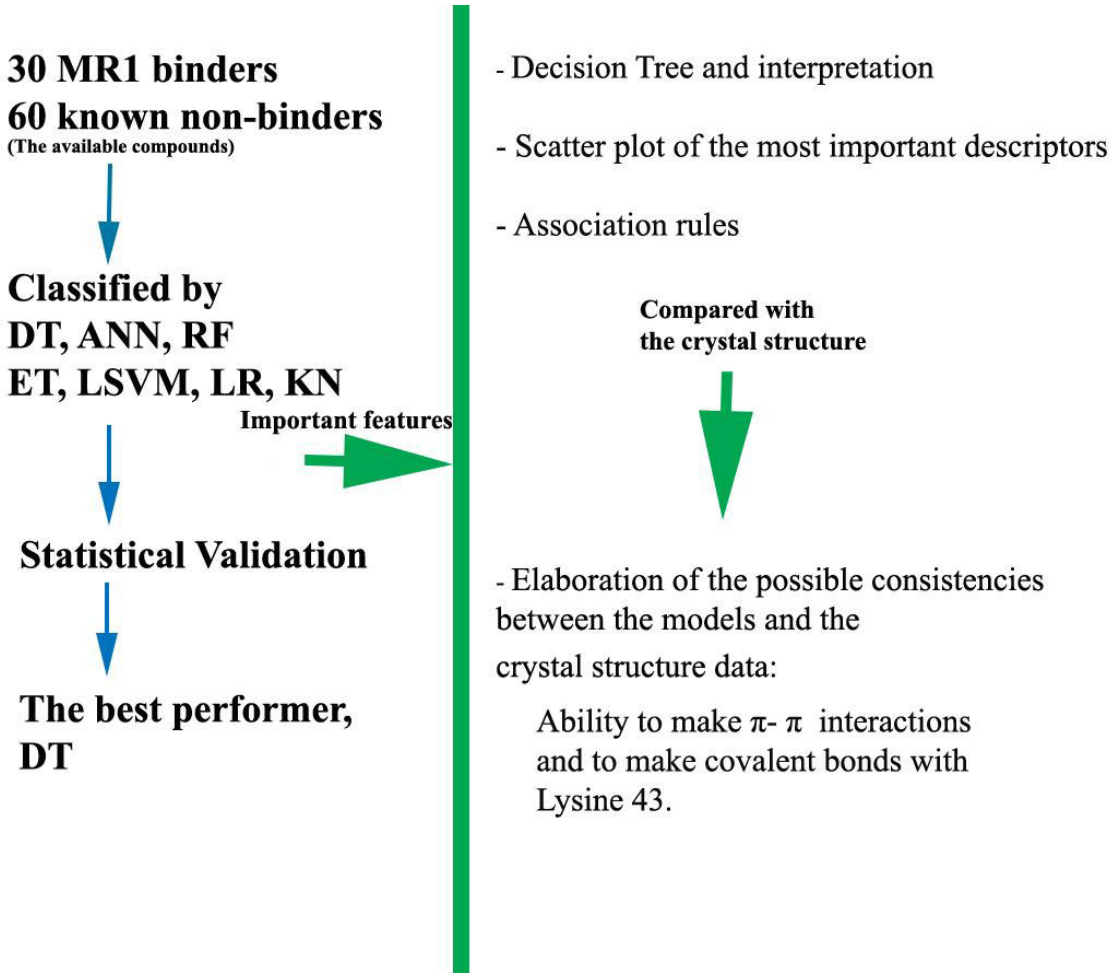


Figure 1. The workflow of the study.

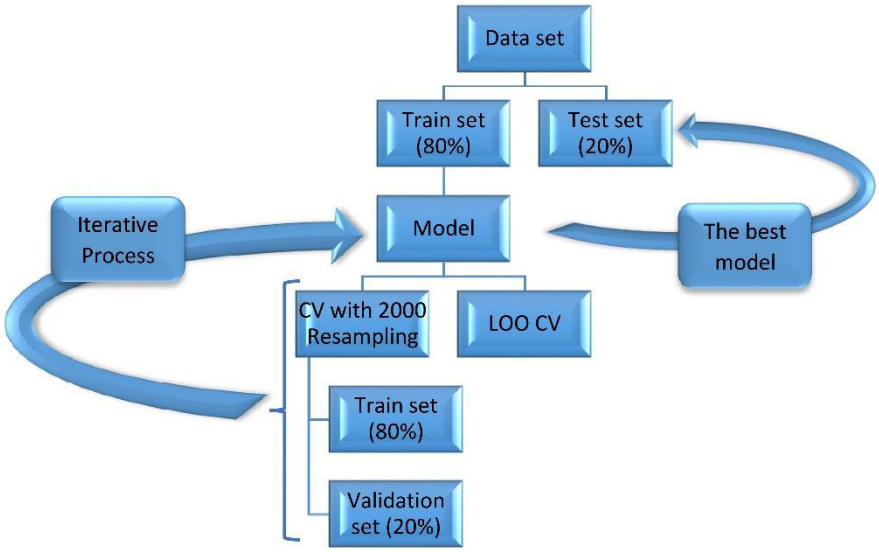


Figure 2. The modeling process. During the search for the optimum hyperparameters and number of descriptors, several models were developed, and the best one for each algorithm was chosen based on the train and CV results. Then, the best models were applied to the test set to assess the prediction power and validity.

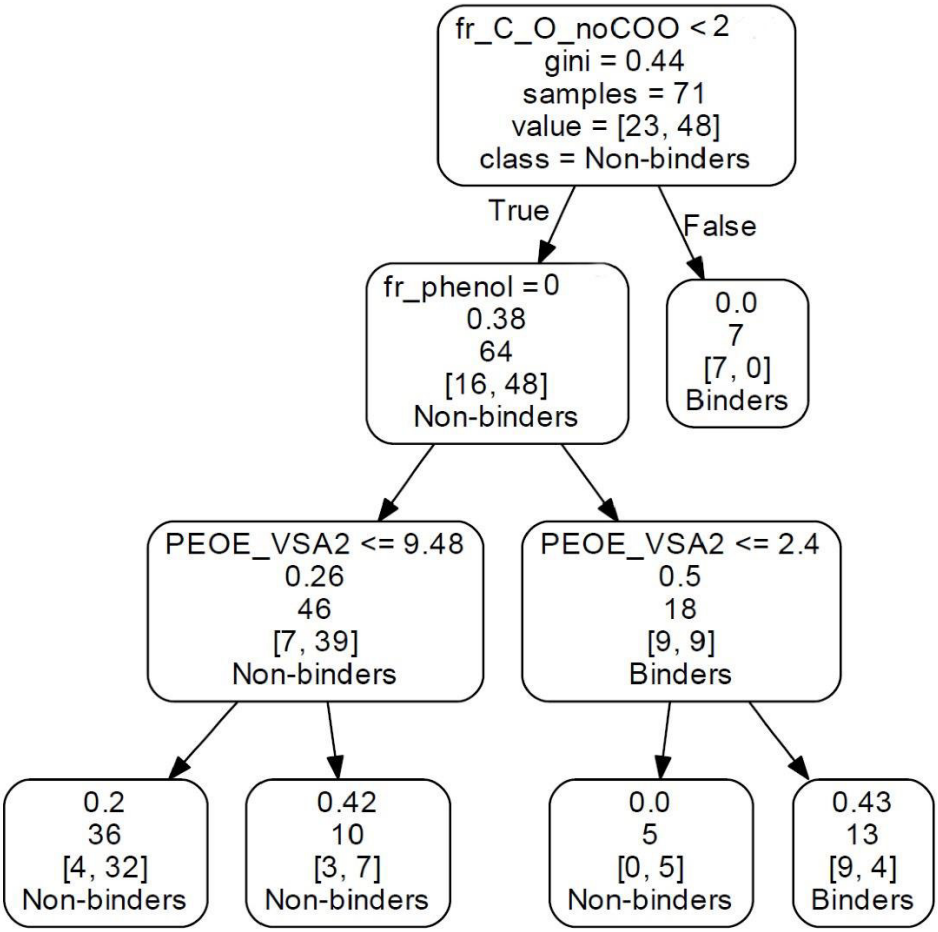


Figure 3. Training of the DT model with three variables and a maximum depth of 3. In each represented rectangle (node), the first row indicates the decision rule. The second one is Gini. Gini is a measure of impurity of a node. The third row shows the number of sample in the node. The fourth row demonstrates the number of compounds. The fifth row indicates the class that majority of the compounds in the node belong to it.

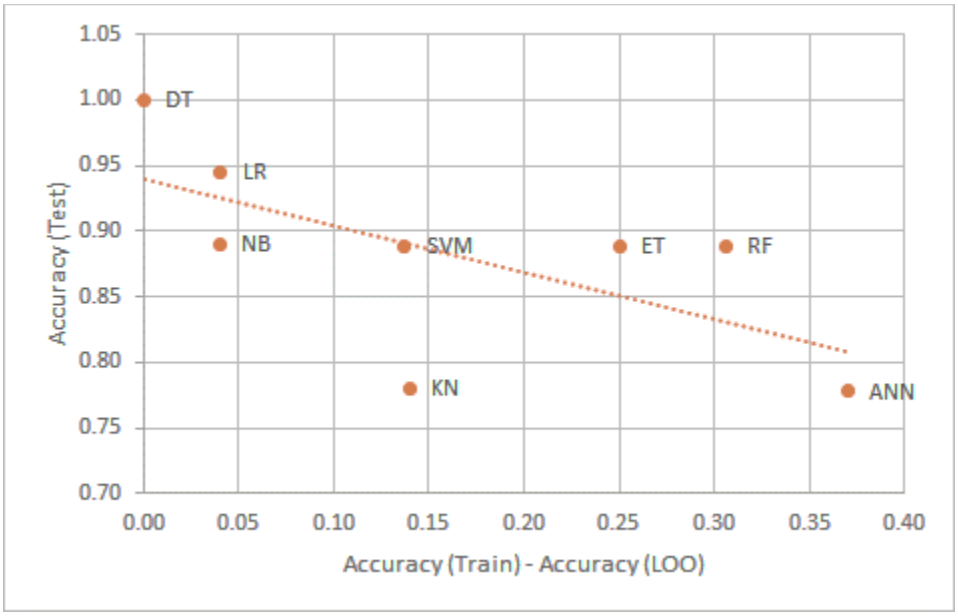


Figure 4. Negative correlation between test set prediction accuracy and the difference between accuracy of the train set prediction and the accuracy of cross-validation (Pearson's correlation coefficient =-0.64).

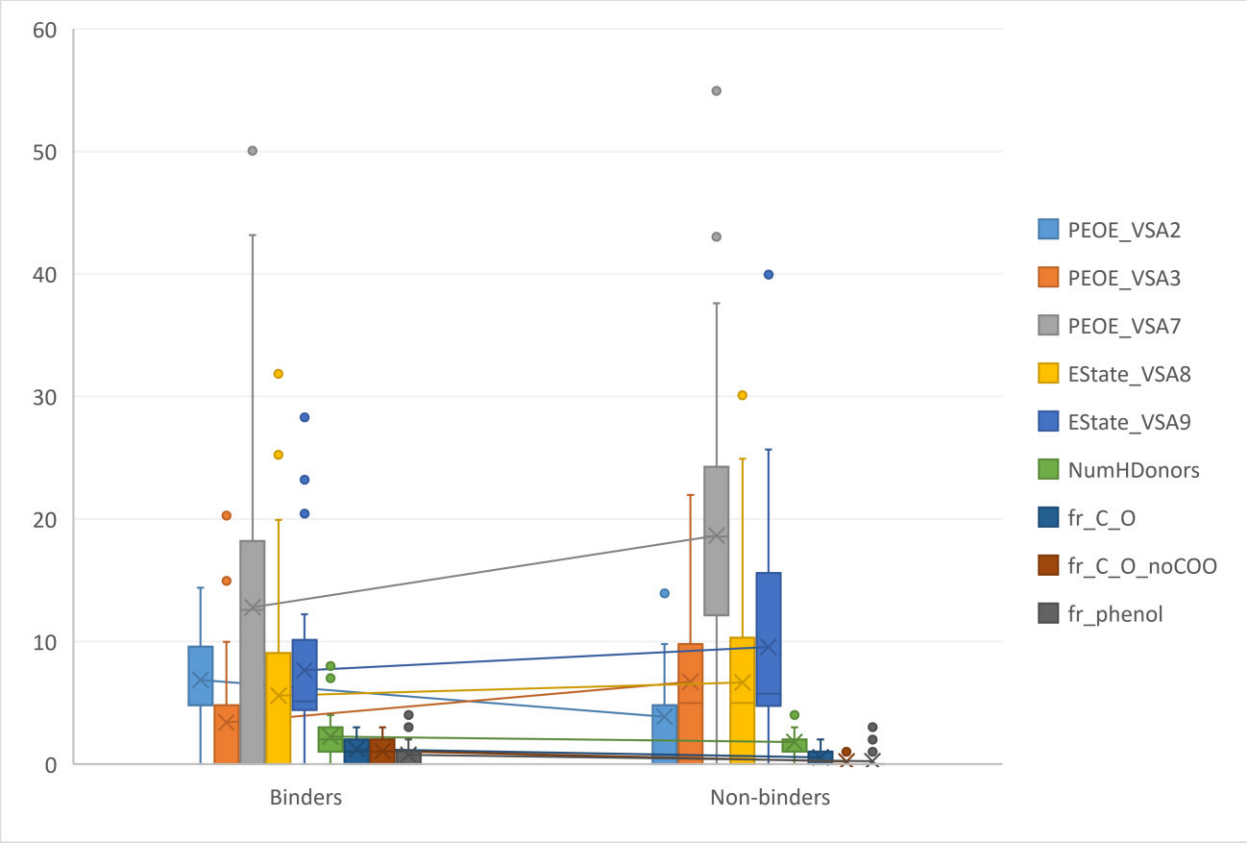


Figure 5. The average value and distribution of the most discriminative descriptors for the binder or non-binder group.

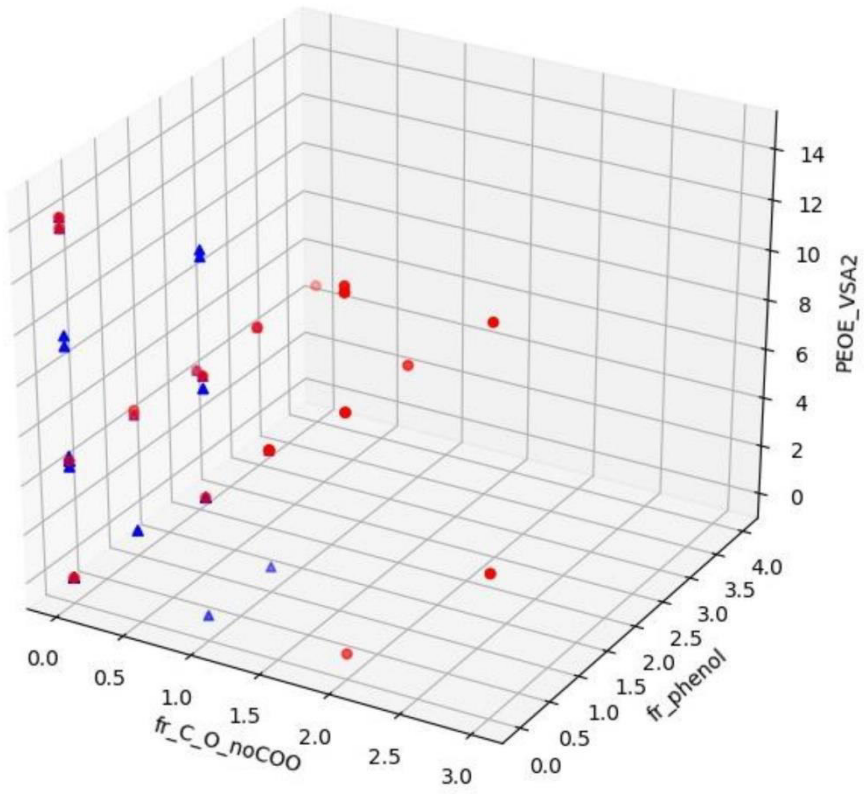
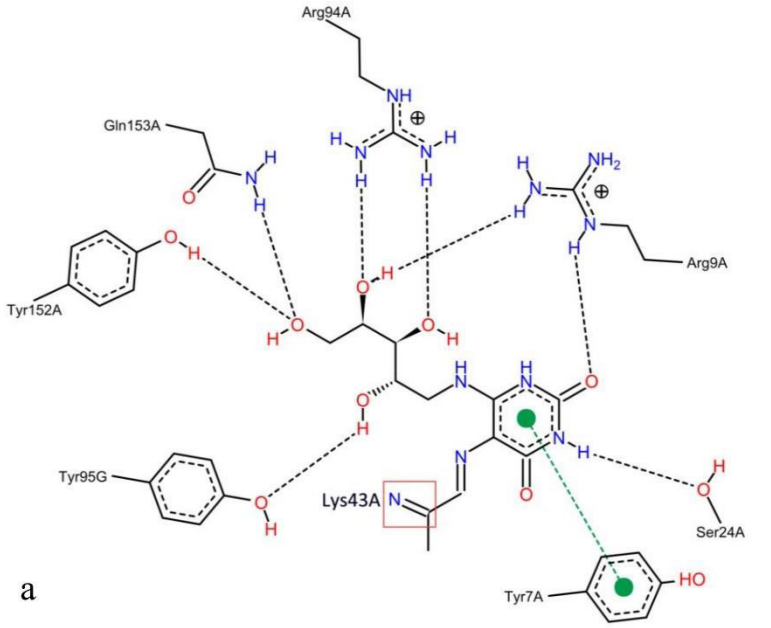
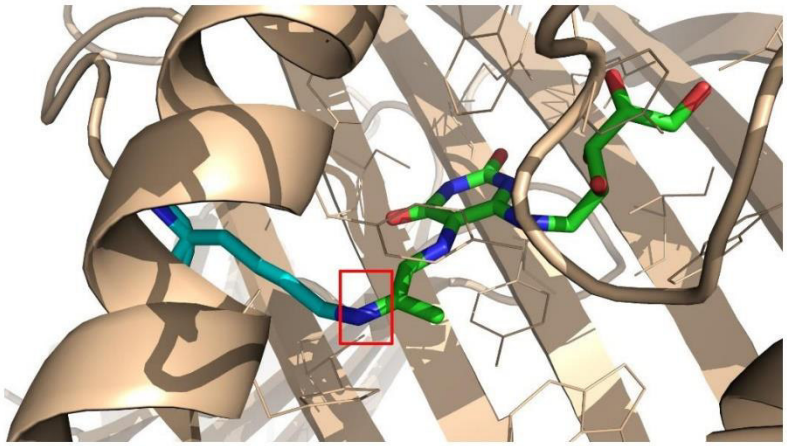


Figure 6. A 3D scatter plot of the three critical descriptors employed by the DT. Binders and non-binders (including train and test set) are shown in red circles and blue triangles, respectively.



a



b

Figure 7. a: The 2D interaction diagram between 5-OP-RU and MR1 (PDB ID: 4nqc). The Schiff base formation after reaction between the ligand carbonyl group and the amino group of the Lys43, which is shown by a red rectangle. The π- π interactions are shown by a green dashed line between rings. A black dashed line shows hydrogen bond. The image was created using PoseView [26] and then edited manually.

b: The 3D interaction diagram between 5-OP-RU and MR1 (PDB ID: 4nqc). The Schiff base formation after reaction between the ligand carbonyl group and the amino group of the Lys43, which is shown by a red rectangle. The image was created using Chimera.