

Supplementary Materials

**Development of classification and regression models for *Vibrio fischeri*
toxicity of ionic liquids: Green solvents for the future**

Rudra Narayan Das and Kunal Roy*

Drug Theoretics and Cheminformatics Laboratory,

Division of Medicinal and Pharmaceutical Chemistry,

Department of Pharmaceutical Technology,

Jadavpur University, Kolkata 700 032 (INDIA)

Email: kunalroy_in@yahoo.com; kroy@pharma.jdvu.ac.in

Phone: +91 98315 94140; Fax: +91-33-2837-1078;

URL: <http://sites.google.com/site/kunalroyindia/>

Supplementary Materials

Supplementary Materials

Model development

For classification analysis, compounds are divided into groups or classes by assuming one or more reference value; for a two group classification, this will be one value. Considering the importance of ionic liquids as potential solvents for a wide variety of usage, we have considered toluene as the reference toxic substance for the determination of *a priori* toxicity class. Toluene is known to be a potential toxic solvent with a pEC₅₀ value of 3.670 mol/L to *Vibrio fischeri*.³⁹ Hence compounds showing pEC₅₀ value greater than 3.670 mol/L have been classified as toxic or positive (P) and less than that are termed as non-toxic or negative (N) substances.

Selection of the training and test sets

Proper selection of training and test sets in QSTR analysis is very crucial for the determination of predictive features of compounds.⁴⁰ In the present study, we have used *k*-means cluster analysis⁴¹ for the division of the dataset into an internal validation set or training set and an external validation set or test set. In *k*-means clustering technique, *k*-means clusters are generated based on the structural similarity of input compounds following a non-hierarchical algorithm such that both the training and test sets are represented by each cluster. We have used standardized matrix of the calculated descriptors for *k*-means cluster division. Five clusters were developed with the 147 compounds from which a training set of 110 (approximately 75%) compounds and test

Supplementary Materials

set of 37 (approximately 25%) compounds were extracted. The serial numbers of the test set compounds present in each clusters has been presented in **Table S4**. The whole dataset has been used for classification analysis whereas some compounds were omitted for carrying out quantitative regression analysis (*vide infra*). Along with the complete compounds consisting of both cations and anions, the dataset also contained singular ionic species as well as non-numeric *i.e.* qualitative toxicity values which have been considered in the classification analysis.

For the development of mathematical regression model, we have considered only the completely defined compounds with numeric toxicity values. It was observed that, the number of only anionic species was 6, the number of compounds showing qualitative toxicity value (*e.g.* pEC_{50} (mol/L) < 1.700) was 17, and 2 anionic species showed the occurrence of qualitative toxicity values. Hence (17+6-2) or 21 compounds have not been considered among the 147 compounds such that the final set for regression model comprises of 126 compounds. After the division of the whole dataset (147 compounds), 20 compounds from training set and 1 compound from test set were removed accordingly. **Table S2** shows the compounds not included in regression analysis.

Descriptor thinning

In the present study, we have calculated 94 indicator variables corresponding to 30 anionic and 64 cationic species. Apart from that, 334 various two-dimensional descriptors were calculated separately for anions and cations using the Dragon software (version 6.0)³⁸ and were assigned to complete compound giving a matrix of 668 variables. Then

Supplementary Materials

variable reduction on 94 indicator parameters was performed by carrying out stepwise multiple linear regression (MLR) method and 33 selected variables were considered for modeling purposes. Among the 668 descriptors, those with undetermined values and constant variances were removed and finally 513 variables were chosen for further work. Further reduction of the number of Dragon descriptors was done during model development, where descriptors were initially chosen separately from each sub-group (*e.g.* constitutional, topological, connectivity, ETA etc.) and then the selected potential descriptors from all such groups were used for building model. Stepwise MLR⁴² and genetic function approximation (GFA)⁴³⁻⁴⁵ analysis were used for selection of variables from each subset of two-dimensional Dragon descriptors.

In the classification analysis, we have used the selected 33 indicator parameters and the subset descriptor pool of various two-dimensional descriptors derived using stepwise MLR technique (as discussed above). Our approach was to consider only those descriptors with potential discriminatory capability. This descriptor pool was scaled so that the indicator variables (0 or 1) become comparable in numerical values with various other two-dimensional descriptors. The performed standardization technique has been described in the next section. In order to judge the best discriminatory property of the variables, we have finally constructed a contribution molecular spectrum plot of the variables of the best model, belonging to the “positive” and “negative” groups. For the regression analysis, we considered only the calculated raw two-dimensional descriptor pool (without performing any standardization) for model development.

Supplementary Materials

Standardization of the variables for classification analysis

Standardization of the variables was performed using the following equation.

$$x' = \frac{x - \overline{x_{training}}}{\sigma_{training}} \quad (S1)$$

where x and x' are the original and standardized (scaled) values of a descriptor respectively, $\overline{x_{training}}$ and $\sigma_{training}$ are the respective mean and standard deviation values of that particular descriptor in the training set. Here, the mean and standard deviation values of the training set has been used to keep uniformity in standardizing the test set variables, so that their individual values remain unaffected when their size is varied.

Performed statistical analyses

We have used linear discriminant analysis (LDA) as a supervised classification tool³² to derive a well predictive toxicity classification model for ionic liquids. A linear discriminant function was developed using the training set which was used to calculate discriminant scores of individual compounds. The calculated and predicted classification categories of the training and test sets respectively were determined at a 50% probability level. In this study, we have performed stepwise selection as a descriptor reduction tool with the F value criteria (F to enter 4.0 and F to remove 3.9) and tolerance value of 0.001, followed by discriminant analysis. The details of the linear discriminant analysis have been documented in next section.

Supplementary Materials

For the regression model, multiple linear regression (MLR)⁴² and partial least squares (PLS)⁴⁶ techniques have been used in combination with the chemometric tools Stepwise Regression and Genetic Function Approximation (GFA)⁴³⁻⁴⁵ for descriptor selection. The PLS method is capable of operating with intercorrelated descriptor pool⁴⁶. In case of the stepwise method⁴², a multiple term equation was first built using the user defined objective function F to enter 4.0, F to remove 3.9. Then, in order to avoid the intercorrelation problem among variables, partial least squares analysis was performed on the selected variables obtained from stepwise regression. The GFA equations were derived at 5000 iterations with no fixed length option for the final equation using lack-of-fit (LOF) as the objective function. The details of the chemometric tools used for model development have been described below.

Details of the chemometric tools

Classification using linear discriminant analysis (LDA)

Classification analysis leads to evaluate the status of belonging of a group of categorical objects. In initial assumption, group of objects are forcefully categorized into two or more categories with respect to a reference. Then linear discriminant analysis (LDA) is performed to judge the appropriateness of the classification of objects following a maximum likelihood method, *i.e.* the determination of the highest likelihood of a member object to be grouped to a particular class. It exhaustively subdivides a p -dimensional

Supplementary Materials

vector space of the independent variables into mutual exclusive regions or groups such that an object falling into a region r will be assigned to group r .³²

LDA reduces dimensionality of data matrix and noisy variable as well to yield lower dimension relevant data keeping the discrimination objective for which it is deployed. It is a supervised method that tries to determine the direction that will aid to the best separation of classes. The objective function considered in the algorithm of LDA is to maximize the ratio of scatter between within-class and between-class observations.

For an observation consisting of C classes (or categories), if i be a class among C comprising of N_i number of samples with a mean vector μ_i , the within-class and between-class scatter matrix can be defined by the following **Equations**:

$$\text{Within-class scatter matrix: } S_w = \sum_{i=1}^C \sum_{j=1}^{N_i} (y_j - \mu_i)(y_j - \mu_i)^T \quad (\text{S2})$$

$$\text{Between-class scatter matrix: } S_b = \sum_{i=1}^C (\mu_i - \mu)(\mu_i - \mu)^T \quad (\text{S3})$$

where, N is the total number of samples of C classes, $N = \sum_{i=0}^C N_i$

i is an arbitrary class among C , $i=1, 2, 3 \dots C$.

y is the sample of interest,

μ is the mean of entire set of samples such that, $\mu = \frac{1}{C} \sum_{i=1}^C \mu_i$

and the superscript ' T ' corresponds to the transpose of matrix.

Supplementary Materials

Then LDA tries to calculate a transformation that will maximize the inter-class (between) variance and minimize the intra-class (within) variance or scatter, *i.e.* maximization of the objective function. Let us consider a mean-centered descriptor matrix X of dimension $m \times p$, and let a be a vector of p dimension. If vector a causes maximization of the ratio, then the function may be written as:

$$J = \frac{a^T S_b a}{a^T S_w a} \quad (\text{S4})$$

This transformation maximizes the separability among classes while reduces the variations within particular groups. Then the linear discriminant function can be defined as below:

$$X_a = x_1 a_1 + x_2 a_2 + x_3 a_3 + \dots + x_p a_p \quad (\text{S5})$$

Here, vector a is defined by the eigenvector of the matrix $S_w^{-1} S_b$ corresponding to the largest eigenvalue²⁹. For classification problems encountering two groups, the equation for discriminant function (**Eq. S5**) reduces to **Eq. S6**.

$$a = S_w^{-1} (\bar{x}_1 - \bar{x}_2) \quad (\text{S6})$$

Supplementary Materials

where \bar{x}_1 and \bar{x}_2 are the two column vectors representing averages of variables belonging to groups 1 and 2. Any unknown sample x_i will be classified into region (or group) h when the mean score $a^T \bar{x}_h$ will be at the smallest distance to $a^T x_i$.³³

The coefficient vector of the discriminant function is orthogonal to its line of likelihood ratio 1 and thus parallel to the gradient of the flat log likelihood ratio surface. Now, for the maximization of the ratio of between to within group scatter matrix, let us consider vector a is given by the first eigenvector λ_1 of $(S_w^{-1}S_b)$ scaled by the scalar quantity $(a^T S_w a)^{-1/2}$. If d represent $(S_w^{1/2}a)$, $S_w^{1/2}$ being the symmetric square root of S_w . Then, the maximum of the numerator $(a^T S_b a)$ may be expressed as:

$$\max d^T S_w^{-1/2} S_b S_w^{-1/2} d$$

$$\text{corresponding to } d^T d = a^T S_w a = 1 \quad (S7)$$

The first eigenvector γ_1 of $(S_w^{-1/2} S_b S_w^{-1/2})$ is used for solving classical eigenvalue problems. Here, both the term $(S_w^{-1} S_b)$ and $(S_w^{-1/2} S_b S_w^{-1/2})$ possess the same eigenvalue, and $a = S_w^{-1/2} \gamma_1$ will be the scaled first eigenvector of $(S_w^{-1} S_b)$.²⁹ During the transformation of the square root matrix $S_w^{-1/2}$, the within-group covariance matrix of the transformed variables $X S_w^{-1/2}$ will be equal to the identity matrix, *i.e.*, signifying same variance of the groups in every direction, if all the classes possess same covariance matrix. In this referred discrimination space or transformed space, euclidean distances and data points become equal to Mahalanobis distances. In a multivariate normal distribution with

Supplementary Materials

covariance matrix Σ , the Mahalanobis distance between any two data points x_i and x_j can be defined as:⁵⁹

$$d_M(x_i, x_j) = \sqrt{\frac{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}{\sqrt{\quad}}} \quad (\text{S8})$$

Square of Mahalanobis distance is a measured parameter during LDA analysis for the determination of likelihood of a compound to be classified in a specific group in the discriminant space.

Stepwise MLR

The stepwise regression⁴² method is aimed at building a multiple-term linear equation step-by-step maintaining a stepping criterion. The basic procedures used for this method are:

- i) Identification of an initial model;
- ii) Iterative ‘stepping’, i.e. alteration of the model of the previous step repeatedly by addition or removal of a predictor variable maintaining the objective function *i.e.* the ‘stepping criteria’ (in our case, $F = 4.0$ for inclusion; $F = 3.9$ for exclusion for the forward and backward selection method); and
- iii) Termination of the search when stepping is no longer possible as per the given stepping criteria, or when a specified maximum number of steps has been reached.

Supplementary Materials

Specifically, all the variables are reviewed at each step and evaluated for the determination of the variable that will have most contribution to the equation; that variable is then included in the model, and the process starts again.

Partial least squares (PLS)

The partial least squares (PLS) appear to be a generalization of regression, which is able to handle data with strongly correlated and/or noisy or numerous independent variables. It builds a statistically more robust and reduced solution than multiple linear regression (MLR) method. A PLS model identifies new variables termed the ‘latent variables or independent scores’ which are linear combinations of the original variables. Cross-validation is used as a strict test for the significance of each consecutive PLS components to avoid over-fitting problem and when the components are non-significant, the process is stopped. The PLS technique ensures that the model obtained is based on its predictivity rather than fitting the dependent variable. PLS is very useful for dealing with large number of variables, and it can also operate with intercorrelated variables. The theory of PLS technique was conceptualized by Wold.³⁹ One important aspect is that the number of components chosen in a PLS analysis should always be lower than the number of independent variables; otherwise it will be simply turned to a multiple linear regression equation if the number of components and variables are equal.

Supplementary Materials

Genetic function approximation (GFA)

Genetic algorithm resembles with the evolution of DNA. The genetic function approximation (GFA) algorithm was initially developed from the inspiration of two apparently dissimilar algorithms:

- (i) Holland's genetic algorithm⁴³ and
- (ii) Friedman's multivariate adaptive regression splines (MARS) algorithm.⁴⁴

Here a one-dimensional string of bits is used to represent individuals. Initially a population of individuals is created by using random initial bits. The quality of an individual is estimated by using a "fitness function" such that the "best" individuals are assigned with the best fitness score. Then such best fitness scored individuals are chosen and allowed for mating in order to transmit their genetic material to offspring through the crossover operation (a process in which each parent contributes pieces of genetic material followed by recombination to create the child). As a result of the discovery of "good" combinations of genes followed by spread through the population, the average fitness of the individuals in the population gets increased after many mating steps have been performed. A natural mapping of Holland's genetic approach by replacement of binary strings of Holland with the strings of basis function to the functional models approaches based on regression evolved into the development of genetic function approximation (GFA) algorithm. The advantages offered by GFA algorithm are as follows:

- i) The GFA method builds many models instead of a single model unlike the techniques like multivariate adaptive regression splines (MARS), classification and regression trees (CART), and principal component analysis

Supplementary Materials

(PCA) methods those incrementally add or delete basis function yielding a final regression model.

- ii) Features are automatically selected for their use in the basis function and full size models are tested for the determination of required number of basis functions.
- iii) Combination of basis functions are generated which makes the privilege of correlation between different features. Scoring of the models is done by using Friedman's "lack of fit" error measure⁴⁵ which prevents overfitting and provides the smoothness of fit to be controlled by the user. The "lack of fit" measure is defined by the following expression:

$$LOF = \frac{LSE}{\left(1 - \frac{c + d \times p}{M}\right)^2} \quad (S9)$$

where 'c' is the number of basis functions (other than constant term), 'd' is smoothing parameter (adjustable by the user), 'M' is number of samples in the training set, 'LSE' is least squares error and 'p' is total numbers of features contained in all basis functions. In our present study we have kept the mutation probabilities at 50% with 5000 iterations being performed. Smoothing parameter d was kept at 1.00. Initial equation length value was selected as 4 and the length of the final equation was not fixed.

Supplementary Materials

Statistical metrics for the regression model

Model quality metrics: R^2 and R_a^2

The quality parameters quantify the fitness of a developed QSAR/QSPR/QSTR model on a pure statistical basis.

The coefficient of determination, R^2 ,⁵¹ measures how closely the observed data tracks the fitted regression line and thus helps to quantify any variation in the data. Errors in either the model or in the data will lead to a bad fit. The maximum possible value for R^2 is 1, which means that there is perfect correlation. R^2 is calculated as the ratio of regression variance to the original variance where the regression variance is calculated as the original variance minus the variance around the regression line as shown by **Eq. S10**:

$$R^2 = 1 - \frac{\sum (Y_{obs(train)} - Y_{calc(train)})^2}{\sum (Y_{obs(train)} - \bar{Y}_{training})^2} \quad (\text{S10})$$

Here, $\bar{Y}_{training}$ is the mean observed response of the training set compounds. Addition of descriptors to the developed QSAR model increases the value of squared correlation coefficient; however, this may not necessarily indicate that the predictive ability of the model improves.

Another parameter used for testing the quality of generated regression equations is adjusted R^2 (R_a^2).⁵¹ R_a^2 (given by **Eq. S11**) is calculated to overcome the drawbacks associated with the value of R^2 . It is a modification of R^2 that adjusts for the number of

Supplementary Materials

explanatory terms in a model. Unlike R^2 , the value of R_a^2 increases only if the new term improves the model more than what would be expected by chance.

$$R_a^2 = \frac{(n-1)R^2 - p}{n-p-1} \quad (\text{S11})$$

However, acceptable values of these statistical parameters are not always sufficient enough to judge model predictivity and alternative methods are employed to assess the predictive ability of the developed QSAR models. Thus, internal and external validation experiments are performed in order to check the predictive potential of the developed models.

Internal and external validation metrics

Q^2 (or Q^2_{int})

For the determination of this parameter, the models are subjected to leave-one-out (LOO) cross-validation technique. In this technique, same operation is repeated in cycles; one compound is omitted from the data set at random in each cycle and then the model is built using the remaining compounds. In this way, a model from reduced set of compounds is formed which is used for the prediction of activity of the omitted compound. The process is iterated until all the compounds are eliminated once from the set of compounds. Q^2 is determined on the basis of the predictability of the model and it

Supplementary Materials

is also termed as cross-validated R^2 . The higher is the value of Q^2 (minimum acceptable threshold value is 0.5 and maximum value is 1), the better is the model predictivity. The cross-validated R^2 (Q^2) is expressed as:⁵²

$$Q^2 = 1 - \frac{\sum (Y_{obs} - Y_{pred})^2}{\sum (Y_{obs} - \bar{Y})^2} \quad (S12)$$

where Y_{obs} represents the observed activity of the training set compounds, Y_{pred} is the predicted activity of the training set compounds and \bar{Y} corresponds to the mean observed activity of the training set compounds.

$$R^2_{pred}$$

R^2_{pred} is termed as the predictive R^2 of a developed model. It is used to judge the external predictive ability of a QSAR model and to calculate this, one need to an external test set of chemicals which may be obtained from the whole dataset by applying suitable algorithm of division or can be obtained from external source as well. The training set compounds are used for the development of different models and the activity/toxicity of the test set compounds are predicted using the models. The external predictive ability of a model was determined by R^2_{pred} which can be defined as:⁵³

Supplementary Materials

$$R_{pred}^2 = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \bar{Y}_{training})^2} \quad (S13)$$

where $Y_{obs(test)}$ and $Y_{pred(test)}$ are the observed activity and predicted activity respectively of the test set compounds and $\bar{Y}_{training}$ corresponds to the mean of observed activity of the training set compounds. R_{pred}^2 value for an acceptable model should be greater than 0.5 and the maximum value is 1.

The r_m^2 metrics

It has been previously shown that⁵⁴ the squared cross-validated correlation coefficient alone might not give the true predictive capability of a model and hence a modified r^2 term was proposed by the present authors' group. The rationale for the development of r_m^2 was to penalize the R^2 value by calculating another R^2 term, r_0^2 and then taking the difference between them. The general formula for the metric r_m^2 can be presented by the following **Equation**:

$$r_m^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \quad (S14)$$

where r^2 and r_0^2 are the squared correlation coefficients between the observed and predicted values of the compounds with and without intercept respectively. In the initial studies, observed values were considered in y-axis whereas predicted values were

Supplementary Materials

considered in the x-axis and the parameter was determined for the validation of training, test as well as overall set (considering both the training and test set) and the corresponding notations were $r_m^2(LOO)$, $r_m^2(test)$, and $r_m^2(overall)$ respectively.^{54,55} Recently, another variation of r_m^2 has been proposed by the present authors' group,⁵⁶ where the axes are interchanged, *i.e.* predicted values are considered in y-axis and observed values are considered in the x-axis giving the parameter $r'_m{}^2$ as follows:

$$r'_m{}^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0'^2}\right) \quad (S15)$$

where $r_0'^2$ bear the same meaning but in the reversed axes. It is interesting to note that during the change of axes, the value of r^2 remain same while it is not true for the case of $r_0'^2$. When the observed values of the test set compounds (y-axis) are plotted against the predicted values of the compounds (x-axis) setting intercept to zero, the slope of the fitted line gives the value of k . Interchange of the axes gives the value of k' . The following equations are employed for the calculation of r^2 , $r_0'^2$, k and k' .

$$r_0'^2 = 1 - \frac{\sum (Y_{obs} - k \times Y_{pred})^2}{\sum (Y_{obs} - \overline{Y_{obs}})^2} \quad (S16)$$

$$r_0'^2 = 1 - \frac{\sum (Y_{pred} - k' \times Y_{obs})^2}{\sum (Y_{pred} - \overline{Y_{pred}})^2} \quad (S17)$$

Supplementary Materials

$$k = \frac{\sum (Y_{obs} \times Y_{pred})}{\sum (Y_{pred})^2} \quad (S18)$$

$$k' = \frac{\sum (Y_{obs} \times Y_{pred})}{\sum (Y_{obs})^2} \quad (S19)$$

In cases where the Y -range is not very wide, either of the r_m^2 and $r'_m{}^2$ metrics may penalize heavily the quality of predictions. Two more parameters have been developed⁵⁶ namely $\overline{r_m^2}$, the average value of r_m^2 and $r'_m{}^2$, and Δr_m^2 , the absolute difference between r_m^2 and $r'_m{}^2$. The average r_m^2 and $r'_m{}^2$ ($\overline{r_m^2}$) has been found to be better metric than the original r_m^2 metric. In general, the difference between r_m^2 and $r'_m{}^2$ values should be low for good models. It has been shown that the value of Δr_m^2 should preferably be lower than 0.2 provided that the value of $\overline{r_m^2}$ is more than 0.5. Like the original r_m^2 metric, the average and difference parameters can be deployed to judge the predictive quality of the training ($\overline{r_m^2}_{(LOO)}$ and $\Delta r_m^2_{(LOO)}$), test ($\overline{r_m^2}_{(test)}$ and $\Delta r_m^2_{(test)}$), and the overall ($\overline{r_m^2}_{(overall)}$ and $\Delta r_m^2_{(overall)}$) set as well.⁵⁶

Randomization

The statistical robustness of the model can further be checked by “Fischer’s Randomization” test (Y-randomization). Two types of randomization tests namely

Supplementary Materials

process randomization and model randomization tools can be performed. Process randomization involves random scrambling of the dependent variables followed by fresh selection of variables from the whole descriptor matrix, whereas in case of model randomization, same set of descriptors as present in the original model is regressed against the scrambled predictor variables. For an acceptable QSAR model, the average correlation coefficient (R_r) of randomized models should be less than the correlation coefficient (R) of non-randomized model. Previously, we have shown⁵⁴ that a parameter R_p^2 (as shown by **Eq. S20**) performs comparison between the squared correlation coefficient (R^2) of the original dataset with the squared average correlation coefficient (R_r^2) obtained from the randomized dataset and penalizes the model R^2 for a small difference between squared mean correlation coefficient (R_r^2) of randomized models and squared correlation coefficient (R^2) of the non-randomized model.

$$R_p^2 = R^2 \times \sqrt{R^2 - R_r^2} \quad (\text{S20})$$

Later a correction to the term R_p^2 was suggested by Todeschini⁵⁷ which is defined as:

$${}^c R_p^2 = R \times \sqrt{R^2 - R_r^2} \quad (\text{S21})$$

Ideally, for a random model, $R_r^2 = 0$ thus making ${}^c R_p^2$ equal to R^2 in such case. When R_r^2 has a considerable value, the value of ${}^c R_p^2$ is penalized reflecting chance correlation. For an acceptable and robust QSAR model, the value of ${}^c R_p^2$ should be greater than 0.5.

Supplementary Materials

Statistical parameters for the classification model

Wilk's lambda (λ) statistics

The Wilk's Lambda is a widely used parameter for the testing of a significance of discriminant model function. It is a distance based parameter and is calculated from the scalar transformations of the covariance matrices of "between" and "within" group variances. In a classification analysis, where at least two groups are present, Wilk's lambda is determined as the ratio of within group sum of squares and total sum of squares, *i.e.* within-category to total dispersion.⁵⁸

$$\text{Wilk's } \lambda = \frac{\text{Within group sum of square}}{\text{Total sum of square}} \quad (\text{S22})$$

Let us consider B_g and W_g are the random $p \times p$ independent variable matrix with the distribution $W_p(q, \Sigma)$ and $W_p(n, \Sigma)$, respectively considering $n > p$. Then the Wilk's λ will be given by the following equation:⁵⁸

$$\lambda = \det\left(\frac{W_g}{B_g + W_g}\right) \quad (\text{S23})$$

Where '*det*' refers to the determinant of the matrix constructed using descriptors. Wilk's lambda focuses on the best discriminating property of the analyzed independent variables and it spans from 0 to 1, where 0 correspond to different values of group means

Supplementary Materials

signifying good level of discrimination achieved by variable and 1 referring to similar group mean values meaning no discrimination achieved by the variables. Hence the value of Wilk's lambda for a good discriminant model should preferably be lower.

Sensitivity, specificity, false positive rate, accuracy, precision and F-measure

These are the calculated statistical parameters determined from the outcome of a classification model. These parameters describe the fraction or percentage of proper classification achieved by a discriminant (or some other classifier) model with respect to the *a priori* class defined by the user. In cases where training and test sets are involved, as in the present study, the discriminant model yields a predicted category for the test set compounds (not used in the model development), as well as a calculated class for the training set (here we are using the term calculated to denote the classes determined for the compounds whose actual class information was used for the development of the model unlike the compounds, the predefined class information of whose were not used during model development). Among the parameters, the *F*-measure is a derived one calculated using sensitivity and precision.

In a categorical analysis, let us assume that positive and negatives are the two predefined classes denoted by 'P' and 'N' respectively, and the results calculated/predicted by the classifier are 'Y' and 'N' respectively. Then four possible outcomes can be notified and defined as follows:⁶⁰

Supplementary Materials

True positive = TP = Positive samples correctly classified as positives.

False negative = FN = Positive samples incorrectly classified as negatives.

True negative = TN = Negative samples correctly classified as negatives.

False positive = FP = Negative samples incorrectly classified as positives.

This correct and erroneous classification results can be presented in the form of a cross-tabulation commonly known as the confusion matrix or contingency table⁶⁰ shown in **Fig. S1** where observed or the predefined class is presented horizontally and the predicted/hypothesized class is presented vertically. From the numerical counts of TP, FN, TN, and FP following parameters can be calculated to judge the model quality.

		True class	
		P	N
Hypothesized class	Y	TP	FP
	N	FN	TN

Fig. S1 Sample confusion matrix

Sensitivity corresponds to the fraction of correct prediction of positive samples and can be defined as the ratio of true positives to the total number of assigned/predefined positives. It is also termed as true positive rate (*tp rate*), recall and hit rate and is defined as:⁶⁰

Supplementary Materials

$$\text{Sensitivity} = \text{tp rate} = \frac{TP}{TP + FN} \quad (\text{S24})$$

Specificity resembles the fraction of correct prediction of negative samples referring to the ratio of number of samples properly classed as negative to the total number of negative sample predefined /assigned and is defined as:⁶⁰

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (\text{S25})$$

Another term may be calculated which is termed as the false positive rate (or false alarm rate) or fp rate, and this is actually the ratio between incorrectly classed negative sample and total assumed negative samples and it actually corresponds to (1-Specificity) value.⁶⁰

$$\text{fp rate} = \frac{FP}{TN + FP} = 1 - \text{Specificity} \quad (\text{S26})$$

Accuracy refers to the correct fraction of positives and negatives classified with respect to the total classification observed and can be defined as:⁶⁰

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (\text{S27})$$

Precision is a fraction corresponding to the positively predicted compounds and can be defined as:⁶⁰

Supplementary Materials

$$\text{Precision} = \text{positive predictive value} = \frac{TP}{TP + FP} \quad (\text{S28})$$

The parameter *F*-measure assesses the positive predictivity of a model and can be defined:⁶⁰

$$F - \text{measure} = \frac{2}{1/\text{Precision} + 1/\text{Sensitivity}} \quad (\text{S29})$$

For a good classification model, the *F*-measure value should be nearer to 1 and theoretically becomes infinite if a model totally fails to predict any TP.

Sensitivity, specificity, accuracy and precision vary from zero (worst prediction) to one (best prediction) and all the parameters can be converted to percentage by multiplying 100 with the corresponding fractions.

Receiver operating characteristics (ROC) curve

ROC curve provides a visual representation of the success and error observed in a classification model. The curve is plotted taking tp rate on the *Y* axis and fp rate on the *X* axis, and the nature of the curve provides easier detection of the correctness of prediction. Apart from toxicity classification problems, ROC curves have been a useful measure in

Supplementary Materials

signal detection theory since the past intended for determining the tradeoff between hit rates and false alarm rates of classifiers.⁶¹

A number of important zones can be classified from which the predictive performance of a developed model can be visualized. **Fig. S2** shows a sample ROC space, where the point (0, 1) corresponds to perfect prediction leading to 100% correct prediction of positive samples and 0% incorrect prediction of negative samples. The points (0, 0) and (1, 1) represent all negative and all positive classifier respectively, which are jointly termed as default classifier. The diagonal joining these two represents a random classifier performance that means, at this line if the probability of a sample to be positive is p , then the probability of it to be negative would be $(1-p)$. Data points located above the diagonal line seem to be good predictive because the true positive rate here is greater than random; on the contrary the points located below the diagonal will be termed as poorly predicted because of the same reason. The left top space of the graph can be termed as liberal performance zone where the rate of true positives is higher although considering false positive errors as the points move horizontally in the right direction. Likewise, the left bottom zone above the diagonal may be considered as conservative performance where though the rate of false positives may be lesser, the rate of true compounds is also not very higher.⁶¹

Supplementary Materials

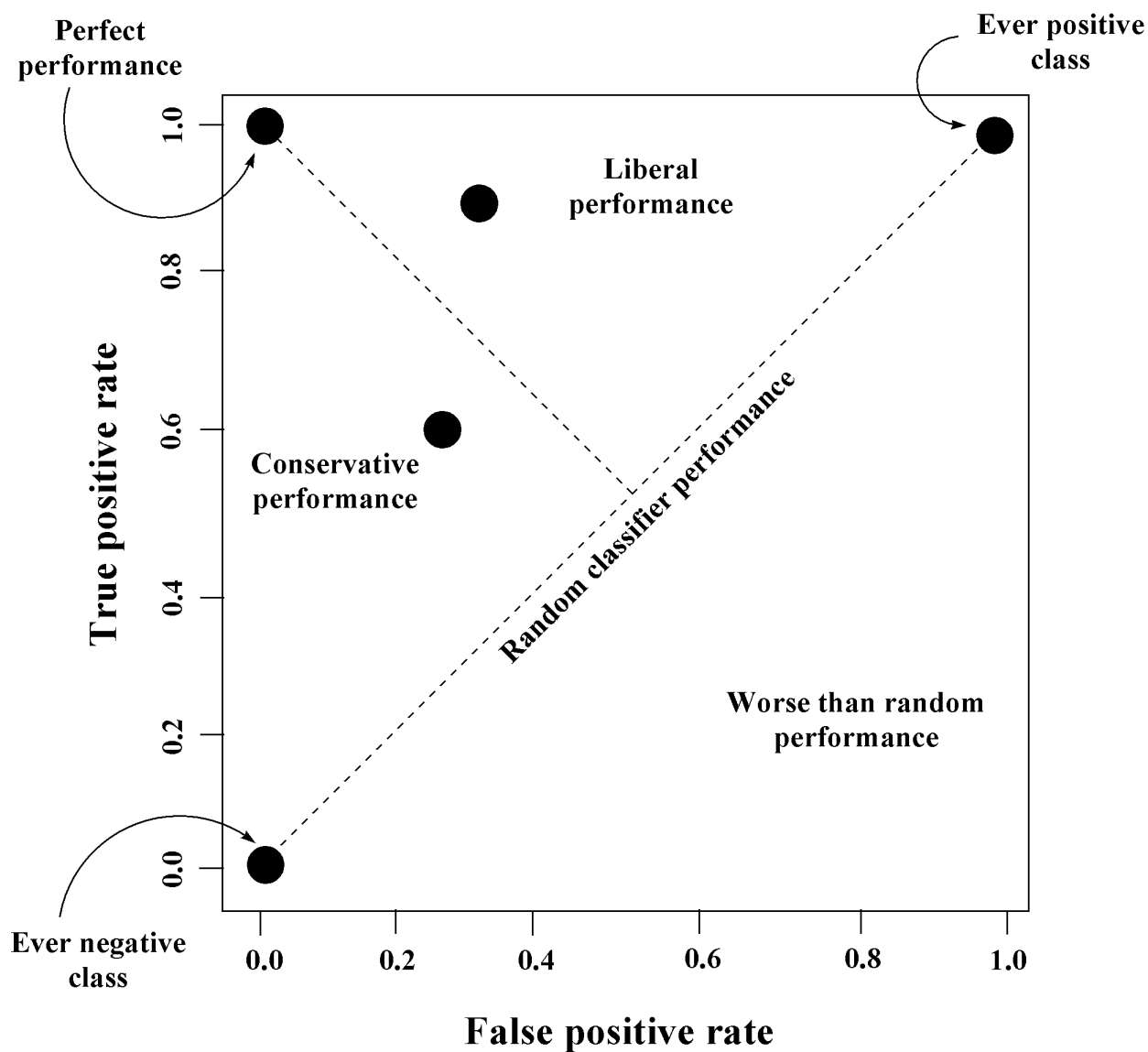


Fig. S2 Sample Receiver operating characteristics (ROC) space

It is observed that an ROC graph yields a two-dimensional graphical result which should be transformed into a single numerical value in order to compare classifiers. Calculation of area under the ROC curve (AUC)⁶² provides suitable means of achieving this object. AUC is still believed to be robust in predicting the discriminant quality of a classifier.

Supplementary Materials

Now as the plot area of the total ROC space is one square unit, the value of AUC will vary from zero to one. It is evident that across the random classifier line, *i.e.*, random guess of classifier will produce an area of 0.5 square unit (between 0, 0 and 1, 1), hence signifying that the acceptable AUC value should be greater than 0.5 (random classifier) and mostly closer to unity (perfect or ideal classifier).

However it has been observed that, the different parameters calculated for the statistical evaluation of ROC analysis, which includes sensitivity, specificity, accuracy, precision, F-measure, and the area under the curve, do not always suffice the required confidence of validation. Hence, new parameter of validation based on better statistical confidence and logic for ROC analysis is always sought for. New parameter such as Boltzmann-Enhanced Discrimination of ROC (BEDROC) based on Robust Initial Enhancement (RIE),⁶³ although showed good performance in identifying actives, later proved to be inefficient in virtual screening studies. Other parameter includes the Enrichment Factor (EF).⁶⁴ Two more newly prescribed parameters for determination of performance of ROC analysis are the ROC graph Euclidean distance (ROCED) and the ROC graph Euclidean distance corrected with Fitness Function (FIT (λ)) (ROCFIT).⁶⁵ Considering the top left corner space as the liberal performance zone, the measurement of distance of a point at this zone from the point (0, 1) can be a good parameter. A function of specificity and sensitivity for a perfect and a real classifier can be defined as their corresponding Euclidean distance d_i as follows:

$$d_i = \sqrt{(Se_p - Se_r)^2 + (Sp_p - Sp_r)^2} \quad (\text{S30})$$

Supplementary Materials

where Se_p and Se_r are the representative sensitivity values of the perfect and the real classifier, and Sp_p and Sp_r are the specificity values of the perfect and real classifier, respectively. Considering the sensitivity, and specificity values of a perfect classifier to be 1, the expression for the Euclidean distance in **Equation S30** reduces to the following form:

$$d_i = \sqrt{(1 - Se_r)^2 + (1 - Sp_r)^2} \quad (\text{S31})$$

Now this distance should be minimum from the point (0, 1) in the ROC graph for both the training and test set compounds to show good classification. The parameter ROCED can be defined from the Euclidean distance values for training and test sets as follows:

$$ROCED = \left(|d_{training} - d_{test}| + 1 \right) \times \left(d_{training} + d_{test} \right) \times \left(d_{test} + 1 \right) \quad (\text{S32})$$

where $d_{training}$ and d_{test} are the Euclidean distances in a ROC curve for training and test compounds respectively. Now, from **Equation S32** it is evident that the parameter ROCED will be minimum when the distances for training and test sets, i.e. $d_{training}$ and d_{test} are minimum corresponding to good classifier performance. Hence, a minimized value of ROCED primarily postulates a similar accuracy for training and test set; secondarily it also explains an almost perfect AUC value (unity) for both the sets, and finally it predicts the test set accuracy to be maximum. For a perfect classifier instance for both training and test sets, ROCED will be equal to zero, whereas it takes a value of 4.5 in an instance of random classifier where $d_{training}=0.5$ and $d_{test}=1$. Pérez-Garrido *et al.*⁶⁵ reported that a classifier having ROCED value greater than 2.5 should not be

Supplementary Materials

considered as acceptable because at this value the Euclidean distance of the training or the test set or both become greater than 0.7, signifying a random classification.

The parameter ROCFIT was defined to provide a fitness basis to the distance based ROCED parameter that circumvents the probable loss of significance of the variables.

ROCFIT can be defined as:⁶⁵

$$ROCFIT = \frac{ROCED}{FIT(\lambda)} \quad (S33)$$

Here the FIT (λ) corresponds to the Wilk's λ parameter. The Wilk's λ parameter is good for explaining the classification of training set, but it does not provide good predictivity for the external or test set chemicals. The ROCFIT parameter eliminates this problem as it contains the term ROCED which address the issue of external predictivity.

Another most popular parameter is the Matthews correlation coefficient,⁶⁶ employed as a measure in machine learning operations. Along with the true positives, and true negatives, this parameter also considers the effect of false positive and negatives. The principal advantage of this parameter is that MCC can be applied to groups of different sizes in the form of a balanced approach. The term MCC can be defined as below:⁶⁶

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (S34)$$

where TP , TN , FP and FN refer the same meaning as discussed before. From **Equation S34** it can be observed that MCC spans over a range varying from -1 to +1; where a

Supplementary Materials

perfect classifier will have the MCC value of unity, a random classifier will be characterized by 0, and -1 represents an inverse prediction. Hence the desired value of MCC should be nearer to one. This parameter in particular addresses the issue of improper explanation of a confusion matrix, and the cases where the dataset sizes are higher.

Pharmacological distribution diagram (PDD)

Pharmacological distribution diagram (PDD) is a frequency distribution plot of a dependent variable where expectancy values of the variable is plotted in the Y-axis against numeric intervals of the variable in the X-axis. In a classification issue, expectancy refers to the probability of categorization of a compound in a specific group for a specific value of the discriminant function. During LDA, a discriminant function (DF) is developed, which is a mathematical equation, used for the calculation of discriminant scores of every individual compounds. Then the discriminant function values of all samples are taken in the abscissa in the form of range, and the expectancy values (probability of activity) are plotted in the ordinate against those ranges. Hence, this graph visually signifies the overlapping regions of the categories *e.g.*, positives and negatives, as well as it shows the regions of DF values that possess maximal probability of finding actives and inactives.⁶⁷ For a classification case comprising of two classes active and inactive (or positive and negative), two terms named ‘active expectancy’ and ‘inactive expectancy’ may be defined as below where the denominator is added with a numerical value of 100 to avoid division by zero:⁶⁷

Supplementary Materials

$$\text{Activity expectancy} = E_a = \frac{\text{Percentage of actives}}{(\text{Percentage of inactives} + 100)} \quad (\text{S35})$$

$$\text{Inactivity expectancy} = E_i = \frac{\text{Percentage of inactives}}{(\text{Percentage of actives} + 100)} \quad (\text{S36})$$

where ‘*a*’ and ‘*i*’ are the number of occurrences of active and inactive compounds at a specific range. It can be evidently understood that for a perfect classification scheme, the active (positive) and inactive (negative) compounds will always be characterized by different ranges of DF values, and hence in an ideal discriminant operation, the actives will always be separated than the inactives whereas overlapping of them will correspond to error in prediction referring to false positives as well as false negatives.

OECD principles for QSAR model development

Development of predictive models using regression techniques as a rational modeling technique is well supported by many international organizations throughout the world. On the basis of the discussions proposed at the Setubal workshop (Portugal, 2002) on the “Regulatory Acceptance of (Q)SARs for Human Health and Environmental Endpoints”,⁶⁸ the Organization for Economic Co-operation and Development (OECD) member countries and the European Commission proposed a five point guidelines at the 37th OECD's Joint Meeting of the Chemicals Committee and the Working Party on

Supplementary Materials

Chemicals, Pesticides and Biotechnology (Joint Meeting), 2004⁶⁹. The guidelines are briefly discussed below.

1. A defined endpoint.
2. An unambiguous algorithm.
3. A defined domain of applicability.
4. Appropriate measures of goodness-of-fit, robustness and predictivity; and
5. A mechanistic interpretation, if possible.

A defined endpoint: OECD principle 1

This principle states that a distinct endpoint should be used for QSAR modeling purpose where the measurements of property/activity/toxicity must not involve different experimental protocols as well as varying treatment condition.

An unambiguous algorithm: OECD principle 2

Here the objective is to guarantee the transparency of the model algorithm which will be easily reproducible by others. It is desired that the method of model development and the logic should be clearly defined and described so that it can be used for new chemicals

Supplementary Materials

(not used in model development) as well. Consistency must also be performed during generation and selection of descriptors to make it an overt one.

A defined domain of applicability: OECD principle 3

Applicability domain (AD) is a postulated hypothetical region in chemical space. It is described and defined by the independent variables as well as dependent variable used to build a mathematical model of interest. Technically, AD represents the chemical space defined by the structural information of the chemicals used in model development, *i.e.*, the training set compounds in a QSAR analysis. In this present study we have used the leverage approach⁷⁰ to define the domain of applicability of the compounds.

Leverage approach checks the structural applicability of external compounds (not used in model development) to lie within the chemical space defined by AD. Reliability is obtained if the external data is interpolated by the model, otherwise during extrapolation of AD to address an external chemical, uncertainty rises. In the leverage approach, the degree of extrapolation is presented in a suitable form of distance, measured with respect to the centroid of the developed chemical space (using training set compounds) and finally compounds can be identified as less or more influential depending upon their closer or extreme relative distances respectively from the centroid. Training set chemicals having higher leverage values may be considered as ‘good leverage’ chemicals because they widens the AD, whereas the same for a test set chemical will be taken as ‘bad leverage’ ones as it tries to extrapolate the model AD⁷¹.

Supplementary Materials

In the present study, compounds with cross-validated standardized residual value greater than 3 standard deviation unit were considered as response outliers. In order to graphically explain the nature of AD, Williams plot was constructed by plotting the HAT diagonal values (h) in the abscissa against standardized cross-validated residuals in the ordinate. A critical leverage value h^* ($h^* = 3p$) was considered with p' denoting the number of descriptors plus one and n representing the number of compounds used in the development of model (training set compounds), to define a limiting condition. Any chemical showing leverage value greater than h^* would be classified as high leverage chemical, and the training set ones will be considered as influential to that AD⁷¹.

Appropriate measures of goodness-of-fit, robustness and predictivity: OECD principle

4

Principle 4 aims at evaluation of the developed QSAR models by carrying out proper statistical validation that will encompass internal performance of the model using a training set (represented by goodness-of-fit and robustness), and its external validation using a test set (shown by external predictivity).

Determination coefficient⁵¹ R^2 is one of the widely used parameters of measuring 'goodness-of-fit' of a model. However, R^2 tends to rise by performing fitting as the number of descriptors is increased and hence other validation parameters are calculated. Adjusted R^2 (R_a^2)⁵¹ is a superior parameter than R^2 as it considers the total number of variables present in the model and only tends to rise if new variable tends to improve

Supplementary Materials

model quality. Cross-validation appears to be a better technique of predicting internal validation and the corresponding parameter is Q^2 . R^2_{pred} is a widely and successfully used parameter for external validation.

A mechanistic interpretation (if any): OECD principle 5

This principle has been proposed to address the possibility of a mechanistic association between the descriptors and the endpoint and also to ensure documentation of such association wherever possible.

Details of the developed PLS model

From **Equation S37** it can be observed that the PLS model, is characterized by encouraging predicted variance of 65.70% and explained variance of 69.20% for the training set. It also showed an acceptable predictive R^2 (R^2_{pred}) value of 0.718.

$$\begin{aligned} pEC_{50}(\text{mol/L}) &= 1.547 + 0.777 \times MSD_{(\text{cation})} - 0.002 \times CENT_{(\text{cation})} + 4.354 \times [\eta]_{(\text{anion})}^{\text{local}} + 0.510 \times SaasC_{(\text{cation})} \\ &- 1.104 \times \frac{\sum \alpha_x}{\sum \alpha_{(\text{anion})}} - 0.735 \times NsOH_{(\text{cation})} - 0.407 \times IC_{2(\text{cation})} - 0.466 \times \chi^{\text{av}}_{(\text{anion})} \end{aligned} \quad (\text{S37})$$

$N_{\text{training}} = 90; N_{\text{test}} = 36; LV_s = 4;$
 $R^2 = 0.717; R_a^2 = 0.692; Q^2 = 0.657; \overline{r_{m(\text{LOO})}^2} = 0.534; \overline{\Delta r_{m(\text{LOO})}^2} = 0.201;$
 $R_{\text{pred}}^2 = 0.718; \overline{r_{m(\text{test})}^2} = 0.579; \overline{\Delta r_{m(\text{test})}^2} = 0.238; \overline{r_{m(\text{overall})}^2} = 0.551; \overline{\Delta r_{m(\text{overall})}^2} = 0.214.$

Additionally, it is observed that though the Δr_m^2 parameter is slightly higher than the proposed tolerance value of 0.2 (0.201 for $\Delta r_m^2_{(\text{LOO})}$, 0.238 for $\Delta r_m^2_{(\text{test})}$, and 0.214 for $\Delta r_m^2_{(\text{overall})}$), the average r_m^2 metrics are above the acceptable value of 0.500 in all the

Supplementary Materials

cases of internal ($\overline{r_{m(LOO)}^2}=0.534$), external ($\overline{r_{m(test)}^2}=0.579$) and overall ($\overline{r_{m(overall)}^2}=0.551$) validation.

The PLS model (**Equation S37**) comprises of four latent variables encoding eight independent variables of which five represent the effects of cation and the rest three describe their importance on anionic species. The parameter $MSD_{(cation)}$ that principally describes molecular branching and also correlates with shape of the molecule indicates that the toxicity of ionic liquids to *V. fischeri* is related to the shape and branching pattern of its cationic fragment. CENT, the centralization index, describes the importance of branching of cationic groups in predicting *V. fischeri* of ILs. $SaasC_{(cation)}$ is an electrotopological state atom index (previously defined) for the cations, and it describes the importance of aromaticity with modeled toxicity values. $NsOH_{(cation)}$ corresponds to the importance of hydroxyl fragment (-OH) in cationic groups. **Equation S37** also shows that the *V. fischeri* toxicity of IL cations is related to the neighborhood symmetry shown by the second order information index parameter $IC_{2(cation)}$, referring to the structural complexity of the vertex in a hydrogen suppressed molecular graph. Three more parameters for anions are present in **Equation S37** of which two are extended topochemical atom (ETA) indices and the rest one is a connectivity parameter. The ETA parameter $[\Sigma\alpha]_x/\Sigma\alpha$ signifies the importance of molecular branching where one central atom is substituted with four other non-hydrogen atoms, and $[r]^{local}$ corresponds to the contribution of branching considering local topology. Finally, the descriptor $\chi_{(anion)}^{4,av}$ is the average fourth order valence connectivity index, and it exemplifies the importance of branching among anionic groups.

Supplementary Materials

Table S1 The twelve principles of green chemistry¹

Principle	Statement
Principle 1. Prevention	It is better to prevent waste than to treat or clean up waste after it has been created.
Principle 2. Atom Economy	Synthetic methods should be designed to maximize the incorporation of all materials used in the process into the final product.
Principle 3. Less Hazardous Chemical Synthesis	Synthetic methods should be designed to use and generate substances that possess little or no toxicity to people or the environment wherever practicable.
Principle 4. Designing Safer Chemicals	Chemical products should be designed to exhibit their desired function while minimizing their toxicity.
Principle 5. Safer Solvents and Auxiliaries	The use of auxiliary substances (e.g., solvents or separation agents) should be made unnecessary whenever possible and innocuous when used.
Principle 6. Design for Energy Efficiency	Energy requirements of chemical processes should be recognized for their environmental and economic impacts and should be minimized. If possible, synthetic methods should be conducted at ambient temperature and pressure.

Supplementary Materials

Principle 7. Use of Renewable Feedstock	A raw material or feedstock should be renewable rather than depleting whenever technically and economically practicable.
Principle 8. Reduce Derivatives	Unnecessary derivatization (use of blocking groups, protection/de-protection, and temporary modification of physical/chemical processes) should be minimized or avoided if possible, because such steps require additional reagents and can generate waste.
Principle 9. Catalysis	Catalytic reagents (as selective as possible) are superior to stoichiometric reagents.
Principle 10. Design for Degradation	Chemical products should be designed so that at the end of their function they break down into innocuous degradation products and do not persist in the environment.
Principle 11. Real-time Analysis for Pollution Prevention	Analytical methodologies need to be further developed to allow for real-time, in-process monitoring and control prior to the formation of hazardous substances.
Principle 12. Inherently Safer Chemistry for Accident Prevention	Substances and the form of a substance used in a chemical process should be chosen to minimize the potential for chemical accidents, including releases, explosions, and fires.

Supplementary Materials

Table S2 List of compounds with their observed and calculated/predicted toxicity values and classifications

Sl. No.	Compounds	Toxicity to <i>Vibrio fischeri</i>		Calculated/ Predicted classification obtained from LDA ¹	Results derived using discriminant equation (Eq. 1) for PDD ²	
		Observed pEC ₅₀ (mol/L) ³¹	Calculated/ Predicted pEC ₅₀ (mol/L) [Model 2/ Eq. 2]		DF	Classification determined using DF
Toxic class (Observed pEC ₅₀ >3.670 mol L ⁻¹) ⁴						
4	[IM][Cap]	3.710	3.050	P	3.899	+
8	[C1IM][Cap]	3.870	3.082	P	4.082	+
23	[C2MIM][FeCl4]	4.490	3.764	P	2.145	+
30	[MOC2MIM][BF4]	4.850	3.329	P	1.520	+
31	[MOC2MIM][N(CN)2]	4.930	3.013	P	0.172	U
41	[C4IM][Cap]	4.000	3.436	P	6.278	+
49*	[C4MIM][8OSO3]	4.170	3.444	N	-2.833	-
54	[C4MIM][FeCl4]	4.490	4.131	P	3.639	+
57	[C6MIM][Br]	4.580	3.397	N	-3.481	-
58*	[C6MIM][Cl]	3.850	3.397	N	-3.821	-
59	[C6MMIM][Cl]	4.260	3.631	P	-0.576	-
60	[C6MIM][PF6]	3.860	3.994	N	-2.672	-
62*	[C6MIM][N(CF3SO2)2]	3.950	3.611	N	-4.126	-
65	[C6EIM][BF4]	3.850	4.306	N	-3.003	-
67*	[C8MIM][Br]	5.370	4.398	P	5.093	+

Supplementary Materials

68	[C8MIM][Cl]	4.950	4.398	P	4.754	+
69*	[C8MIM][PF6]	5.180	4.994	P	5.903	+
70*	[C8MIM][BF4]	4.600	4.956	P	5.389	+
71*	[C8MIM][N(CF3SO2)2]	5.170	4.612	P	4.448	+
72	[C9MIM][BF4]	5.280	5.604	P	-0.765	-
73	[C10MIM][Cl]	5.870	5.802	P	-0.612	-
74*	[C10MIM][BF4]	6.180	6.360	P	0.024	U
75	[C10MIM][FeCl4]	6.430	7.206	P	8.476	+
76	[C14MIM][Cl]	6.150	5.637	P	2.457	+
77*	[C16MIM][Cl]	5.770	5.554	P	3.952	+
78	[C18MIM][Cl]	4.650	5.471	P	5.424	+
89*	[C4MMPy][N(CN)2]	3.880	3.379	N	-7.982	-
90	[C6MPy][Br]	3.930	3.715	P	4.923	+
91	[C6MPy][Cl]	4.560	3.715	P	4.584	+
92*	[C8Py][Cl]	4.310	4.436	N	-2.355	-
93	[C8MPy][Br]	5.210	4.860	P	1.706	+
94*	[C8MMPy][Br]	4.880	5.312	P	5.348	+
95	[C8MMPy][N(CF3SO2)2]	5.640	5.526	P	4.703	+
100	[C4MPyRR][P(C2F5)3F3]	4.300	2.987	P	-1.541	-
108	[C4(CH3)2N-Py][Br]	3.680	3.592	P	5.076	+
110	[C4(CH3)2N-Py][N(CF3SO2)2]	4.150	3.806	P	4.431	+
111	[C6(CH3)2N-Py][N(CF3SO2)2]	4.620	4.879	P	-0.024	-
114	[TMG][TFA]	5.530	2.889	N	-5.732	-
117*	[TMG][Cap]	3.710	3.268	P	5.784	+
118*	[Melamine][TFA]	3.760	3.029	N	-5.271	-

Supplementary Materials

121	[C16MMMN][Br]	5.960	5.624	P	3.655	+
122	[C16MMMN][Cl]	5.600	5.624	P	3.315	+
123	[C16BnMMN][Cl]	6.310	5.364	P	7.506	+
124*	[C12BnMMN][Cl]	6.230	5.530	P	4.276	+
140	[C14XXXP][P(C2F5)3F3]	4.300	3.730	P	0.696	+
143[†]	[12OSO3]	4.500	2.844	P	4.754	+

Non-toxic class (Observed pEC₅₀<3.670 mol L⁻¹)^b

1	[IM][TFA]	1.930	2.671	N	-7.617	-
2*	[IM][Ace]	2.390	2.674	N	-8.361	-
3	[IM][TfO]	2.190	2.657	N	-8.626	-
5	[C1IM][TFA]	2.430	2.702	N	-7.435	-
6	[C1IM][Ace]	1.930	2.705	N	-8.179	-
7	[C1IM][TfO]	1.650	2.688	N	-8.443	-
9	[C1IM][For]	3.170	2.583	N	-8.814	-
10[†]	[C1MIM][MetSO4]	<1.240	2.713	N	-8.818	-
11	[CNC1MIM][N(CF3SO2)2]	2.190	2.784	N	-6.362	-
12	[EOC1MIM][N(CF3SO2)2]	3.000	3.004	N	-5.459	-
13[†]	[EOC1MIM][Cl]	<1.670	2.790	N	-5.153	-
14	[C2MIM][EtSO4]	1.980	2.840	N	-7.651	-
15*	[C2MIM][(2-OPhO)2B]	3.020	2.344	N	-6.139	-
16*	[C2MIM][Cl]	1.560	2.361	N	-6.943	-
17	[C2MIM][(C2F5)2PO2]	2.950	2.903	N	-4.791	-
18[†]	[C2MIM][(OCCOO)2B]	<1.720	2.517	N	-9.494	-
19[†]	[C2MIM][MetSO4]	<1.720	2.799	N	-8.286	-
20*	[C2MIM][(CN)4B]	2.440	2.532	N	-9.285	-

Supplementary Materials

21	[C2MIM][SCN]	1.850	2.804	N	-7.310	—
22	[C2MIM][TFA]	1.720	2.823	N	-6.781	—
24 [†]	[EOC2MIM][Br]	<1.700	3.087	N	-4.002	—
25	[EOC2MIM][N(CF3SO2)2]	3.040	3.301	N	-4.647	—
26	[OHC2MIM][N(CF3SO2)2]	1.920	1.925	N	-6.244	—
27*	[OHC2MIM][I]	2.110	1.711	N	-5.303	—
28*	[MOC2MIM][N(CF3SO2)2]	3.160	2.985	P	0.579	+
29	[MOC2MIM][Cl]	1.820	2.770	P	0.885	+
32*	[C3MIM][BF4]	2.060	3.081	N	-5.598	—
33	[C3MIM][N(CF3SO2)2]	3.230	2.737	N	-6.539	—
34	[OHC3MIM][N(CF3SO2)2]	2.110	2.140	N	-5.459	—
35 [†]	[OHC3MIM][Cl]	<1.700	1.926	N	-5.153	—
36*	[MOC3MIM][N(CF3SO2)2]	2.760	3.287	N	-4.647	—
37 [†]	[MOC3MIM][Cl]	<1.700	3.073	N	-4.342	—
38	[C4IM][TFA]	2.720	3.057	N	-5.238	—
39	[C4IM][Ace]	2.680	3.060	N	-5.982	—
40	[C4IM][TfO]	3.230	3.043	N	-6.247	—
42	[C4IM][For]	2.810	2.938	N	-6.618	—
43*	[C4MIM][PF6]	2.710	3.324	N	-4.299	—
44*	[C4MIM][BF4]	2.570	3.285	N	-4.813	—
45	[C4MIM][Br]	2.550	2.727	N	-5.108	—
46*	[C4MIM][Cl]	2.590	2.727	N	-5.448	—
47	[C4MIM][I]	2.410	2.727	N	-4.813	—
48	[C4MIM][N(CN)2]	2.330	2.969	N	-6.160	—
50	[C4MIM][TfO]	2.400	3.175	N	-6.295	—

Supplementary Materials

51	[C4MIM][N(CF3SO2)2]	3.220	2.942	N	-5.754	—
52	[C4MIM][N(CF3)2]	2.530	3.242	N	-4.452	—
53	[C4MIM][pTS]	2.480	2.973	N	-4.477	—
55*	[C4EIM][BF4]	3.200	3.513	N	-4.554	—
56	[C5MIM][BF4]	2.860	3.593	N	-4.001	—
61*	[C6MIM][BF4]	2.820	3.955	N	-3.185	—
63	[C6MIM][P(C2F5)3F3]	3.350	3.900	P	1.199	+
64*	[C6MIM][2-SO2PhCO)N]	3.330	3.439	N	-3.050	—
66	[C7MIM][BF4]	3.560	4.409	N	-2.370	—
79†	[MPy]	2.930	2.283	N	-7.643	—
80	[OHC3Py][N(CF3SO2)2]	1.990	2.164	N	-5.575	—
81	[C4Py][Br]	2.610	2.751	N	-5.224	—
82	[C4MPy][Br]	3.250	2.930	P	-1.553	—
83	[C4MMPy][Br]	3.410	3.137	N	-6.930	—
84	[C4Py][Cl]	2.710	2.751	N	-5.564	—
85	[C4Py][Al2Cl7]	2.990	3.292	N	-2.838	—
86*	[C4MPy][BF4]	2.980	3.488	P	-1.257	—
87	[C4Py][N(CN)2]	2.700	2.993	N	-6.276	—
88	[C4MPy][N(CN)2]	3.340	3.172	N	-2.605	—
96	[EOC2MPyRR][N(CF3SO2)2]	3.030	3.098	N	-5.760	—
97	[OHC3MPyRR][N(CF3SO2)2]	2.090	1.938	N	-6.571	—
98†	[C4MPyRR][Cl]	<1.700	2.484	N	-6.560	—
99*	[C4MPyRR][N(CF3SO2)2]	3.460	2.699	N	-6.866	—
101	[C6MPyRR][Cl]	3.000	3.147	N	-4.914	—
102	[C6MPyRR][N(CF3SO2)2]	3.600	3.361	N	-5.220	—

Supplementary Materials

103	[EOC1MMorp][N(CF3SO2)2]	2.620	3.036	N	-6.338	—
104[†]	[C4MMorp][Br]	<1.700	2.723	N	-5.986	—
105	[C4MMorp][N(CF3SO2)2]	3.510	2.937	N	-6.631	—
106	[C4MPiper][Br]	1.730	2.565	N	-6.281	—
107	[C4MPiper][N(CF3SO2)2]	3.440	2.779	N	-6.926	—
109*	[C4(CH3)2N-Py][Cl]	3.480	3.592	P	4.736	+
112[†]	[Choline][Cl]	<1.000	1.453	N	-6.283	—
113	[Choline][N(CF3SO2)2]	1.850	1.667	N	-6.589	—
115	[TMG][Ace]	2.830	2.891	N	-6.477	—
116	[TMG][TfO]	2.480	2.874	N	-6.741	—
119	[Melamine][Ace]	3.610	3.032	N	-6.015	—
120*	[Melamine][TfO]	3.550	3.015	N	-6.279	—
125	[C4N][Cl]	3.310	4.335	N	-7.214	—
126[†]	[C1N][Br]	<1.000	2.151	N	-7.267	—
127[†]	[C2N][Br]	<1.000	2.490	N	-7.108	—
128*	[C4N][Br]	2.730	4.335	N	-6.875	—
129	[C6EEEN][Br]	3.540	3.306	N	-4.945	—
130	[C4EMMN][N(CF3SO2)2]	2.790	2.576	N	-6.204	—
131[†]	[C4EMMN][Cl]	<1.700	2.362	N	-5.898	—
132	[EOCOC1EMMN][N(CF3SO2)2]	2.680	3.065	N	-5.789	—
133	[OHC2MMN][Ace]	1.710	1.858	N	-6.151	—
134[†]	[OHC2MMN][OHAce]	<1.700	1.901	N	-6.978	—
135[†]	[(MOC2)2N][NH2O3S]	<1.700	3.029	N	-5.187	—
136	[MOC2EMMN][N(CF3SO2)2]	2.650	2.661	N	-5.909	—
137*	[C4P][Br]	3.290	2.640	N	-7.934	—

Supplementary Materials

138*	[(C4)3EP][[(CH3CH2)2PO4]	2.930	2.815	N	-7.365	—
139	[C14XXXP][Br]	2.590	3.227	N	-3.984	—
141†	[N(CF3SO2)2]	<1.700	2.315	N	-10.018	—
142†	[(2-OPhO)2B]	3.240	2.085	N	-8.908	—
144†	[8OSO3]	3.540	2.817	N	-7.097	—
145†*	[BF4]	<1.700	2.659	N	-9.076	—
146	[C2S][N(CF3SO2)2]	2.740	2.520	N	-7.280	—
147	[EOCOC1MPyRR][N(CF3SO2)2]	2.770	3.735	N	-5.435	—

¶ Calculated (training set) and predicted (test set) classification obtained from the linear discriminant analysis. ‘P’ denotes toxic (positive) and ‘N’ denotes non-toxic (negative) observations. The validation parameters are sensitivity, specificity, accuracy, precision and F-measure, and are presented in **Table 4** for training and test sets separately.

‡ The classification derived using **Eq. 1** according to the mentioned ranges of DF; here ‘+’ is toxic, ‘-’ is non-toxic, and ‘U’ is undetermined toxicity.

^a The validation parameters for the *a priori* toxic group determined from DF are as follows:

Training set: Undetermined = 3.23%; False non-toxicity (Positives incorrectly classified) = 29.03%; Overall accuracy = 67.74%; Adjusted accuracy = 70.00%.

Test set: Undetermined = 6.67%; False non-toxicity (Positives incorrectly classified) = 40.00%; Overall accuracy = 53.33%; Adjusted accuracy = 57.14%.

^b The validation parameters for the *a priori* non-toxic group determined from DF are as follows:

Supplementary Materials

Training set: Undetermined = 0.00%; False toxicity (Negatives incorrectly classified) = 2.53%; Overall accuracy = 97.47%; Adjusted accuracy = 97.47%.

Test set: Undetermined = 0.00%; False toxicity (Negatives incorrectly classified) = 9.09%; Overall accuracy = 90.91%; Adjusted accuracy = 90.91%.

* Test set compounds.

† Compounds not included in QSAR analysis due to lack of proper numerical value and toxicities are predicted using **Equation 2**.

Supplementary Materials

Table S3 Categorical list of all calculated descriptors

Category of descriptors	Names and notations of the descriptors
Indicator variables for anions	(2-OPhO) ₂ B, (2-SO ₂ PhCO)N, (C ₂ F ₅) ₂ PO ₂ , (CH ₃ CH ₂) ₂ PO ₄ , (OCCOO) ₂ B, I2OSO ₃ , 8OSO ₃ , Ace, Al ₂ Cl ₇ , B(CN) ₄ , BF ₄ , Br, Cap, Cl, EtSO ₄ , FeCl ₄ , For, I, MetSO ₄ , N(CF ₃) ₂ , N(CF ₃ SO ₂) ₂ , N(CN) ₂ , NH ₂ O ₃ S, OHAcce, P(C ₂ F ₅) ₃ F ₃ , PF ₆ , pTS, SCN, TFA, TfO.
Indicator variables for cations	(C ₄) ₃ EP, (MOC ₂) ₂ N, C ₁₀ MIM, C ₁₂ BnMMN, C ₁₄ MIM, C ₁₄ XXXP, C ₁₆ BnMMN, C ₁₆ MIM, C ₁₆ MMMn, C ₁₈ MIM, C ₁ IM, C ₁ MIM, C ₁ N, C ₂ MIM, C ₂ N, C ₂ S, C ₃ MIM, C ₄ (CH ₃) ₂ N-Py, C ₄ EIM, C ₄ EMMN, C ₄ IM, C ₄ MIM, C ₄ MMorp, C ₄ MMPy, C ₄ MPiper, C ₄ MPy, C ₄ MPyRR, C ₄ N, C ₄ P, C ₄ Py, C ₅ MIM, C ₆ (CH ₃) ₂ N-Py, C ₆ EEEN, C ₆ EIM, C ₆ MIM, C ₆ MMIM, C ₆ MPy, C ₆ MPyrrol, C ₇ MIM, C ₈ MIM, C ₈ MMPy, C ₈ MPy, C ₈ Py, C ₉ MIM, Choline, CNC ₁ MIM, EOC ₁ MIM, EOC ₁ MMorp, EOC ₂ MIM, EOC ₂ MPyRR, EOCOC ₁ EMMN, EOCOC ₁ MPyRR, IM, Melamine, MOC ₂ EMMN, MOC ₂ MIM, MOC ₃ MIM, MPy, OHC ₂ MIM, OHC ₂ MMN, OHC ₃ MIM, OHC ₃ MPyRR, OHC ₃ Py, TMG.
Constitutional	MW, AMW, Sv, Se, Sp, Si, Mv, Me, Mp, Mi, nAT, nSK, nBT, nBO, nBM, SCBO, RBN, RBF, nDB, nTB, nAB, nH, nC, nN, nO, nP, nS, nF, nCl, nBR, nI, nB, nHM, nHet, nX.
Topological	ZM1, ZM1V, ZM1Kup, ZM1Mad, ZM1Per, ZM1MulPer, ZM2, ZM2V, ZM2Kup, ZM2Mad, ZM2Per, ZM2MulPer, ON0, ON0V, ON1, ON1V, Qindex, BBI, DBI, SNar, HNar, GNar, Xt, Dz, Ram, BLI, Pol, LPRS, MSD, SPI, PJI2, ECC, AECC,

Supplementary Materials

	DECC, MDDD, UNIP, CENT, VAR, ICR, SMTI, SMTIV, GMTI, GMTIV, Xu, CSI,
	Wap, S1K, S2K, S3K, PHI, PW2, PW3, PW4, PW5, MAXDN, MAXDP, DELS,
	TIE, Psi _{i_s} , Psi _{i_A} , Psi _{i_0} , Psi _{i_1} , Psi _{i_t} , Psi _{i_0d} , Psi _{i_1d} , Psi _{i_1s} ,
	Psi _{e_A} , Psi _{e_0} , Psi _{e_1} , Psi _{e_t} , Psi _{e_0d} , Psi _{e_1d} , Psi _{e_1s} , BAC, LOC.
	X0, X1, X2, X3, X4, X5, X0A, X1A, X2A, X3A, X4A, X5A, X0v, X1v, X2v, X3v,
Connectivity	X4v, X5v, X0Av, X1Av, X2Av, X3Av, X4Av, X5Av, X0sol, X1sol, X2sol, X3sol,
	X4sol, X5sol, XMOD, RDCHI, RDSQ, X1Kup, X1Mad, X1Per, X1MulPer.
	ISIZ, IAC, AAC, IDE, IDM, IDDE, IDDM, IDET, IDMT, IVDE, IVDM, S0K,
Information	HVcpx, HDcpx, Uindex, Vindex, Xindex, Yindex, IC0, IC1, IC2, IC3, IC4, IC5,
indices	TIC0, TIC1, TIC2, TIC3, TIC4, TIC5, SIC0, SIC1, SIC2, SIC3, SIC4, SIC5, CIC0,
	CIC1, CIC2, CIC3, CIC4, CIC5, BIC0, BIC1, BIC2, BIC3, BIC4, BIC5.
Extended	
topochemical	$\Sigma\alpha$, $\Sigma\alpha/N_v$, $\Sigma\varepsilon$, $\Sigma\varepsilon/N$, $\Sigma\beta_s$, $\Sigma\beta'_s$, $\Sigma\beta_{ns}$, $\Sigma\beta'_{ns}$, $\Sigma\beta$, $\Sigma\beta'$, η , η' , η^{local} , $[\eta']^{local}$, η_F , η'_F ,
atom (ETA)	η_F^{local} , $[\eta']_F^{local}$, η_B , η'_B , $[\Sigma\alpha]_P/\Sigma\alpha$, $[\Sigma\alpha]_V/\Sigma\alpha$, $[\Sigma\alpha]_X/\Sigma\alpha$.
indices	
	SsCH ₃ , SdCH ₂ , SssCH ₂ , StCH, SdsCH, SaaCH, SsssCH, SddC, StsC, SdssC, SaasC,
	SaaaC, SssssC, SsNH ₂ , SssNH, SdNH, SsssN, SdsN, SaaN, StN, SsNH ₃ ⁺ , SssNH ₂ ⁺ ,
	SdNH ₂ ⁺ , SsssNH ⁺ , SssssN ⁺ , SddsN, SaasN, SaaNH, SsOH, SdO, SssO, SaaO,
Atom-type E-	SsPH ₂ , SssPH, SsssP, SdsssP, SddsP, SsssssP, SsSH, SdS, SssS, SaaS, SdssS,
state indices	SddssS, SsssssS, SsF, SsCl, SsBr, SsI, SsBH ₂ , SssBH, SsssB, SssssB-, NsCH ₃ ,
	NdCH ₂ , NssCH ₂ , NtCH, NdsCH, NaaCH, NsssCH, NddC, NtsC, NdssC, NaasC,
	NaaaC, NssssC, NsNH ₂ , NssNH, NdNH, NsssN, NdsN, NaaN, NtN, NsNH ₃ ⁺ ,
	NssNH ₂ ⁺ , NdNH ₂ ⁺ , NsssNH ⁺ , NssssN ⁺ , NddsN, NaasN, NaaNH, NsOH, NdO,

Supplementary Materials

NssO, NaaO, NsPH₂, NssPH, NsssP, NdsssP, NddsP, NsssssP, NsSH, NdS, NssS,

NaaS, NdssS, NddssS, NssssssS, NsF, NsCl, NsBr, NsI, NsBH₂, NssBH, NsssB,

NssssB-

Molecular

properties

Uc, Ui, Hy, AMR, TPSA(NO), TPSA(Tot), MLOGP, MLOGP2, ALOGP, ALOGP2.


Supplementary Materials

Table S4 Serial numbers of the test set compounds of the dataset in different clusters

Cluster Number	Number of occurrence of test set compounds	Sl. number of the test set compounds
1	10	15, 28, 36, 49, 62, 64, 71, 99, 117, 138
2	2	77, 124
3	10	16, 27, 46, 58, 67, 92, 94, 109, 128, 137
4	14	2, 20, 32, 43, 44, 55, 61, 69, 70, 74, 86, 89, 118, 120
5	1	145

Supplementary Materials

Table S5 Definition of selected descriptors present in the best models

Sl. No.	Name	Definition
1	MSD	<p>Mean square distance index:³⁴ A distance matrix parameter. For a connected molecular graph G, MSD is defined as follows:</p> $MSD = \frac{\left(\sum_{i=1}^A \sum_{j=1}^A (\delta_{ij}^2) \right)^{0.5}}{A \times (A-1)} = \left(\frac{\sum_{k=1}^D ({}^k f \times k^2)}{\sum_{k=1}^D {}^k f} \right)$ <p>where δ_{ij} is the topological distance, A is the number of atoms, D is the topological diameter, and ${}^k f$ is the graph distance count of order k. Value of MSD index decreases with increasing molecular branching in an isomeric series.</p> <p>An electrotopological state atom index descriptor.³⁴ It corresponds to the contribution of aromatic carbon atom attached to two more aromatic carbon and one non-hydrogen atom presented by the following fragment: . The value of this parameter increases as the occurrence of such fragment in a molecule rises. It describes the importance of such fragments referring to the aromaticity of a compound.</p> <p>It is the core count of extended topochemical atom (ETA) indices.³⁵⁻³⁷ For non-hydrogen vertex, α value for individual atom can be defined as described below. For a connected molecular graph G, the values of α are summed over the atoms.</p>
2	SaaS	
3	$\Sigma\alpha$	

Supplementary Materials

$$\alpha = \frac{Z - Z^v}{Z^v} \times \frac{1}{PN - 1}$$

where PN stands for period number, while Z and Z^v represent atomic number and valence electron number respectively. Here hydrogen atom (H) being considered as the reference and α for hydrogen is taken to be zero. Increase in α value corresponds to increase in molecular bulk.

Centralization index:³⁴ A topological parameter derived from distance matrix and is defined as:

$$CENT = \Delta\sigma^* = 2 \cdot W - A \cdot \sigma^*$$

4 CENT

where σ^* is the minimum value of vertex distance degree, A is the number of atoms, and W is the Wiener index which is defined as,

$$W = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \delta_{ij} \text{ where } \delta_{ij} \text{ is the shortest topological distance for } N \text{ number of}$$

vertices. $CENT$ increases with decreased branching.

Gutman molecular topological index:³⁴ A distance matrix based topological index.

For a connected molecular graph G , $GMTI(S_G)$ can be defined as below.

5 GMTI

$$S_G = \sum_{i=1}^A \sum_{j=1}^A \delta_i \delta_j \cdot d_{ij}$$

where $\delta_i \delta_j \cdot d_{ij}$ is the topological distance between the vertices v_i and v_j weighted by the product of the endpoint vertex degrees. $GMTI$ reduces with raised branching in isomeric molecular series.

Supplementary Materials

Kier and Hall's connectivity index,³⁴ termed as the valence connectivity index of order five. For a connected molecular graph G , following is the m^{th} (higher order) order valence connectivity index:

$$6 \quad {}^5\chi^v = \sum_{k=1}^K \left(\prod_{a=1}^n \delta_a \right)_k^{-0.5} ; \text{ for } {}^5\chi^v, m=5$$

where k runs over all of the m^{th} order subgraphs constituted by n atoms, K is the total number of m^{th} order subgraphs present in the molecular graph G , δ is the vertex degree, the product is taken over all the vertices involved in constituting each subgraph. ${}^5\chi^v$ index is inversely proportional to branching and unsaturation content for five order bond fragments.

$[\eta']^{\text{local}}$ is a variant of the composite index η .³⁵⁻³⁷ The composite index is defined as given below:

$$\eta = \sum_{i < j} \left[\frac{\gamma_i \gamma_j}{r_{ij}^2} \right]^{0.5}$$

7 $[\eta']^{\text{local}}$ where γ_i is the VEM (valence electron mobile) vertex count of the i^{th} vertex and $\gamma_i = \alpha_i / \beta_i$, α_i being the core count for the i^{th} vertex and β_i stands for VEM count considering all bonds connected to the atom and lone pair of electrons (if any); r_{ij} is the topological distance between vertex i and j . Now, when $r_{ij}=1$, *i.e.* only interactions due to unit topological distance is considered, the corresponding parameter is termed as η^{local} and in order to avoid dependence of vertex, it is divided by the vertex count N_v , which gives a more reliable parameter $[\eta']^{\text{local}}$. It is therefore defined as:

Supplementary Materials

$$[\eta']^{local} = \frac{\eta^{local}}{N_v} = \frac{\sum_{i<j, r_{ij}=1} (\gamma_i \gamma_j)^{0.5}}{N_v}$$

$[\eta']^{local}$ corresponds to branching and holds a proportional relationship with it.

Solvation connectivity index of fourth order.³⁴ Considering the characteristic dimension of the molecules by atomic parameters, a formal definition can be made as:

$${}^m\chi_q^s = \frac{1}{2^{m+1}} \times \sum_{k=1}^K \frac{\left(\prod_{a=1}^n L_a \right)_k}{\left(\prod_{a=1}^n \delta_a \right)_k^{0.5}} ; \text{for the present case } m=4$$

8

${}^4\chi^s$

where L_a refers to the principal quantum number of the a^{th} atom in the k^{th} subgraph and δ_a is the corresponding vertex degree; m is the order of the subgraph and K is the total number of such subgraphs; the subscript q indicates the type of molecular subgraph; n corresponds to the number of non-hydrogen vertices in m^{th} order subgraph. $1/(2^{m+1})$ is a normalization factor used to make a relationship between this solvation connectivity parameter and Kier and Hall's normal connectivity index; it is placed such that the values of ${}^m\chi$ and ${}^m\chi^s$ for compounds having only second row atoms coincide. Solvation connectivity index defines the solvation entropy and describe dispersion interactions of a compound in solution.

Supplementary Materials

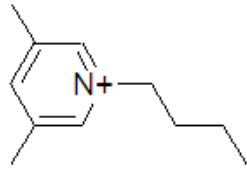
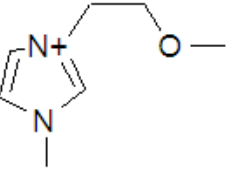
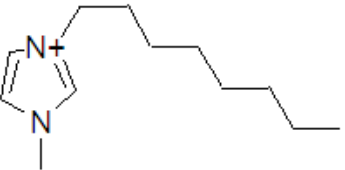
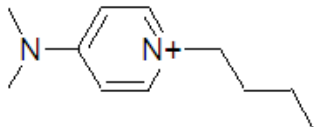
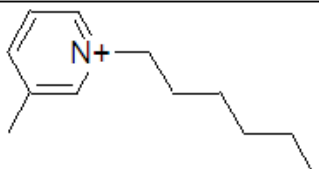
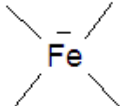
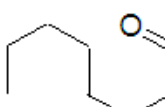
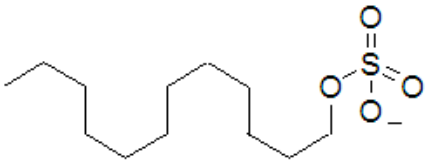
Chemical name	Formula code	Structure
Butyldimethylpyridinium	C ₄ MMPy	
Methoxyethylmethylimidazolium	MOC ₂ MIM	
Octylmethylimidazolium	C ₈ MIM	
Butyl(dimethylamino)pyridinium	C ₄ (CH ₃) ₂ N-Py	
Hexylmethylpyridinium	C ₆ MPy	
Tetrachloroferrate(III)	FeCl ₄ ⁻	
Caprylate	Cap	
<i>o</i> -Dodecylsulfate	12OSO ₃ ⁻	

Fig. S3 Structural representation of the best ionic indicator variables present in the classification model

Supplementary Materials

References

See main text.