Machine Learning - Based q-RASPR Modeling of Power Conversion Efficiency of Organic Dyes in Dye-Sensitized Solar Cells

Souvik Pore, Arkaprava Banerjee, Kunal Roy*

Drug Theoretics and Chemoinformatics Laboratory,

Department of Pharmaceutical Technology, Jadavpur University,

188 Raja S C Mullick Road, 700032, Kolkata, India

Supplementary Materials SI-1

*Corresponding author

Prof. Kunal Roy, Phone: +91 9831594140; Fax: +91-33-2837-1078;

Email: kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

Machine learning methods

a) Ridge regression: It is a popular method, which is used to address the multicollinearity problem of MLR models without removing any independent variables. In this method, a small amount of bias (penalty) is added to get better predictions. It is an important regularization technique that helps to reduce the complexity of a model, and it is known as Tikhonov regularization. The generalized equation of ridge regression is: $L(x, y) = Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n}(w_i)^2)$ where w_i is weightage of individual feature and λ is penalty term.[42,43]

b) Linear Support Vector Machine (LSVM): LSVM is a form of SVM algorithm, where the data domain can be classified linearly without any kind of transformation. The main steps in LSVM are mapping of data domain into response data, and then division of the data domain. The generalized equation for LSVM is: $\hat{y} = w^T X + b$.[44]

c) Support Vector Machine (SVM): It is a classification machine-learning algorithm, but it can also be used for regression problems, as known as Support Vector Regression (SVR). The main idea behind SVM is to draw a decision boundary between observations to perform its predictions. For nonlinear SVM, we have to transform the data into a feature space (using a kernel function) before mapping with the response, and the generalized equation for SVM (non-linear) is represented as follows: $\hat{y} = w^T \phi(X) + b$, where $\hat{y}$ is predictions, w is the vector of weights, X is a vector of input features, ϕ is a kernel function and b is bias. This decision boundary is known as a plane for a three-dimensional space and known as a hyperplane for higher order space where a large number of features are present. The SVM method considers both margins which are formed by the area between the decision boundary and the closest training compound and the hyperplane for predictions. The margin is represented by the following equation: $margin = \frac{1}{w^T w}$. SVM tries to maximize the distance between the two closest training compounds on either side of decision boundary.[45]

d) Random forest (RF): The RF algorithm builds multiple decision tree models and combines their outcome for more accurate and stable prediction. This method helps to overcome overfitting problem of a decision tree models. The RF algorithm is based on an ensemble learning method known as Bagging (bootstrap aggregating) which is a resampling technique applied to a dataset. In bootstrapping, observations are selected by random sampling with replacement, and random feature subsets are selected. In bagging, a large number of datasets

are created by bootstrapping the original dataset, multiple decision tree models are formed using these datasets, and finally average of predictions are taken.[45]

e) Gradient boosting (GB): Boosting is also an ensemble method that helps to form a strong learner by combining many weak learners. It is also a tree-based method in which decision trees are generated sequentially, and every tree tries to correct its predecessor. Gradient boosting (GB) is one of the boosting methods in which it tries to fit the current predictor with residual error made by the previous predictor.[42]

f) XGBoost: It stands for Extreme Gradient Boosting and was first implemented by the researchers of the University of Washington. This method was built based on the same algorithm of GB, but the main drawback of GB is that it searches for minimizing the loss function across all possible splits to create a new branch of a decision tree. Thus, GB method becomes time-consuming when thousands of features are present because there are thousands of possibilities to split the node of a decision tree. XGBoost handles these drawbacks by taking information of feature distribution across all data points in a single leaf node and by this way it reduces the search space. This method cannot generate multiple decision trees in parallel but can generate multiple branches of a decision tree in parallel.[46]

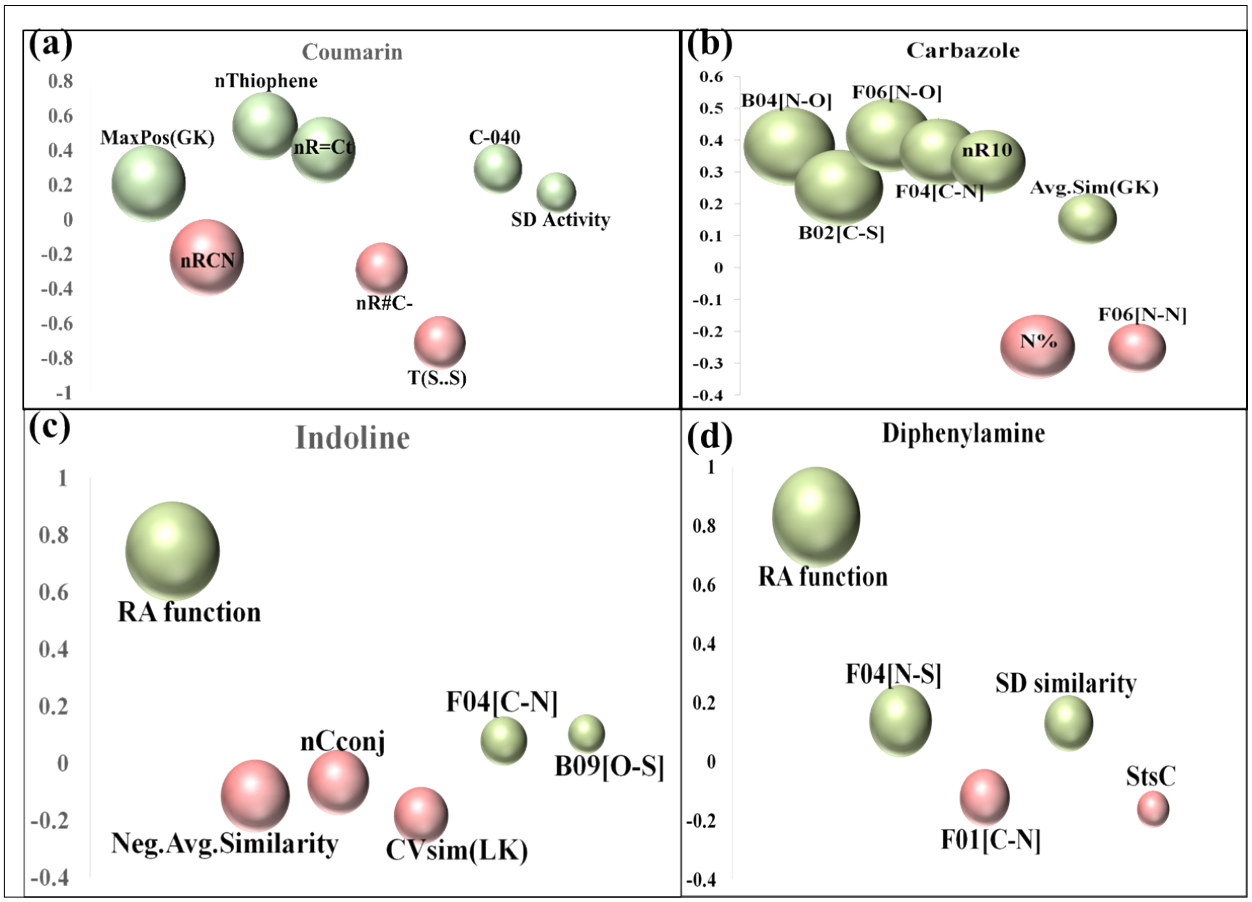


Figure S1. Bubble plot showing regression coefficients for individual descriptors in the PLS model (size of the bubble shows variable importance score)

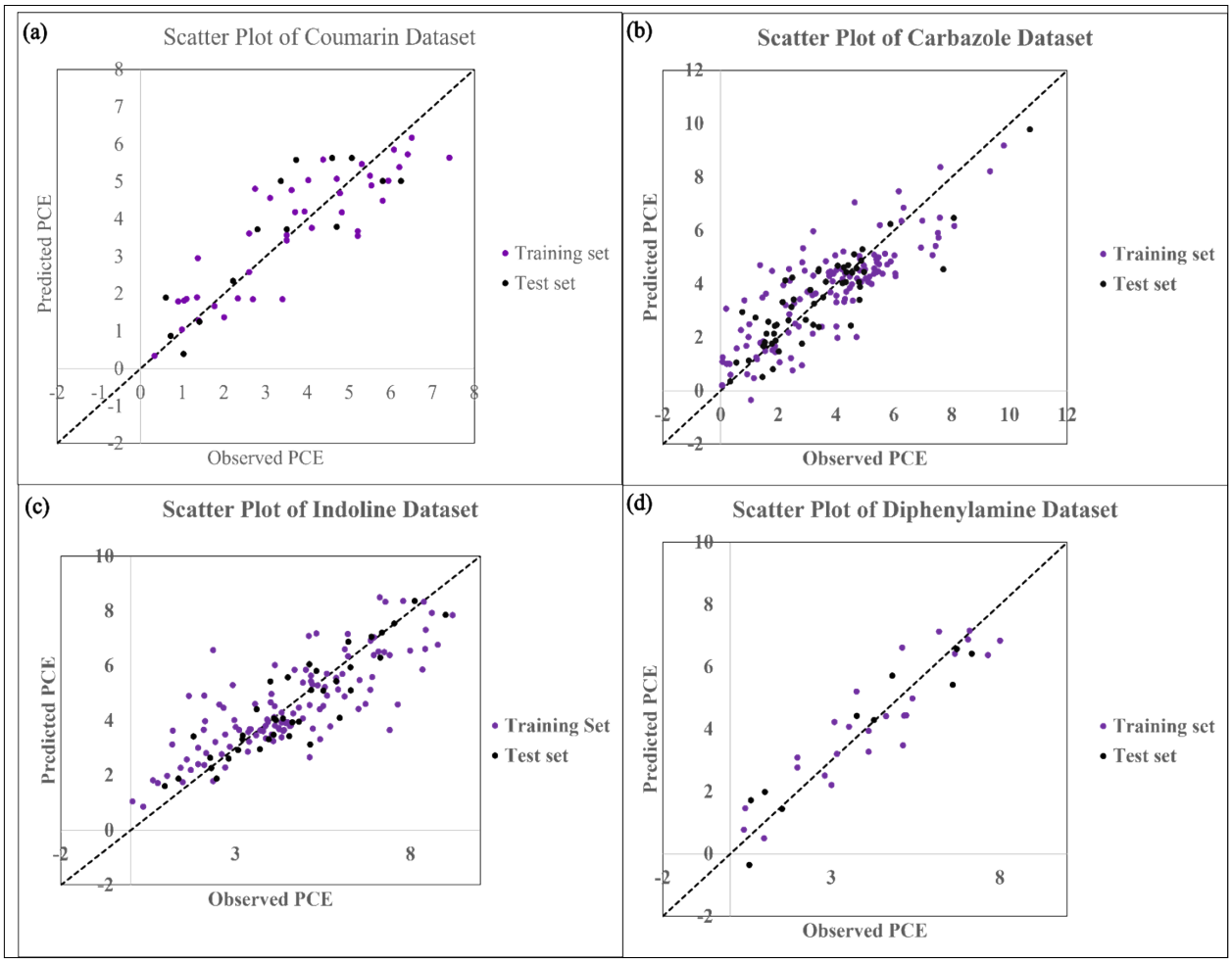


Figure S2. Scatter Plots indicating the prediction quality of the developed q-RASPR PLS models

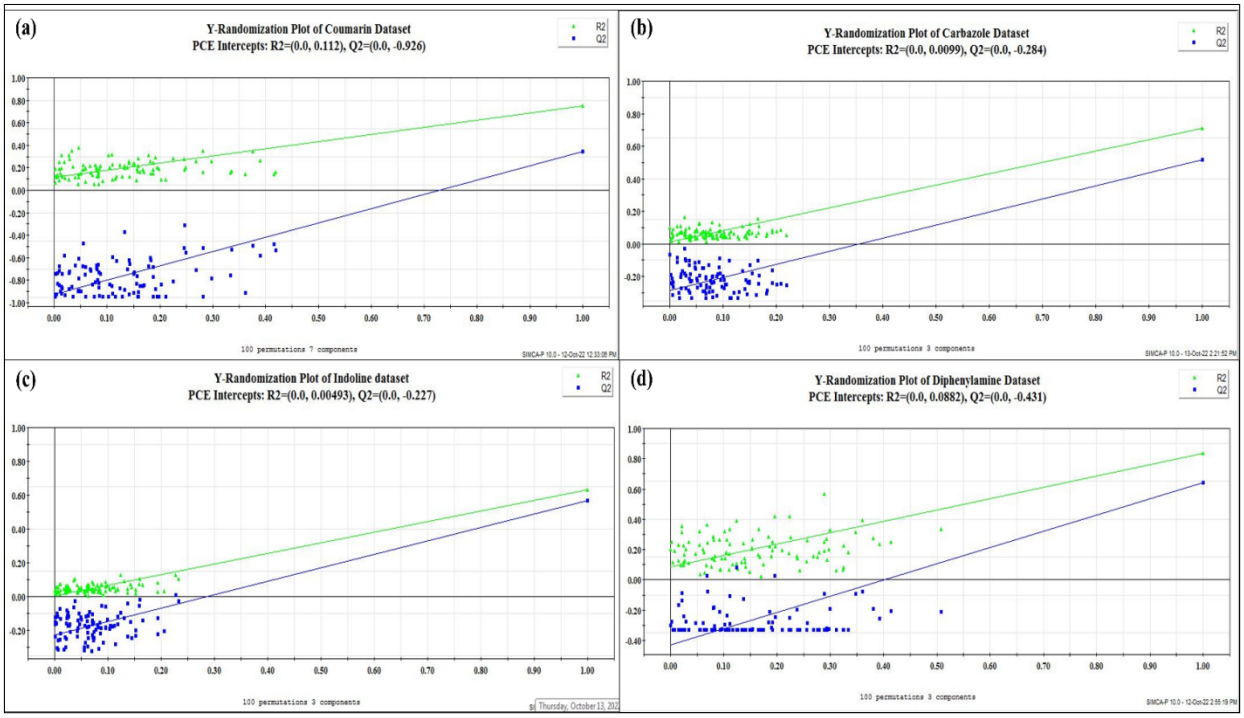


Figure S3. PLS randomization plots

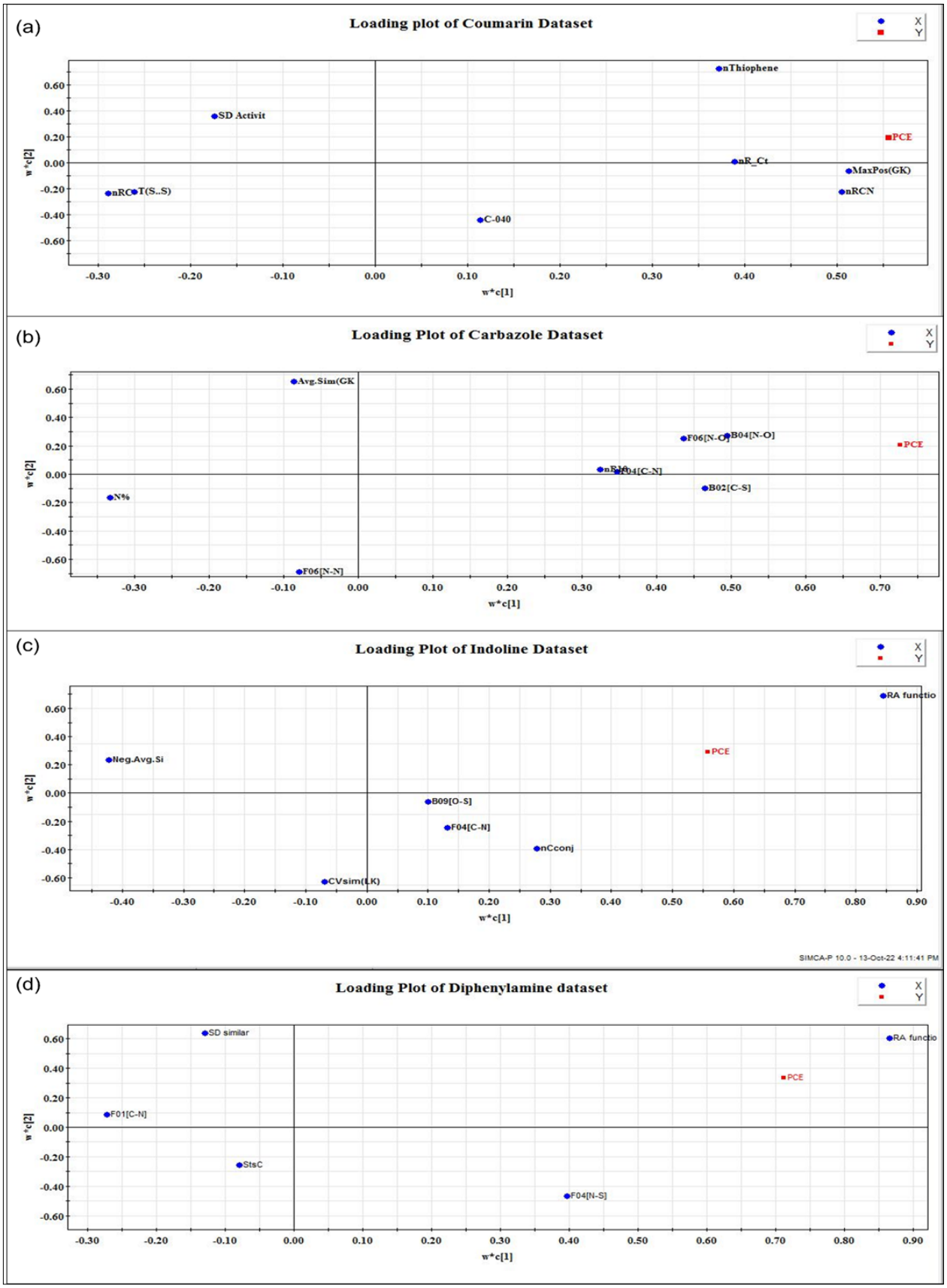


Figure S4. Loading Plots showing significance of the descriptors to PCE

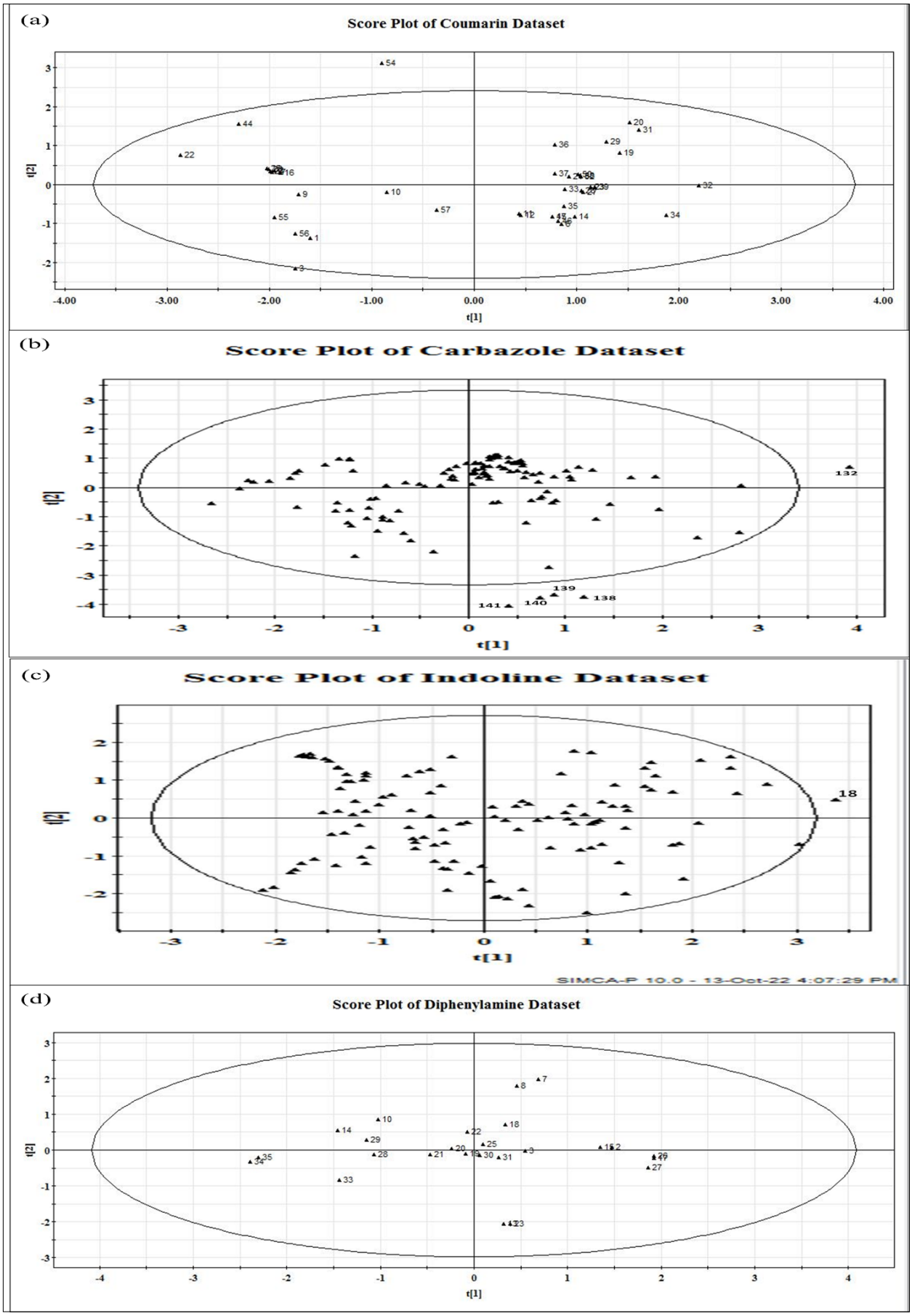


Figure S5. Score Plots indicating the applicability domain of PLS models

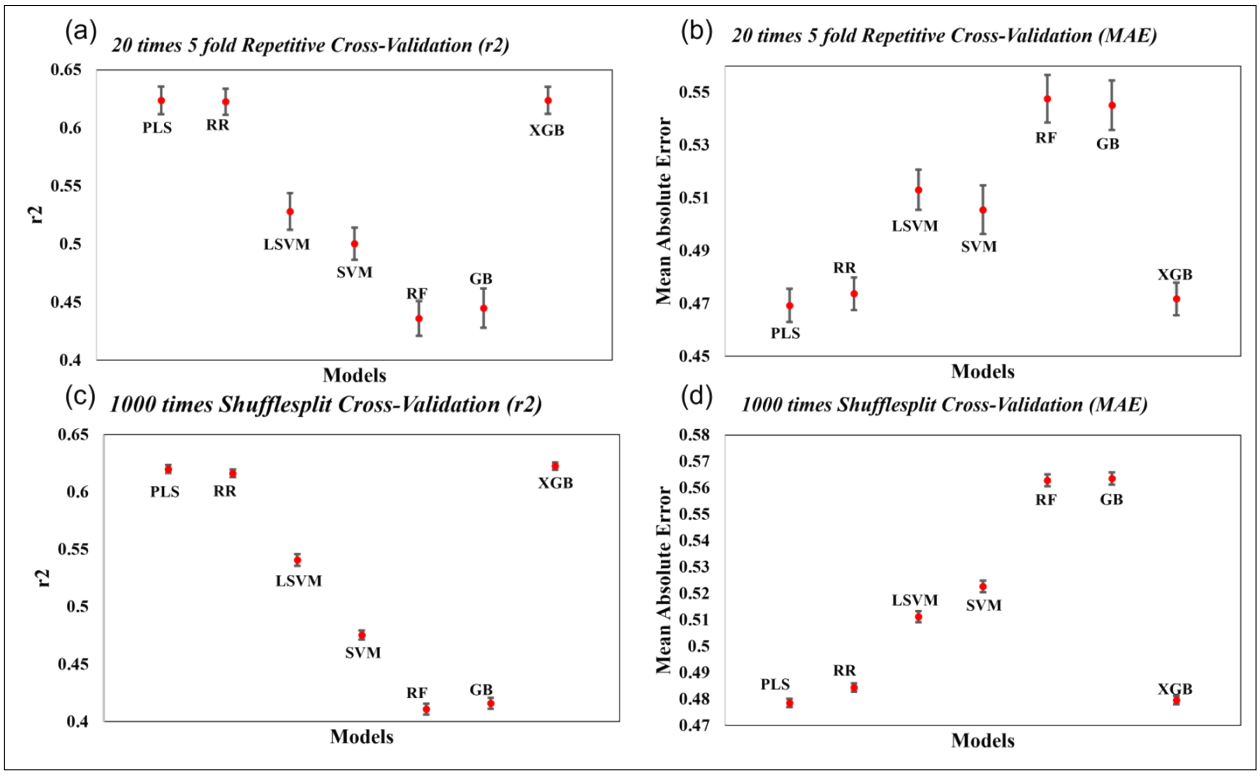


Figure S6. Cross-validation statistics based on 20 times 5-fold repetitive CV and 1000 shuffle split CV method (Mean ± SEM) for the carbazole dataset

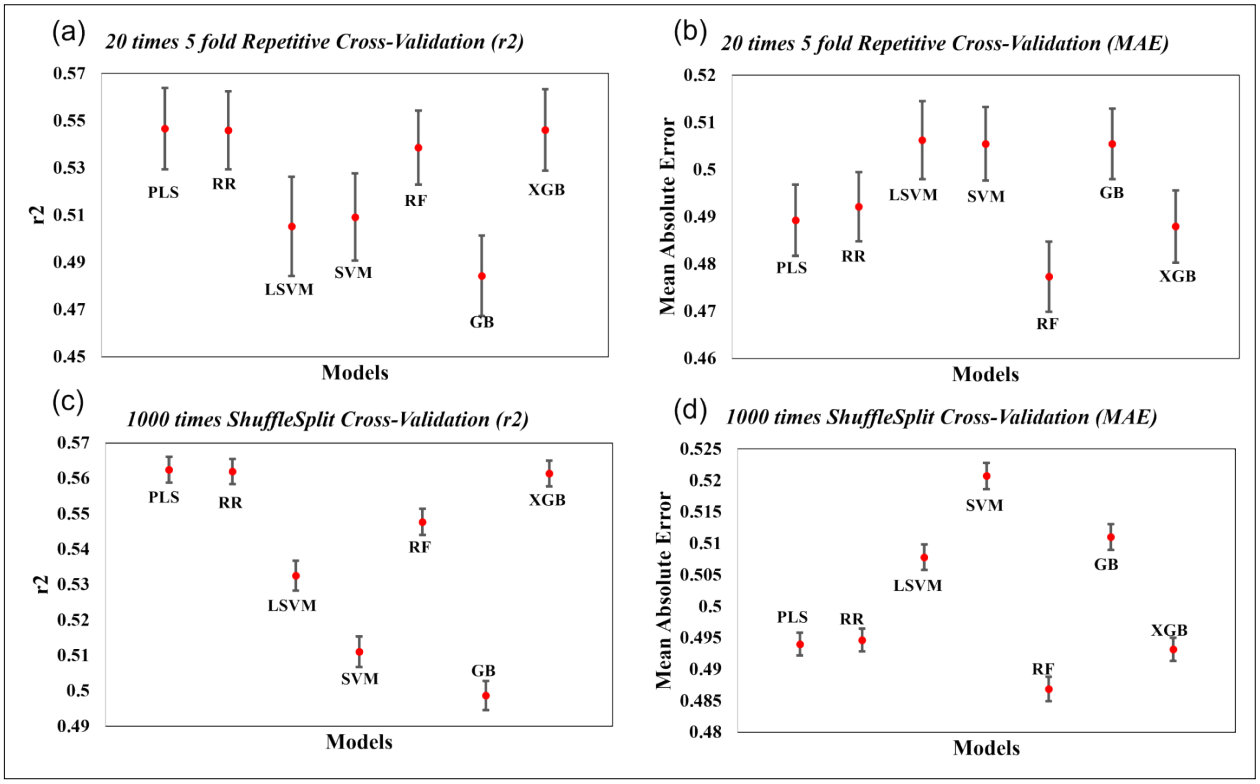


Figure S7. Cross-validation statistics based on 20 times 5-fold repetitive CV and 1000 shuffle split CV method (Mean ± SEM) for the indoline dataset

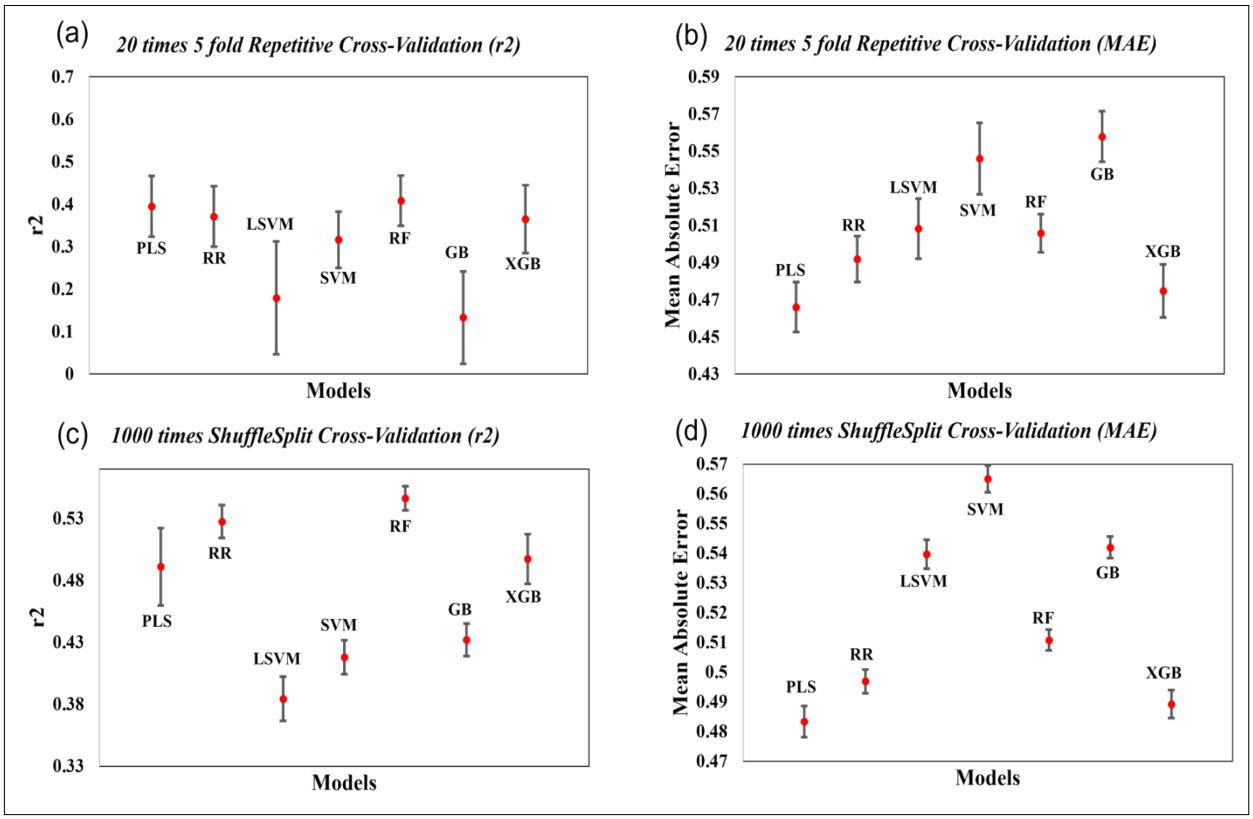


Figure S8. Cross-validation statistics based on 20 times 5-fold repetitive CV and 1000 shuffle split CV method (Mean ± SEM) for the diphenylamine dataset

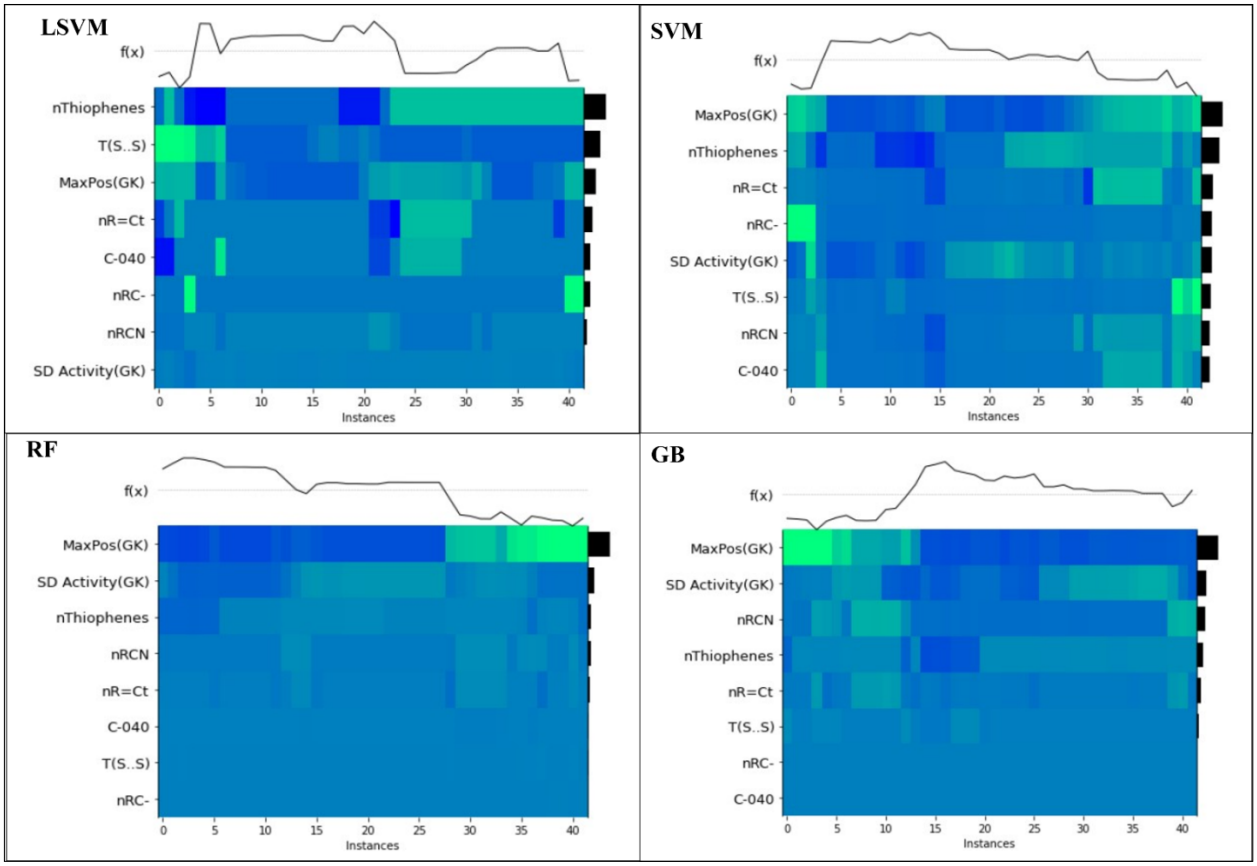


Figure S9. Heatmap plots of LSVM, SVM, RF and GB models for the coumarin dataset, indicating the relative importance of descriptors

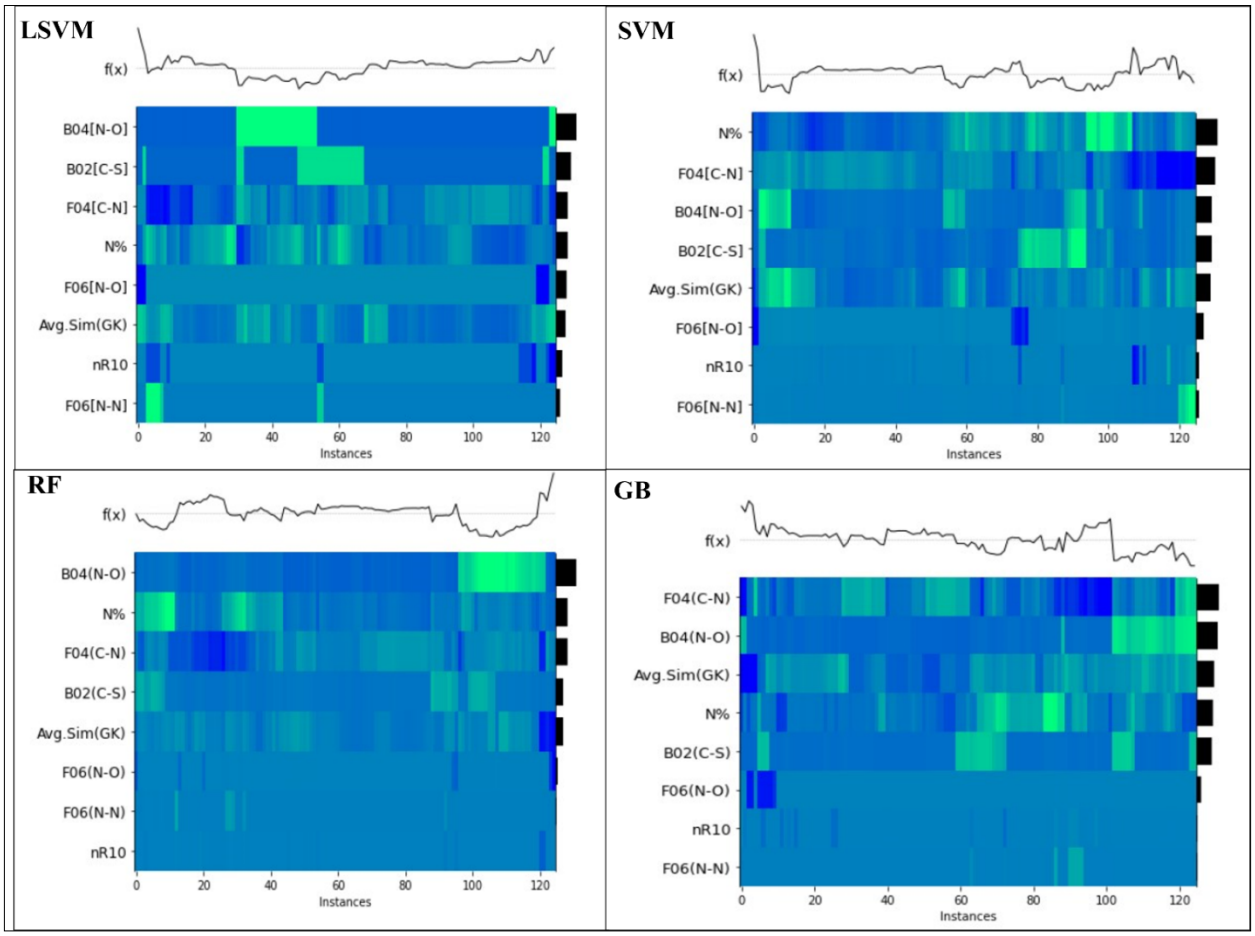


Figure S10. Heatmap plots of LSVM, SVM, RF and GB models for the carbazole dataset, indicating the relative importance of descriptors

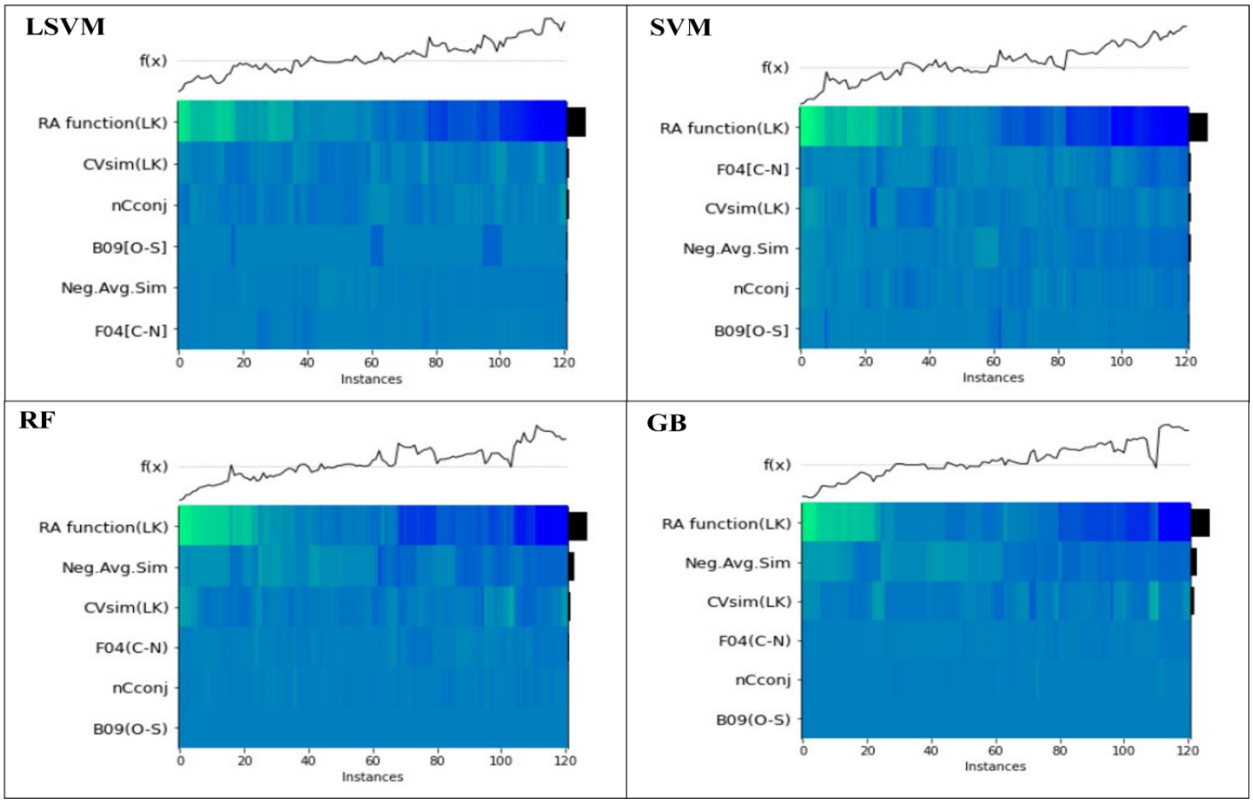


Figure S11. Heatmap plots of LSVM, SVM, RF and GB models for the indoline dataset, indicating the relative importance of descriptors

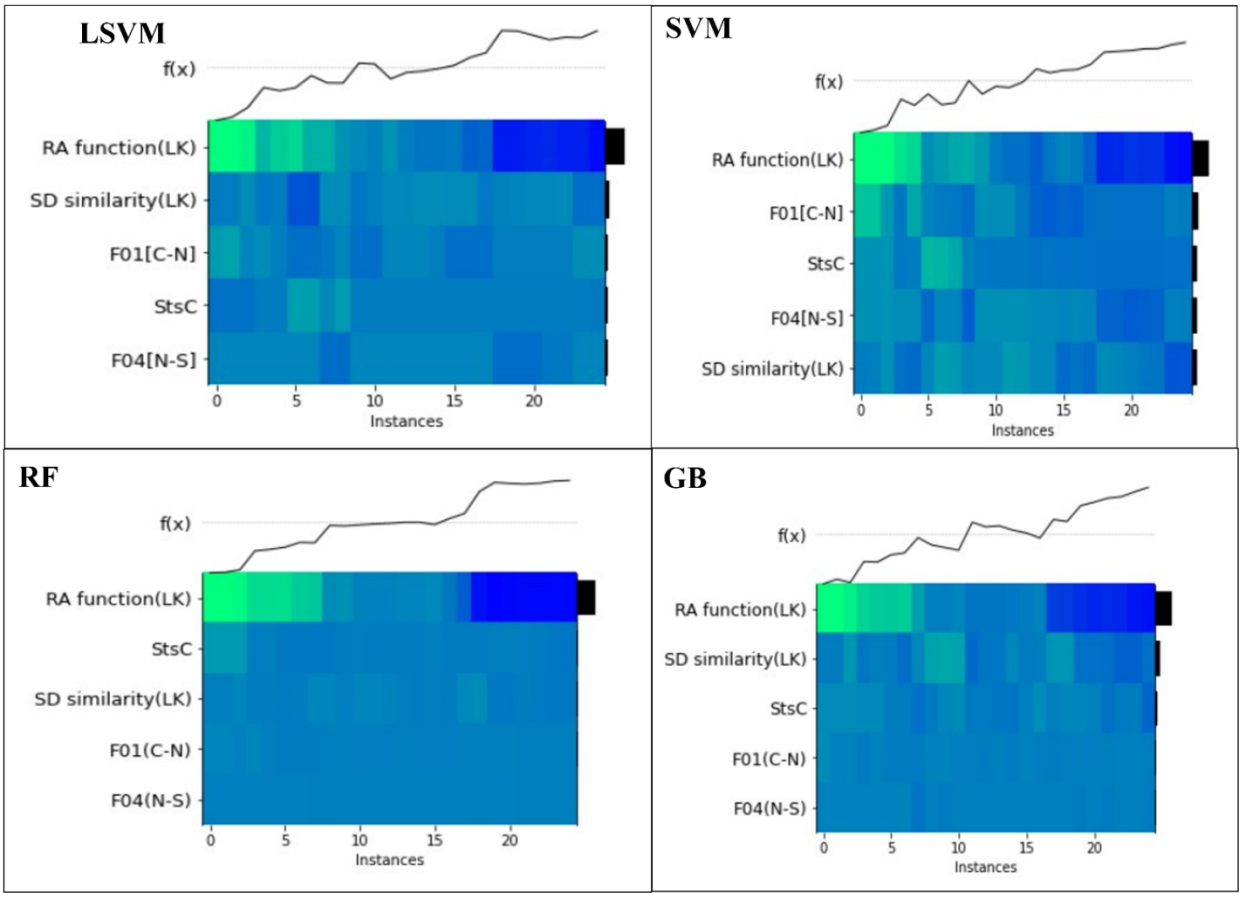
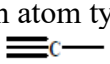


Figure S12. Heatmap plots of LSVM, SVM, RF and GB models for the diphenylamine dataset, indicating the relative importance of descriptors

Table S1. Initial pool of descriptors used for the read-across analysis

Dataset	Descriptors
Coumarin	nRCN : number of aliphatic nitriles in the structure F08[N-S] : Frequency of N and S atoms at topological distance 8 B09[S-S] : Presence or absence of S and S atoms at topological distance 9 nCconj : number of non-aromatic conjugated C(sp2) atoms in the structure nArNR2 : number of aromatic tertiary amines in the structure B08[N-S] : Presence or absence of N and S atoms at topological distance 8 C-034 : it is an atom centered fragment based descriptor which indicate the fragment - R–CR..X (R=any group linked through carbon atom, X=electronegative atoms like N, S, P, O, Halogens) nThiophenes : indicate the number of Thiophene rings in the structure nR#C- : number of a non-terminal carbon atom with the 'sp' hybridization nR=Ct : number of an aliphatic tertiary carbon atom with the 'sp2' hybridization T(S..S) : a 2D atom pair descriptor that indicates the sum of the topological distance between two sulfur atoms where they are part of two thiophene rings C-040 : is an atom-centered fragments descriptor that represents fragments like R–C(=X)-X/R–C#X/X = C = X (R: any group linked through carbon; X: any electronegative atom like N, S, P, O, halogen; #: triple bond)
Carbazole	F08[O-O] : Frequency of O and O atoms at a topological distance of 8 NaasC : representing the Number of atoms of aasC **(-C(-)-)** F06[N-N] : representing the Frequency of N and N atoms at topological distance 6 F06[C-C] : representing the Frequency of C and C atoms at topological distance 6 nR10 : representing the number of 10-membered rings in the structure F04[C-N] : representing the Frequency of C and N atoms at topological distance 4 B08[O-S] : representing the Presence or absence of O and S atoms at a topological distance 8 B04[N-O] : representing the Presence or absence of N and O atoms at topological distance 4 N% : total percentage of N atoms in the structure F06[N-O] : representing the Frequency of N and O atoms at topological distance 6 B02[C-S] : representing the Presence or absence of C and S atoms at topological distance 2 B10[C-S] : representing the Presence or absence of C and S atoms at topological distance 10 B06[N-S] : representing the Presence or absence of N and S atoms at topological distance 6 F06[O-S] : representing the Frequency of O and S atoms at topological

	distance 6 B04[O-S] : representing the Presence or absence of O and S atoms at topological distance 4
Indoline	SaaaC : it is an atom type E-state indices indicate the Sum of aaaC E-states (aaCa where a is aromatic bond) B07[N-N] : representing the Presence or absence of N and N atoms at topological distance 7 B06[N-N] : representing the Presence or absence of N and N atoms at topological distance 6 B04[S-S] : representing the Presence or absence of S and S atoms at topological distance 4 F04[C-N] : representing the Frequency of C and N atoms at topological distance 4 F07[N-S] : representing the Frequency of N and S atoms at topological distance 7 nCrq : number of ring quaternary C(sp3) atoms in the structure F10[C-N] : representing the Frequency of C and N atoms at topological distance 10 F07[N-O] : representing the Frequency of N and O atoms at topological distance 7 NsssN : it is an atom type E-state descriptors indicate the Number of atoms of type sssN () B05[O-S] : representing the Presence or absence of O and S atoms at topological distance 5 B09[O-S] : representing the Presence or absence of O and S atoms at topological distance 9 B05[S-S] : representing the Presence or absence of S and S atoms at topological distance 5 F04[S-S] : representing the Frequency of S and S atoms at topological distance 4 F07[N-N] : representing the Frequency of N and N atoms at topological distance 7 nCconj : representing the number of non-aromatic conjugated C (sp2) atoms in the structure F10[O-S] : representing the Frequency of O and S atoms at topological distance 10 B02[N-O] : representing the Presence or absence of N and O atoms at topological distance 2
Diphenylamine	F01[C-N] : representing the Frequency of C and N atoms at topological distance 1 F08[C-N] : representing the Frequency of C and N atoms at topological distance 8 StsC : It is an atom type E-state descriptor that indicates the sum of tsC E-states () nPyrimidines : It is functional group count descriptor indicate the number of Pyrimidines in the structure nCsp : It is a constitutional descriptor indicate the number of sp hybridized carbon atoms in the structure B08[N-N]: representing the Presence or absence of N and N atoms at topological distance 8 C-041 : It is an atom centered fragment corresponds to X–C(=X)-X,

	where, X can be any electro negative atom O, N, S, P, Se and halogens connected with the carbon atom nHAcc : It is a functional group count descriptor indicate the number of acceptor atoms for H-bonds (N,O,F) nR#C- : nR#C- : number of a non-terminal carbon atom with the 'sp' hybridization F04[N-S] : representing the Frequency of N and S atoms at topological distance 4 ETA_dBeta : It is an extended topochemical atom descriptor, measuring the relative unsaturation content (Δβ)

Table S2. List of read-across derived descriptors and their definition

Descriptors	Description	Mathematical equation
RA function	It is a read-across weighted average prediction score for a target compound which is obtained based on the similarity between selected close source compound and target (or query) compound.	$$RA\ function = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$ $$w_i = \frac{S_i}{\sum_{i=1}^{n} S_i}$$ w_i = weightage given to individual selected close source compounds S_i= Similarity between individual selected close source compounds and target compounds x_i=observed response value of the selected close source compounds
SD Activity ($S_{weighted}$)	It is the standard deviation of the observed response value of selected close source compounds for each query compounds.	$$S_{weighted} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x}_{wtd})^2}{\sum_{i=1}^{n} w_i}} \times \frac{n}{n-1}$$ $$x_{wtd} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$ n = effective number of selected close source compound w_i = weightage given to individual selected close source compounds

SE	It is the standard error of the observed response values of selected close source compounds for each query compounds.	$SE = \frac{SD\ Activity}{\sqrt{n}}$
CVact ($CV_{activity}$)	It is the coefficient of variance of the observed response (or activity) values of the selected close source compounds for each query compounds.	$CV_{activity} = \frac{S_{weighted}}{\bar{x}_{wtd}}$
CVsim ($CV_{similarity}$)	It is the coefficient of variance of the similarity values of the selected close source compounds for each query compounds.	$CV_{similarity} = \frac{SD\ Similarity}{\bar{f}}$ $\bar{f}$= average similarity value of the selected close source compounds
MaxPos	Maximum similarity level to Positive close source compounds based on mean value of training set observed response.	
MaxNeg	Maximum similarity level to Negative close source compounds based on mean value of training set observed response.	
Abs MaxPos-MaxNeg (Abs.Diff.)	Absolute difference between *MaxPos* and *MaxNeg*.	$AbsDiff = \lvert MaxPos - MaxNeg \rvert$
Avg.Sim. (Average Similarity)	It is the mean of similarity values of the selected close source compounds for each query compounds.	$Avg.Sim.(\bar{f}) = \frac{\sum_{i=1}^{n} f_i}{n}$ f_i = similarity values of the selected close source compounds n=number of selected close source compounds
SD Similarity	It is the standard deviation of the similarity values of selected close source compounds for each query compounds.	$SD\ Similarity = \sqrt{\frac{\sum_{i=1}^{n}(f - \bar{f})^2}{n-1}}$ $\bar{f}$= average similarity value of the selected close source compounds n=number of selected close source compounds
g_m [Banerjee-Roy Coefficient]	A novel concordance measure	$g_m = (-1)^n \lvert Posfrac - 0.5 \rvert$ n=1 when *MaxPos<MaxNeg* n=2 when *MaxPos>=MaxNeg* Posfrac=Fraction of the close source compounds having a response value greater than the training set mean

		response
gm*Avg.Sim	Product of the values of g_m and *Average Similarity*	
gm*SD Similarity	Product of the values of g_m and *SD Similarity*	
Pos.Avg.Sim	Average similarity value of the selected positive close source compounds based on the mean observed response value of training set	
Neg.Avg.Sim	Average similarity value of the selected negative close source compounds based on the mean observed response value of training set	