CROATICA CHEMICA ACTA
CCACAA, ISSN 0011-1643, e-ISSN 1334-417X
Croat. Chem. Acta **87** (1) (2014) 29–37.
http://dx.doi.org/10.5562/cca2284

Original Scientific Article

QSPR Modeling of Odor Threshold of Aliphatic Alcohols Using Extended Topochemical Atom (ETA) Indices†

Pallabi Pal, Indrani Mitra, and Kunal Roy*

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

RECEIVED APRIL 19, 2013; REVISED NOVEMBER 26, 2013; ACCEPTED DECEMBER 17, 2013

Abstract. The present work establishes a quantitative structure-property relationship (QSPR) between topochemical features and odor threshold (OT) of aliphatic alcohols. A data set of 53 aliphatic alcohols was chosen for the analysis employing different chemometric techniques, among which, genetic function approximation with spline option (GFA-spline) showed the most acceptable results in terms of internal and external validation metric values. The extended topochemical atom (ETA) indices, developed by the present authors' group, were considered as descriptors for model development. Additionally, selected non-ETA descriptors were also tried for model development. It was observed that the models with ETA indices significantly surpass the predictive ability of the models developed using other descriptors. The final model suggests that molecular branching and electronic parameters significantly influence the odor potency of the molecules. Additionally, increased lipophilicity and reduced electronegativity increase the odorant property. The model thus developed may effectively be used for prediction of odor threshold of any untested aliphatic alcohols. (doi: 10.5562/cca2284)

Keywords: QSPR, ETA, topochemical, odorant, odor threshold

INTRODUCTION

Odorant compounds constitute a significant portion of the organic chemistry. Those which are present in the environment, facilitate to identify their presence in air with their typical odor characteristic. The threshold of olfaction presents a key feature to all the odor active compounds, the value of which may differ due to variation in the protocols for measurement. Olfaction, a physico-chemical property can be defined as the least concentration of any air borne chemical that is perceived by half of the healthy tested individual.[1] However, two chemicals having same odor threshold may not produce the same level of annoyance in the surroundings, as it depends on the type of odor of those chemicals.[2] This demonstrates the presence of a complex mechanism of action of the odorant receptors (ORs). An odorous molecule present in the environment is supposed to bind to a number of ORs at a time. Thus, the ambiguous nature of odorant receptors along with various characteristics of olfactory data has enhanced the urge to gain information about threshold data for odor of various compounds which have its wide application in the field of bioscience, food chemistry and environmental pollution.[3-5] In relation to the mechanism behind odorant binding, earlier it was proposed that odorants of similar property activates common receptor subtypes. But later, it was proved to be wrong as it was seen that homologous oxygenated aliphatic molecules though having similar molecular properties do not share similar quality of odor. Thus, it was proposed that the theory behind the olfactory mechanism lies in the combinatorial effects of different types of receptors.[6] Again, it was found that majority of the mammalian olfactory receptors belong to the class A of G-protein coupled receptors (GPCRs) superfamily, although a convincing reason behind the anomalous behavior of these olfactory receptors is still unknown.[6]

Thus, a great deal of attention of the present research has been oriented towards the development of models that enable the prediction of the odor threshold of compounds and aid in understanding the facts behind their binding possibilities by avoiding time consuming and costly experimental setup. In this aspect, the *in-silico* prediction of olfactory threshold using quantitative structure-property relationship (QSPR)[7] technique gets highlighted. It is a method by which the structural information of any chemical compound can be correlated

† Dedicated to Professor Douglas Jay Klein on the occasion of his 70th birthday.

* Author to whom correspondence should be addressed. (E-mail: kunalroy_in@yahoo.com)

to its respective property value. This involves the extraction of chemical information in the form of descriptors, followed by their correlation with property values of individual compounds giving a predictive mathematical equation.[8] From this computational technique, it is easier to know the structural fragments that alter the physicochemical properties of compounds which further help to design new potential molecules with low odor threshold. Moreover, these developed predictive models can also assist in the screening of potent odorant moieties from large database of compounds, which reduces the requirement of time consuming synthesis and testing analysis of a large number of odorous compounds for different purposes. The QSPR paradigm is now supported by the Registration, Evaluation, and Authorization of Chemicals (REACH) norms,[9] a legislatory initiative of the European Commission and also by the organization for economic cooperation and development (OECD).[10,11] Again, these types of *in-silico* predictive models are used by the Food and Drug Administration (FDA)[12] for minimizing the rate of false negatives and false positives saving incalculable costs for manufacturers. The Council for International Organizations of Medical Sciences (CIOMS)[13] also recommends the methods such as in silico mathematical models, computer simulation, and the use of *in vitro* biological systems before animal experiments for the advancement of biological knowledge. However, all QSPR/QSAR[14] (activity)/QSTR[15] (toxicity) models should be sufficiently validated before their application for prediction of new data.

The QSPR/QSAR approach has been widely used by various research groups for successful prediction of odor potency. Luan *et al.*[16] established that support vector machine (SVM) can be an effective tool in QSAR studies for developing classification based model of fragrance properties. 91 organic compounds were used for building both linear and non-linear models where the non-linear model (SVM) showed superior predictability than the linear model developed using the linear discriminant analysis. Again, QSAR approach was taken up by Du *et al.*[17] where the researchers utilized 64 volatile organic compounds for prediction of odor detection thresholds and nasal pungency thresholds (NPTs) for the olfaction and nasal trigeminal chemosensory systems. The best model was developed using local lazy regression method which proved to be effective even when the experimental property values are unknown.

In the present study, QSPR analysis has been carried out for establishing a relationship between the odor threshold (OT) data of 53 aliphatic alcohols and their structural attributes. Different types of descriptors were calculated for the purpose, but, a very simple class of 2D descriptors showed to be the most important one for prediction of OT values. These 2D descriptors belong to the extended topochemical atom (ETA)[18,19] indices which have been shown previously to be very much effective while prediction of other properties like solubility[20] and CMC[21] values. The calculation of ETA parameters does not involve the requirement of computationally exhaustive conformational analysis and alignment procedure. Thus less computational time is required for calculation of these variables than the complex 3D descriptors. The first generation ETA descriptors were developed based on TAU descriptors representing the valence electron mobile (VEM) environment.[22,23] The development history and formalism of the first generation ETA indices have been detailed in a book chapter by Roy and Das.[19] It provides information regarding the electronic features, size, shape, branching, and functionality of molecules. Moreover, the second generation variables can describe the electron richness, unsaturation, polar surface area and ability of hydrogen-bond formation of a given molecule. The first and second generation ETA indices are now available in PaDEL-Descriptor (version 2.11),[24] an open source software available at http://padel.nus.edu.sg/software/padeldescriptor.

MATERIALS AND METHODS

The Dataset

The dataset for the present QSPR study has been collected from a report by Anker *et al.*[25] The negative logarithmic value of the average of observed highest and lowest odor thresholds for each compound has been considered as the response variable for the current analysis. In total, there are 53 dataset compounds comprising of different aliphatic alcohols, the odor thresholds of which are expressed in mol/L. The list of compounds has been given in Table 1.

Model Development

For generation of reliable QSPR models based on odor threshold data, firstly descriptors were calculated which have considerable contribution for modulating the values of the physicochemical property concerned. The independent variables comprised of descriptors from PaDEL-Descriptor,[24] Cerius2[26] and Dragon software[27] platform. 107 descriptors were finally considered after elimination of highly correlated variables and also those whose variance was less than 0.0001. This pool of descriptors that has been utilized for model development has been shown in Table S1 in the Supplementary Materials section. The non ETA descriptors[28] include topological, structural, physicochemical, electronic and spatial types whereas the ETA descriptors include both first and second generation variables. All the ETA parameters are further discussed in Table S2 of the

Table 1. List of 53 aliphatic alcohols with their corresponding observed and predicted/calculated log(1/*T*) value

Sl. no.	Compounds	Obsereved log(1/*T*)[a]	Pred./Calc. log(1/*T*)[b]
1	Ethanol	3.20	3.07
2	1-propanol	4.00	3.99
3	1-butanol	4.70	4.49
4[c]	2-methyl-1-propanol	3.70	4.03
5	1-pentanol	4.50	4.42
6[c]	3-methyl-1-butanol	4.60	4.65
7	2,2-dimethyl-l-propanol	4.70	4.41
8	1-hexanol	4.90	4.83
9[c]	2-methyl-1-pentanol	5.00	5.06
10	3-methyl-1-pentanol	4.60	4.85
11	4-methyl-I-pentanol	4.60	4.85
12[c]	1-heptanol	5.41	5.96
13[c]	2,2-dimethyl-l-pentanol	6.68	5.96
14	1-octanol	6.32	5.86
15[c]	1-nonanol	6.37	6.17
16	1-decanol	7.30	6.54
17[c]	1-undecanol	5.84	6.61
18	1-dodecanol	5.87	6.88
19	2-propanol	3.00	3.28
20	2-butanol	3.40	3.87
21	2-methyl-2-propanol	3.70	3.79
22	2-pentanol	3.90	4.47
23	3-methyl-2-butanol	5.16	4.38
24	2-hexanol	5.62	4.78
25	3-methyl-2-pentanol	4.60	4.85
26	4-methyl-2-pentanol	4.50	4.86
27	2-heptanol	6.28	5.70
28	2-octanol	6.72	5.85
29	2-nonanol	6.31	5.92
30	2-decanol	6.68	6.62
31[c]	2-methyl-3-pentanol	5.21	5.06
32	3-methyl-3-pentanol	4.50	4.86
33	3-ethyl-3-pentanol	6.27	5.70
34	2,4-dimethyl-3-pentanol	5.67	5.74
35	4-heptanol	5.28	5.77
36[c]	6-undecanol	6.84	6.61
37[c]	2,4-dimethyl-1-pentanol	5.56	5.96
38	2,3-dimethyl-2-butanol	5.35	4.80
39	3-methyl-2-hexanol	5.68	5.74
40	3-pentanol	4.10	4.45
41	3-heptanol	5.55	5.75
42[c]	2-methyl-3-hexanol	6.26	5.96
43	3-octanol	5.90	5.88
44	2-methyl-3-heptanol	5.46	5.90
45	3-methyl-3-heptanol	6.72	5.85
46	5-methyl-3-heptanol	6.35	5.86
47	6-methyl-3-heptanol	5.59	5.89
48	3-nonanol	5.94	5.96
49	3-decanol	5.81	6.73
50	4-octanol	5.33	5.90
51	2-methyl-4-heptanol	5.33	5.90
52	4-nonanol	5.76	5.97
53	5-nonanol	5.94	5.96

[a] Observed log(1/*T*) value, expressed in mol/L.
[b] Calculated values for training set compounds and predicted values for test set compounds (expressed in mol/L).
[c] Test set compounds.

Supplementary Materials section. Compounds were divided into two classes: one comprising 42 compounds which has been considered as training set and the other is the test set comprising 11 compounds (size ration 4:1). Division of whole dataset into training and test set plays an important feature in model development, since the quality of the QSPR model depends highly on the selection of training and test sets.[29] In the present study, *k*-means clustering technique,[30] available in the SPSS software,[31] was employed for the splitting of the dataset. Five clusters were generated according to the features available for the respective compounds. 42 compounds were selected from the total cluster so that the training set encompasses the entire range of chemical space of the whole dataset. Figure S1 in Supplementary Materials shows the plot of the first three principal components of the variables and depicts that each test set compound remains in close vicinity of at least one training compound.[32,33]

Chemometric Tools Employed for Model Development

The training set was utilized for model development and the test set for subsequent model validation. At first the total pool of independent variables comprising both first and second generation ETA descriptors was used. Different algorithms were also utilized for model development keeping the division of the training and test set compounds unaltered. These include GFA-MLR[34] (genetic function approximation followed by multi linear regression) and G/PLS[35] (genetic/partial least squares) methodologies. Both linear and spline options were considered for each method. For the GFA-MLR models, the selection of the best model was done based on the lowest LOF[36] (lack-of-fit) score using 5000 crossovers. The G/PLS models were derived in Cerius2 software, at 1000 iterations using scaled variables The smoothness parameter value was kept at a value of 1.0. Further, the compounds of the training set were utilized for the generation of QSPR models with all other non ETA parameters, using the same algorithms for model generation. And lastly, the whole descriptor pool was employed for the third category of model development where both ETA and non-ETA variables were considered. It included a total of 107 descriptors.

Validation of the QSPR Models

All the final QSPR models, developed using three sets of descriptors (ETA, non-ETA and both ETA and non-ETA), were selected based on the significant values of different statistical parameters[37] such as determination coefficient (R^2), explained variance (R^2_a) and variance ratio (F) at specified degrees of freedom (df) describing the quality of the model (the F value should be significant

at $p < 0.01$). The standard errors of all regression coefficients should be sufficiently low so that corresponding '*t*' values are significant at $p < 0.01$. The error involved or the accuracy involved in the model development can also be understood from the value of standard error of estimate (*s*) and rmse values. The definitions of different statistical metrics for equation quality are given in Supplementary Materials section.

All the developed models were subjected to extensive statistical validation involving internal, external and overall strategies and thereby complying with the proper OECD guidelines. For ensuring the internal validation of the deduced QSPR models, leave-one-out cross-validation technique has been employed, the results being presented by the cross-validated squared correlation coefficient (Q^2 or $Q^2_{(LOO)}$). The Q^2 value takes care of the statistical significance of the model since it is calculated by using the LOO predicted values of the training set compounds that are generated during each leave-one-out (LOO) cross validated cycle. Furthermore, the r_m^2 metrics[38] were calculated for describing the performed internal (LOO) validation. Equations (1) and (2) give the formulae for computing the r_m^2 metrics.

$$\overline{r_m^2} = (r_m^2 + r_m'^2)/2 \quad (1)$$

$$\Delta r_m^2 = \left| r_m^2 - r_m'^2 \right| \quad (2)$$

where $r_m^2 = r^2 \times (1 - \sqrt{(r^2 - r_0^2)})$ and

$r_m'^2 = r^2 \times (1 - \sqrt{(r^2 - r_0'^2)})$

in which r^2 represents the squared correlation coefficient between the observed and predicted (LOO predicted) data of compounds with intercept (*i.e.*, for the regression line, observed = slope × predicted + intercept) , r_0^2 represents the same without the intercept (*i.e.*, for the regression line, observed = new_slope × predicted), whereas $r_0'^2$ holds same meaning as r_0^2 when the axes are interchanged. For asserting the accuracy of the validation strategies, the standard value of $\overline{r_m^2}$ and Δr_m^2 were considered as value >0.5 and <0.2 respectively.

The test set compounds have been employed to check the predictive ability of the model and thus verifying its external predictive potential. The resultant R^2_{pred}[39] value thus determines the predictability of the developed model in determining the odor threshold values of similar type of untested compounds. Here also the r_m^2 metrics for external validation have been applied.[40] Thus, $r^2_{m(test)}$ and $\Delta r^2_{m(test)}$ determine the proximity between the predicted value and the original experimental value of the test set compounds of the dataset. Additionally, $r^2_{m(overall)}$ metrics[38] have further been verified for all the QSPR models. The calculation is similar to that of Equations 1 and 2 where all the compounds of the dataset have been considered instead of the training set compounds. The parameters, $r^2_{m(overall)}$ and $\Delta r^2_{m(overall)}$, signify the overall performance of the deduced models.

Additionally, the robustness of the models has been also ascertained by the Y-randomization test[41] available in the Cerius2 software. Here, many models were developed after randomizing the values of dependent variable while keeping the descriptor matrix intact. A QSPR model is said to be robust if the value of R^2 of the non-random model is more than the square of average value of R (R_r^2) of the randomized models. For the present study, both process as well as model randomization tests have been performed for the final model developed using only the ETA variables at 90 % and 99 % confidence levels respectively. Finally, an additional metric, the ${}^cR_p^2$ value has been calculated using the following formula (Equation 3), which shows the reliability of the model and the process by which it has been established.[42]

$${}^cR_p^2 = R\sqrt{\left(R^2 - R_r^2\right)} \quad (3)$$

According to the point 3 of OECD[10,11] principles of QSAR models development, the applicability domain of a QSPR model must also be well defined since a single *in-silico* predictive model cannot be universally accepted for all types of compounds. The domain of applicability is a theoretical space covering the model descriptors and response variables of the training set. In this study, the domain of applicability has been determined following the leverage approach (Williams plot).[43] The plot has leverage values (*h*) on the *x* axis with standardized residual values on the *y* axis. The leverage (*h*) of a compound in the original variable space is calculated based on the HAT matrix as $\mathbf{H} = (\mathbf{X}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X})$, where **H** is an $(n \times n)$ matrix that orthogonally projects vectors into the space spanned by the columns of **X**. The leverage values of all the compounds were calculated using the Statistica software[44] which help to determine whether that compound fits in the applicability domain of the model or not. Here, the critical leverage value, *h**, was calculated using the mathematical formula: $h^* = 3((p+1)/n)$ and standardized residual limit for the boundary of the applicability domain was set at $\pm 2.5\sigma$.

RESULTS AND DISCUSSION

Diverse models were developed using two different chemometric tools namely, GFA and G/PLS algorithms employing three different set of descriptors (ETA, non-ETA and combination of ETA and non-ETA). In the

Table 2. Results obtained using different chemometric tools

Sl no.	Type of descriptors	Statistical method used	Model no.	Descriptors	LVs	Different quality and validation metrics		
						R^2	Q^2	R^2_{pred}
1.	ETA	GFA-linear	1	$[\Sigma\alpha]_P/\Sigma\alpha$, η'		0.755	0.715	0.725
		GFA-spline	**2**[(a)]	**$\Sigma\beta'_s$,<0.74427–η'>, $\Sigma\alpha/N_v$**		**0.809**	**0.778**	**0.813**
		GPLS-linear	3	$\Delta\varepsilon_D$, $\Delta\varepsilon_A$, $\Delta\beta'$	2	0.747	0.709	0.704
		GPLS-spline	4	<$\Delta\psi_A$–0.09197>,<0.13043–$[\Sigma\alpha]_Y/\Sigma\alpha$>, <$[\eta'_F]^{local}$–0.04902>,<0.765–$\eta'$>,<0.46875–$\Sigma\beta'$>	3	0.820	0.789	0.774
2.	Non-ETA	GFA-linear	**5**	LogP, Jurs_WPSA-2		0.784	0.736	0.784
		GFA-spline	6	<6.94975–$^0\chi$>,<21.83–Shadow_YZ>		0.814	0.779	0.628
		GPLS-linear	7	Jurs_WPSA_2, $^1\chi$,S_ssCH2, SC-2	3	0.801	0.741	0.691
		GPLS-spline	8	<2-Atype_C_1>,<Jurs_DPSA_3–21.7477>, <3.13–LogP>,<ShadowYZ–27.3894>	3	0.819	0.749	0.807
3.	ETA+ non-ETA	GFA-linear	9	LogP, $[\eta'_F]^{local}$, Jurs_WPSA-2		0.806	0.767	0.782
		GFA-spline	10	<η'–0.74427>,<3.21–LogP>, <6.48264-Shadow-Ylength>		0.807	0.755	0.668
		GPLS-linear	11	Wiener, $[\eta'_F]^{local}$, LogP, Jurs_WPSA_2	3	0.798	0.758	0.785
		GPLS-spline	12	<9-SC_O>,<nCs-4>,<0.21429–$[\Sigma\alpha]_Y/\Sigma\alpha$>, <Zagreb-20>	3	0.802	0.748	0.759

[(a)] Equation (4).

present study, four sets of models were developed for each descriptor set. Comparison among the models is described briefly in Table 2. Most of the models showed encouraging statistical parameters proving the reliability of the models and the process of its development. The equation (Equation 4, model 2; see below) bearing best prediction ability [with respect to both internal and external validation measures (Q^2 = 0.778, R^2_{pred} = 0.813)] has been selected as the best model pertaining to odor threshold values. From Table 3 it could be noted that the 2D descriptor, ETA, plays a significant role in correlating the property value ($\log(1/T)$) with that of the structural features of each molecule.

Though good quality models were developed out of non-ETA variables, but it was noted that when ETA parameters were added to the descriptor matrix, QSPR models having improved prediction power were obtained. For example, model 5 (Table 3) was developed using GFA-linear algorithm. All the non-ETA descriptors were utilized for model development, among which the lipophilic factor LogP and 3D descriptor Jurs_WPSA_2 were selected for the generation of the model. The corresponding R^2 and Q^2 values were 0.784 and 0.736. But, when an ETA parameter, $[\eta'_F]^{local}$ was introduced in the model (Equation 12) using the same algorithm and definite division of dataset, it was observed that the R^2 and Q^2 values were enhanced to 0.806 and 0.767 respectively. In case of models with non ETA parameters, though the R^2 values were satisfactory but the R^2_{pred} values were moderately low. The criterion of a good QSPR model is not only to have good R^2 and Q^2 values but the prime necessity is that the model should bear good prediction capacity. The prediction ability of the models also gets enhanced on the addition of ETA descriptors. This was observed for Equations 13 and 12 where the presence of only one ETA parameter in each model (η' and $[\eta'_F]^{local}$ respectively) significantly increases the ability of the model in predicting the $\log(1/T)$ value of different compounds. Thus, the use of ETA parameter in a QSPR model is noteworthy. In this context, it was noticed that a model with only ETA descriptors provided good statistical quality along with better predictability in comparison to all other QSPR models on odor threshold that were developed using non ETA and combined pool of descriptors (ETA and non-ETA).

Discussion of the best model

Among all the employed modeling techniques, the GFA-spline algorithm gave the best results. These models explained nonlinearity of the developed correlation. The spline terms are denoted in the parenthesis, *e.g.* <*f*(*x*) – *a*> where '*f*(*x*)' denotes the variable while '*a*' is known as the knot of the spline representing an optimum value of the independent variable. For each case, the total spline term has been considered as zero if the summation of the knot and the value lie in the negative range.[34]

Three different models were developed using the GFA-spline tool using three different descriptor matrices. Among them, model developed using only the ETA variables showed the best results towards the prediction of odor threshold values. In QSAR studies, more emphasis is now given to the predictive quality of a model. Model 2 in Table 3 shows the highest Q^2 (internal validation metric) and highest R^2_{pred} (external validation metric) among all the tabulated models and hence has been selected as the best model.

$$\begin{aligned}
&\log\left(1/T\right) = -10880(\pm 4622) + \\
&65021(\pm 27734) \times \Sigma\alpha/Nv - 43243(\pm 18489) \times \\
&\Sigma\beta's + 12.491(\pm 2.492) \times \langle 0.74427 - \eta' \rangle \\
&n_{training} = 42, LOF = 0.279, s = 0.463, \\
&F\left(df\right) = 53.82\left(3,38\right), R^2 = 0.809, \\
&R^2_a = 0.794, PRESS = 9.478, Q^2 = 0.778, \\
&rmsep(\text{int}) = 0.475, \overline{r^2_{m(LOO)}} = 0.685, \\
&\Delta r^2_{m(LOO)} = 0.156, n_{test} = 11, R^2_{pred} = 0.813, \\
&rmsep(ext) = 0.415, \overline{r^2_{m(test)}} = 0.679, \\
&\Delta r^2_{m(test)} = 0.188, \overline{r^2_{m(overall)}} = 0.692, \\
&\Delta r^2_{m(overall)} = 0.161
\end{aligned} \tag{4}$$

The standard errors of the regression coefficients are shown within parentheses. Equation 4 denotes the best QSPR model along with the results of the statistical and validation parameters. GFA was performed with 5000 iterations using 42 compounds as the training set (n_{training}) and validated with the 11 test set compounds (n_{test}). Among 100 models, the one bearing the least Friedman's LOF score (0.279) has been selected as the final model as this fitness function denotes the degree of over fitting of the model. The model could explain 79.4 % of the variance (adjusted coefficient of variation) and could predict 77.8 % of the variance (leave-one-out predicted variance). The prediction error involved in the model development has been shown in terms of the standard error of estimate (s) and the square sum of predictive residual (PRESS) measures which are lower (0.463 and 9.478 respectively) for the best model. The statistical quality of the model can be well explained by the determination coefficient (R^2), the value of which should be as near as possible to 1 for a good model. In the present case, the R^2 value is 0.809 which signifies that the descriptors involved in the final model could well encode the structural parameters of the compounds required to explain response variable. All the regression coefficients are significant at $p < 0.01$ as evidenced from the corresponding t value at df = 38. The F value of the model is significant at $p < 0.01$. The values of all descriptors appearing in Equation (4) are given in Table S3 of Supplementary Materials section.

Apart from all these, the final model (Equation 4) has been validated thoroughly using internal and external validation metrics. The values of Q^2 (0.778) and $r^2_{m(LOO)}$ metrics calculated for the model show encouraging internal validation statistics. However Q^2 calculated using leave-one-out (LOO) cross-validation technique with the calibration set, alone cannot satisfactorily judge the predictability of the model. Hence, $r^2_{m(LOO)}$ (0.685) and $\Delta r^2_{m(LOO)}$ (0.156) were further calculated which can well define the quality of prediction by the developed QSPR model. The prediction ability of the model (Equation 4) has been checked with external validation technique applied on the test set molecules. The R^2_{pred} value, calculated using the predicted and observed activity of the test compounds is equal to 0.813 and is considerably good with respect to the ideal values of 1 (threshold value is 0.5). Moreover, the values of R^2, Q^2 and R^2_{pred} are quite near to each other which support the reliability of the model. Further, the r^2_m metrics for test set *i.e.* $r^2_{m(test)}$ and $\Delta r^2_{m(test)}$ bearing values of 0.679 (> 0.5) and 0.188 (< 0.2) respectively also fulfil the criterion of a model with good prediction potency. Selection of the best model has been done by additionally calculating $r^2_{m(overall)}$ (0.692) and $\Delta r^2_{m(overall)}$ (0.161) metrics utilizing the whole dataset.

The deviations of the prediction data of test set compounds from that of the observed data has been expressed as root mean square error in prediction ($\text{rmsep}_{\text{ext}} = 0.415$). The error involved in the prediction of responses of the training set compounds using cross validation technique has been also marked by the value $\text{rmsep}_{\text{int}}$ (0.475).[45] The $\text{rmsep}_{\text{int}}$ value has been calculated based on leave-one-out predictions values while the $\text{rmsep}_{\text{ext}}$ value has been computed from the predicted values of the test set compounds. Both the values are quite low and are close to each other. The predicted values of individual compounds of the dataset calculated by the best QSPR model (Equation 4) have been provided in Table 1. The proximity of the observed and calculated / predicted responses of the compounds of both the training and test sets have been shown in the scatter plot. (Figure 1)

The absence of chance correlation between the response variable and the descriptors during model development has been analyzed using the Y-randomization test. For the best model (Equation 4), the square of average correlation coefficient of the randomized models (R_r^2) is much less than the actual R^2 value of the non-random model which finally resulted in significant values for the ${}^{c}R_p^2$ parameter (model = 0.678 and process = 0.634). A value of ${}^{c}R_p^2$ more than 0.5 signifies robustness in favor of the model and also for the process involved.

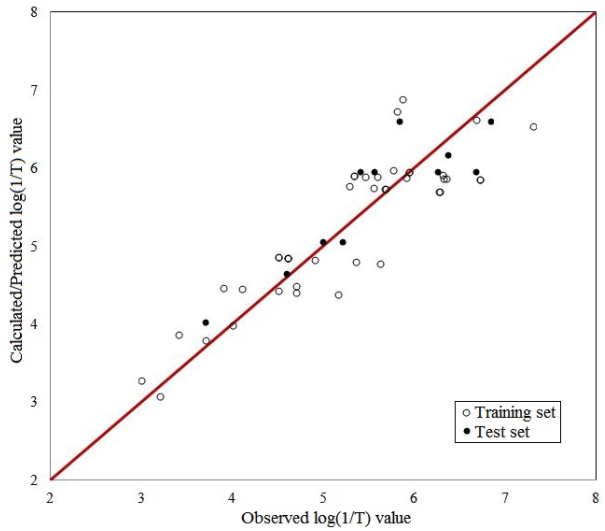

Figure 1. Scatter plot showing both the training and test set compounds.

The best QSPR model (Equation 4) involving odor threshold data consists of three 2D independent variables encoding the essential physicochemical features incorporated in the structures of the compounds under concern. The first descriptor $^{\Sigma\alpha}/_{Nv}$ has the highest regression coefficient among the three variables. It is a first generation ETA descriptor where α individually denotes the size of any atom. Thus, the Σα represents the molecular bulk of a molecule, whereas N_V stands for the total number of non-hydrogen atoms. Moreover, the presence of positive sign in the coefficient of the descriptor clearly denotes that the $\log(1/T)$ increases with increase in the molecular size of alcohols. This has been rightly observed for the highest potent molecule, **C16** (1-decanol) and also for long chain alcohols like 1-dodecanol (**C18**), 2-decanol (**C30**), 3-decanol (**C49**). Since, lipophilicity increases with molecular bulk, it can be said that a molecule should be more lipophilic for it to be a potent odorant.

$\Sigma\beta's$, ranking second among the three descriptors in the value of regression coefficient, modulates the threshold value of odor inversely. Here, the basic parameter β includes the electronic features of molecules where βs denotes the contribution of σ electrons. The descriptor describes the contribution of electronegativity (electron richness) towards prediction of odor threshold values. Thus, lesser is the electronegativity, better is the odorant property as seen for compound **23** (3-methyl-2-butanol). Again, the first generation composite ETA index, η', within the spline term denotes overall topological environment of a molecule. Although, its regression coefficient value is the least among the three descriptors, yet its presence plays a significant role for obtaining a good correlation value. From the equation, it can be inferred that a positive value of the spline term is obtained only for values of the descriptor greater that 0.744 (knot of the spline). Such a condition is essential in order to obtain molecules with significant odor potency. It simultaneously denotes molecular branching and electronic distribution features present in a compound. Taking all the descriptors together, it may be concluded that the unfavorable value of the most important variable, $^{\Sigma\alpha}/_{Nv}$ is responsible for the reduced potency of derivatives like ethanol (**C1**), 2-propanol (**C19**) and 2-butanol (**C20**), although they possess acceptable values for other descriptors.

Domain of Applicability

It is a prime requisite of any QSPR model to determine the applicability domain since the prediction of any compound can be appropriate only if the test compound falls within the domain of applicability of the model. Figure 2 shows the Williams plot by which the applicability domain of the final model (Equation 4) with only ETA descriptors has been ascertained. From the plot, it can be marked out that training set compound, **C1** (Ethanol) lies outside the domain *i.e.*, the leverage value of the referred compound is more than that of the critical hat value (h^*) which is equal to 0.286. Thus, ethanol can be considered as an influential chemical with respect to the developed QSPR model, since, avoiding which can lower the correlation value. Here, the entire test set compounds lie within the applicability domain of the model denoting reliable prediction.

COMPARISON OF THE BEST MODEL WITH PREVIOUSLY REPORTED MODELS

Junkes *et al.*,[46] established a relationship between semi-empirical topological index and odor threshold values

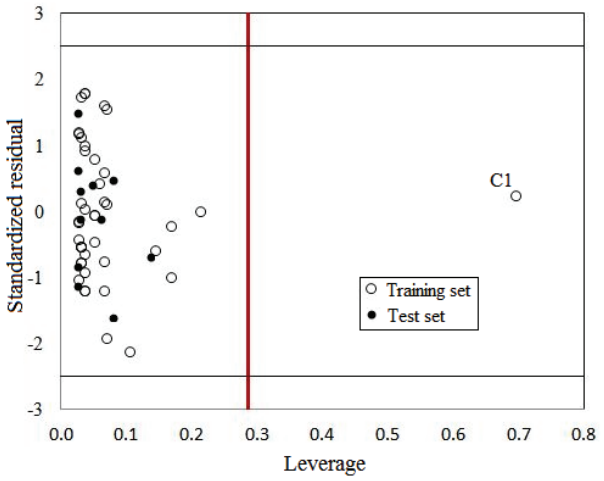

Figure 2. Williams plot denoting the applicability domain of the best QSPR model (Equation (4), model 2).

using the same dataset of the present study, taking 49 aliphatic compounds out of it. Their best model showed to have a R^2 value of 0.714 and Q^2 value corresponding to 0.747. Again, Anker and Jurs[25] developed a QSPR model using the same set of compounds which yielded a squared correlation coefficient value of 0.863 and four compounds were denoted as response outliers by the model. Here, we have developed a QSPR model (Equation 4) with the training set compounds which were selected based on clustering technique, using ETA indices and employing genetic function approximation approach which showed determination coefficient value of 0.809. 42 out of 53 compounds were taken for the model generation and rest were predicted using the best model (R^2_{pred} = 0.813). The final model showed acceptable values for the various validation parameters. The applicability domain for the model has also been reported. Since, the descriptors can easily depict the topological as well as chemical nature of the compounds at a same time; hence, it is useful to utilize this model for *in-silico* prediction of odor threshold. Comparison between different QSPR models on odor threshold, developed by different research group has been summarized in Table S4 of the Supplementary Materials section.

CONCLUSION

The present study successfully demonstrates the application of ETA indices to predict the odor threshold of a series of aliphatic alcohols. The model constructed using GFA (spline) technique showed acceptable internal stability along with good external prediction quality. Even the closeness between the experimental observation data and prediction values of $\log(1/T)$ for all the compounds was reflected in the significant values of r_m^2 metrics. Thus, it is well understood that the ETA parameters possess sufficient diagnostic power in defining the changes in the property values with variation in the structure of the compounds with the –OH functional group which fall within the domain of applicability of the developed model. The mechanistic interpretation of the best QSPR model (Equation 4) suggests that increased lipophilicity and reduced electronegativity potentiate odorant property. Molecular branching and electronic distribution properties of each compound may also be studied further to understand the mechanism behind binding of the odorants to receptors. Hence, ETA descriptors which are simple and easily interpretable requiring less time for calculation can be applied for developing reliable QSPR models for prediction of odor threshold.

Supplementary Materials. – Supporting informations to the paper are enclosed to the electronic version of the article. These data can be found on the website of *Croatica Chemica Acta* (http://public.carnet.hr/ccacaa).

Acknowledgements. Financial assistance from the UGC, New Delhi and ICMR, New Delhi is thankfully acknowledged.

REFERENCES

1. M. Benzo, G. Gilardoni, C. Gandini, G. Caccialanza, P. V. Finzi, G. Vidari, S. Abdod, and P. Layedra, *J. Chromatogr. A* **1150** (2007) 131–135.
2. J. A. Nicell, *Atmos. Environ.* **37** (2003) 4955–4964.
3. W. S. Cain, R. Schmidt, and A. A. Jalowayski, *Int. Arch. Occup. Environ. Health* **80** (2007) 721–731.
4. B. Siegmund and B. Pollinger-Zierler, *J. Agric. Food Chem.* **54** (2006) 5984–5989.
5. J. E. Cometto-Muñiz, W. S. Cain, M. H. Abraham, and J. Gil-Lostes, *Physiol. Behav.* **95** (2008) 658–667.
6. X. Deupi and B. Kobilka, *In Mechanisms and Pathways of Heterotrimeric G Protein Signaling*, Elsevier Academic Press Inc, San Diego, 2007, pp. 137–166.
7. X. Zeng, H. Wang, and Y. Wang, *Chemosphere* **86** (2012) 619–625.
8. I. Mitra, A. Saha, and K. Roy, *Eur. J. Med. Chem.* **45** (2010) 5071–5079.
9. A. P. Worth, A. Bassan, J. De Bruijn, A. G. Saliner, T. Netzeva, G. Patlewicz, M. Pavan, I. Tsakovska, and S. Eisenreich *SAR QSAR Environ. Res.* **18** (2007) 111–125.
10. OECD Principles for the Validation of (Q)SARs. http://www.oecd.org/dataoecd/33/37/37849783
11. OECD, Environment Directorate, Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology. http://www.olis.oecd.org/olis/2004doc.nsf/LinkTo/NT00009192/$FILE/JT00176183
12. R. Benigni and R. Zito, *Mutat. Res.* **566** (2004) 49–63.
13. Z. Bankowski and N. Howard-Jones, *International Guiding Principles for Biomedical Research Involving Animals* [Committee of International Organizations of Medical Science (CIOMS), Geneva,] (1986).
14. P. K. Ojha and K. Roy, *Chemom. Intell. Lab. Syst. Volume* **109** (2011) 146–161.
15. S. Kar and K. Roy, *J. Hazard. Mater.* **177** (2010) 344–351.
16. F. Luan, H. T. Liu, Y. Y. Wen, and X. Y. Zhang, *Flavour Fragr. J.* **23** (2008) 232–238.
17. H. Du, J. Wang, Z. Hu, M. Liu, and X. Yao, *Sensor Actuat B-Chem.* **138** (2009) 55–63.
18. K. Roy and G. Ghosh, *Internet Electron. J. Mol. Des.* **2** (2003) 599–620.
19. K. Roy and R. N. Das, *On Extended Topochemical atom (ETA) Indices for QSPR studies*, in: E. A. Castro and A.K. Hagi (Eds.) *Advanced methods and applications in chemoinformatics: research progress and new applications,* IGI Global, Hershey, 2011, pp. 380–411.
20. R. N. Das and K. Roy, *Struct. Chem.* **24** (2013) 303–331.
21. K. Roy and H. Kabir, *Chem. Eng. Sci.* **73** (2012) 86–98.
22. D. K. Pal, C. Sengupta and A. U. De, *Indian J. Chem.* **27B** (1988) 734–739.
23. K. Roy and A. Saha, *Indian J. Chem.* **44B** (2005) 1693–1707.
24. C. W. Yap, *J. Comput. Chem.* **32** (2011) 1466–1474.
25. L. S. Anker and P. C. Jurs, *Anal. Chem.* **62**, (1990) 2676–2684.
26. Cerius 2 Version 4.10, Accelrys Inc., San Diego, CA, USA. Software. (2005) http://www.accelrys.com.
27. DRAGON Version. 6, TALETE srl, Italy, http://www.talete.mi.it/products/dragon_molecular_descriptors.htm.
28. R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, Wiley, VCH, Weinheim, 2000.
29. P. P. Roy, J. T. Leonard, and K. Roy, *Chemom. Intell. Lab. Syst.* **90** (2008) 31–42.

30. B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, Arnold Press, London, 2001.
31. SPSS 9.0, SPSS Inc, Chicago, **2011**; http://www.spss.com
32. G. W. Snedecor and W. G. Cochran, *Statistical Methods*, Oxford & IBH Publishing Co. Pvt. Ltd. New Delhi, 1967, pp. 381–418.
33. J. Camacho, J. Pico, and A. Ferrer, *Chemometr. Intell. Lab.* **100** (2010) 48–56.
34. D. Rogers and A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **4** (1994) 854–866.
35. W. J. Dunn and D. Rogers, *Genetic Algorithms*, in: J. Devillers (Eds.), *Molecular Modeling*, Academic Press, London, 1996, pp. 109–130.
36. J. H. Friedman, *ANN. STAT.* **19** (1991) 1–141.
37. K. Roy and I. Mitra, *Comb. Chem. High T. Scr.* **14** (2011) 450–474.
38. P. K. Ojha, I. Mitra, R. N. Das, and K. Roy, *Chemmom. Intell. Lab. Sys.* **107** (2011) 194–205.
39. A. Golbraikh and A. Tropsha, *J. Mol. Graph. Mod.* **20** (2002) 269–276.
40. K. Roy, I. Mitra, S. Kar, P. K. Ojha, R. N. Das, and H. Kabir, *J. Chem. Inf. Model.* **52** (2012) 396–408.
41. A.Tropsha, P. Gramatica, and V. K. Gombar, *QSAR Comb.Sci.* **22** (2003) 69–77.
42. I. Mitra, A. Saha, and K. Roy, *Mol. Simulat.* **36** (2010) 1067–1079.
43. P. Gramatica, *QSAR Comb. Sci.* **26** (2007) 694-701.
44. STATISTICA Version 7.1, STATSOFT Inc. USA; http://www.statsoft.com/
45. V. Consonni, D. Ballabio, and R. Todeschini, *J. Chemometrics* **24** (2010) 194–201.
46. B. D. S. Junkes, A. C. S. Arruda, R. A. Yunes, L. C. Porto, and V. E. F. Heinzen, *J. Mol.Model.* **11** (2005) 128–134.

Supplementary Materials

Quality measures in fitting of a QSAR model

A QSAR model is needed to be checked for its quality before applying it for screening of new molecules. Several statistical parameters are available for assessing the quality of the model. Initially the acceptability of a QSAR model depends upon three statistical parameters: (i) standard error of estimate (s), (ii) squared correlation coefficient (R^2) and (iii) explained variance (R_a^2) based on the MLR technique. The error in the estimation of individual activity values of the compounds under study using the MLR method can be quantified based on their residual data. The standard error of estimate (SEE or s) for the residuals is calculated by taking the root-mean square of the residuals. The standard error of the estimate is a measure of the accuracy of fitting. Lower values of SEE correspond to improved model acceptability.

$$s = \sqrt{\frac{\sum \left(Y_{obs} - Y_{calc}\right)^2}{n - p - 1}} \qquad \text{(S1)}$$

In Eq. S1, Y_{obs} and Y_{calc} are the actual and estimated scores respectively, while n is the number of scores and p is the number of descriptors. Again, variation in the data is quantified by the correlation coefficient (R), which measures how closely the observed data tracks the fitted regression line. An R^2 of 0 means that there is no relationship between activity and the parameters selected for the study, while an R^2 of 1 means a 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.

$$R^2 = 1 - \frac{\sum (Y_{obs} - Y_{calc})^2}{\sum (Y_{obs} - \overline{Y_{training}})^2} \qquad \text{(S2)}$$

In Eq. S2, $\overline{Y_{training}}$ is the mean observed activity of the training set compounds. Previously, QSAR models were only based on the fitting description of the mathematical equation using the correlation coefficient. The prime drawbacks of the R^2 parameter lies in the facts that it does not provide any information on whether: (i) the independent variables are a true cause of the changes in the dependent variable, (ii) the correct regression was used, (iii) the most appropriate set of independent variables has been chosen, (iv) the model might be improved by using transformed versions of the existing set of independent variables and (v) whether any collinearity exists in the data or not. However, adjusted R^2 (R^2_a representing Eq. S3) is a modification of R^2 that adjusts for the number of explanatory terms in a model. Unlike R^2, the R^2_a increases only if the new term improves the model more than would be expected by chance. The adjusted R^2 can be negative, and will always be less than or equal to R^2.

$$R_a^2 = \frac{(n-1)R^2 - p}{n - p - 1} \qquad \text{(S3)}$$

In Eq. S3, n is the number of compounds and p is the number of descriptors. 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. The addition of descriptors to the model increases the value of R^2, but this may not indicate an improvement in model quality. So to optimally determine the predictive quality, the models are required to be further validated using various validation techniques.

Validation strategies

Both internal and external validation statistics constitute the primary methods for validation of the developed QSAR models. Both the methods have been widely used by different groups of researchers for assessing the predictive ability of the developed model. Several metrics are used to check the predictivity of the QSAR models. For the validation of QSAR models, three strategies are primarily adopted: (i) internal validation using the training set molecules, and (ii) external validation based on the test set compounds.

Internal validation (Leave-one-out cross-validation)

Internal validation deals with validation of a QSAR model based on the molecules involved in the QSAR model building process (training set data). In this technique, one compound is eliminated from the data set at random in each cycle and the model is built using the rest of the compounds. The model thus formed is used for predicting the activity of the eliminated compound. The process is repeated until all the compounds are eliminated once. On the basis of the predicting ability of the model, the predicted residual sum of squares (PRESS) (Eq. S4), the value of standard deviation of error of prediction (SDEP) (Eq. S5) and the cross-validated R^2 (Q^2) metrics (Eq. S6) for the model are determined. The higher is the value of Q^2 (more than 0.5) the better is the model predictivity.

$$PRESS = \sum (Y_{obs(train)} - Y_{pred(train)})^2 \quad \text{(S4)}$$

$$SDEP = \sqrt{\frac{PRESS}{n}} \quad \text{(S5)}$$

$$Q^2 = 1 - \frac{\sum (Y_{obs(train)} - Y_{pred(train)})^2}{\sum (Y_{obs(train)} - \overline{Y_{training}})^2} \quad \text{(S6)}$$

In the above equations, $Y_{obs(train)}$ and $Y_{pred(train)}$ refer to the observed activity and the predicted activity of the training set molecules calculated based on the LOO technique. From Eq. S6, it can be stated that the mean response value of the training set molecules and the distance of the mean from the response values of the individual molecules play a crucial role in determining the value of Q^2. As the value of the denominator ($\sum (Y_{obs(train)} - \overline{Y}_{training})^2$) on the right hand side of the equation increases, the value of Q^2 also increases. Thus, even for large difference in the predicted and observed response values, acceptable Q^2 values may be obtained if the molecules exhibit a significantly wide range of response data. Hence, a large value of Q^2 does not necessarily indicate that the predicted activity data lies in close proximity to the observed ones although there may exist a good overall correlation between the values. Thus to obviate this error and to

better indicate the model predictive ability, the r_m^2 metrics [$\overline{r_m^{2}}_{(LOO)}$ and $\Delta r_m^{2}{}_{(LOO)}$] (Eqs. S7 and S8) for internal validation, introduced by our research group, are calculated.

$$\overline{r_m^{2}} = \frac{(r_m^{2} + r'^{2}_m)}{2} \quad \text{(S7)}$$

$$\Delta r_m^{2} = | r_m^{2} - r'^{2}_m | \quad \text{(S8)}$$

Here, $r_m^{2} = r^2 \times (1 - \sqrt{(r^2 - r_0^{2})})$ and $r'^{2}_m = r^2 \times (1 - \sqrt{(r^2 - r'^{2}_0)})$. Squared correlation coefficient values between the observed and predicted values of the test set compounds with intercept (r^2) and without intercept (r^2_0) were calculated for determination of r_m. Change of the axes gives the value of r'^{2}_0 and the r'^{2}_m metric was calculated based on the value of r'^{2}_0.

External validation

Despite being the most popular technique for validation of a QSAR model, internal validation is not the sufficient condition for the model to have a high predictive power. The cross-validation technique only provides a reasonable approximation of the ability of the model to predict the activity of new molecules. Thus to precisely judge the external predictive potential of the developed model, a sufficiently large dataset demands proper external validation by removing a portion of the whole dataset as test set. The R^2_{pred} (Eq. S9) parameter exclusively reflects the degree of correlation between the observed and predicted property data.

$$R^2_{pred} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \overline{Y_{training}})^2} \quad \text{(S9)}$$

Here, $Y_{obs(test)}$ is the observed activity of the test set compounds and $Y_{pred(test)}$ is the predicted activity of the test set compounds. R^2_{pred} value for an acceptable model should be greater than 0.5 (maximum value 1).

In order to determine the proximity between the observed and predicted response data, the r_m^2 metrics, viz. $\overline{r_m^{2}}_{(test)}$ and $\Delta r_m^{2}{}_{(test)}$ (similar to those employed for internal validation) are calculated for the test molecules.

External predictive ability of a QSAR model may further be determined by a comparison of the observed activity values and the model predictions (activity calculation of molecules not included in development of the model) through calculation of a parameter referred to as root mean square error in prediction (rmsep) depicted in Eq. S10.

$$rmsep = \sqrt{\frac{\sum (y_{obs(test)} - y_{pred(test)})^2}{n_{ext}}} \quad \text{(S10)}$$

Here, n_{ext} refers to the number of test set compounds.

Table S1. Pool of descriptors calculated using different software

Chemometric software	Category of descriptors	Descriptor
Cerius2 software[26]	Spatial	PMI-mag, Jurs_SASA, Jurs_PPSA_1, Jurs-PNSA_1, Jurs-DPSA_1, Jurs_PPSA_2, Jurs_PNSA_2, Jurs_DPSA_2, Jurs_PPSA_3, Jurs_PNSA_3, Jurs_DPSA_3, Jurs_FPSA_2, Jurs_WPSA_1, Jurs_WNSA_1, Jurs_WPSA_2, Jurs_WNSA_2, Jurs_WPSA_3, Jurs_WNSA_3, Jurs_RPCS, Jurs_RNCS, Jurs_TPSA, Jurs_TASA, Jurs_RPSA, Jurs_RASA, Vm, Area, RadOfGyration, Shadow-XY, Shadow-XZ, Shadow-YZ, Shadow-XYfrac, Shadow-XZfrac, Shadow-YZfrac, Shadow-nu, Shadow-Xlength, Shadow-Ylength, Shadow-Zlength
	Structural	Rotlbonds, Hbond acceptor, Chiral centers
	Thermodynamic	LogP, MR, AlogP, MolRef, AlogP98, Atype_C_1, Atype_C_2, Atype_C_3, Atype_C_6, Atype_C_8, Atype_H_46, Atype_H_47, Atype_H_52.
	Electronic	HOMO, Sr, LUMO, Dipole-mag, Apol
	Topological	S_sCH$_3$, S_ssCH$_2$, S_sssCH, S_sOH, JX, Zagreb, SC-0, SC-1, SC-2, SC-3_P, SC-3_C, ${}^{0}\chi$, ${}^{1}\chi$, ${}^{2}\chi$, ${}^{3}\chi_p$, ${}^{3}\chi_c$, ${}^{0}\chi^{v}$, ${}^{1}\chi^{v}$, ${}^{2}\chi^{v}$, ${}^{3}\chi_p^{v}$, ${}^{3}\chi_c^{v}$, Wiener
PaDEL-Descriptor software[24]	Topological	$\Sigma\alpha/N_v$, $[\Sigma\alpha]_P/\Sigma\alpha$, $[\Sigma\alpha]_Y/\Sigma\alpha$, $[\Sigma\alpha]_X/\Sigma\alpha$, $\Sigma\beta'$, $\Sigma\beta'_s$, η', η'_F, $[\eta']^{local}$, $[\eta'_F]^{local}$, η'_B , $\Delta\varepsilon_A$, $\Delta\varepsilon_C$, $\Delta\varepsilon_D$, $\Delta\psi_A$, $\Delta\beta'$, $\Sigma\beta'_{ns(\delta)}$
Dragon software[27]	Constitutional	MW, AMW, RBN, RBF
	Functional group count	nCp, nCs, nCt, nCq, nOHp, nOHs, nOHt

Table S2. Brief description of ETA descriptors

The ETA indices provide potential information about electronic features and the contribution of size, shape, branching, and functionality of a molecule. The second generation indices can have a better power to encode the structural features responsible for electron richness, unsaturation, polar surface and ability of hydrogen-bond formation of a given molecule. The variables are denoted by some basic parameters such as α which is related to the size or bulk, ε which provides information about electronegativity of atoms and β that is related to electronic contribution. The following are the ETA indices that have been utilized in the present work.

Sl. No.	Generation	Notation	Significance
1	1st generation ETA	$\Sigma\alpha/N_v$	Measure of molecular bulk
2		$[\Sigma\alpha]_P/\Sigma\alpha$	Contribution of terminal substituents on the carbon skeleton
3		$[\Sigma\alpha]_Y/\Sigma\alpha$	Contribution of branching
4		$[\Sigma\alpha]_X/\Sigma\alpha$	Contribution of quaternary atom
5		$\Sigma\beta'$	A measure of contribution of electron richness and electronegative atom
6		$\Sigma\beta'_s$	Contribution of electronegativity
7		η'	Provides information on the overall topological environment in a molecule
8		η'_F	Contribution of heteroatom and multiple bonds
9		$[\eta']^{local}$	Measures the contribution of locally bonded atoms where local boding refers to the atoms connected with single covalent bonds
10		$[\eta'_F]^{local}$	Measure of local functionality contribution
11		η'_B	A measure of branching contribution relative to the molecular size
12	2^{nd} generation ETA	$\Delta\varepsilon_A$	A measure of contribution of unsaturation and electronegative atom count
13		$\Delta\varepsilon_C$	A measure of contribution of electronegativity
14		$\Delta\varepsilon_D$	A measure of contribution of hydrogen bond donor atoms
15		$\Delta\psi_A$	A measure of hydrogen bonding propensity of the molecules
16		$\Delta\beta'$	A measure of relative unsaturation content relative to molecular size
17		$\Sigma\beta'_{ns(\delta)}$	A measure of lone electrons entering into resonance relative to molecular size

Table S3. Value of the descriptors appearing in Eq. (4)

Sl. No.	Log(1/T)	$\Sigma\alpha/N_v$	$\Sigma\beta'_s$	η'	$<0.74427-\eta'>$

1	3.2	0.44444	0.41667	0.46247	0.2818
2	4	0.45833	0.4375	0.58753	0.15674
3	4.7	0.46667	0.45	0.68469	0.05958
4	3.7	0.46667	0.45	0.74216	0.00211
5	4.5	0.47222	0.45833	0.765	0
6	4.6	0.47222	0.45833	0.82341	0
7	4.7	0.47222	0.45833	0.91998	0
8	4.9	0.47619	0.46429	0.83367	0
9	5	0.47619	0.46429	0.89127	0
10	4.6	0.47619	0.46429	0.89529	0
11	4.6	0.47619	0.46429	0.89024	0
12	5.41	0.47917	0.46875	0.89373	0
13	6.68	0.47917	0.46875	1.04459	0
14	6.32	0.48148	0.47222	0.94714	0
15	6.37	0.48333	0.475	0.99526	0
16	7.3	0.48485	0.47727	1.03904	0
17	5.84	0.48611	0.47917	1.07922	0
18	5.87	0.48718	0.48077	1.11636	0
19	3	0.45833	0.4375	0.64801	0.09626
20	3.4	0.46667	0.45	0.74427	0
21	3.7	0.46667	0.45	0.84569	0
22	3.9	0.47222	0.45833	0.82145	0
23	5.16	0.47222	0.45833	0.88571	0
24	5.62	0.47619	0.46429	0.88674	0
25	4.6	0.47619	0.46429	0.95452	0
26	4.5	0.47619	0.46429	0.9466	0
27	6.28	0.47917	0.46875	0.94357	0
28	6.72	0.48148	0.47222	0.99401	0
29	6.31	0.48333	0.475	1.03945	0
30	6.68	0.48485	0.47727	1.08083	0
31	5.21	0.47619	0.46429	0.95275	0
32	4.5	0.47619	0.46429	0.99033	0
33	6.27	0.47917	0.46875	1.05383	0
34	5.67	0.47917	0.46875	1.07028	0
35	5.28	0.47917	0.46875	0.94979	0
36	6.84	0.48611	0.47917	1.12851	0
37	5.56	0.47917	0.46875	1.00656	0
38	5.35	0.47619	0.46429	1.0528	0
39	5.68	0.47917	0.46875	1.01103	0
40	4.1	0.47222	0.45833	0.82471	0
41	5.55	0.47917	0.46875	0.9484	0
42	6.26	0.47917	0.46875	1.00823	0
43	5.903	0.48148	0.47222	0.99899	0
44	5.46	0.48148	0.47222	1.05632	0

45	6.72	0.48148	0.47222	1.09211	0
46	6.35	0.48148	0.47222	1.0613	0
47	5.59	0.48148	0.47222	1.05198	0
48	5.939	0.48333	0.475	1.04441	0
49	5.81	0.48485	0.47727	1.0857	0
50	5.33	0.48148	0.47222	1.00112	0
51	5.33	0.48148	0.47222	1.05493	0
52	5.76	0.48333	0.475	1.04694	0
53	5.939	0.48333	0.475	1.04775	0

Table S4. Comparison with previously reported work

Ref.	Modeling technique	Descriptors	$n_{training}$	R^2	Q^2	n_{test}	R^2_{pred}
Jukes *et al.*[46]	MLR	Semi-empirical topological index (I_{ET})	49	0.765	0.558	-	-
Anker and Jurs[25]	MLR	Spatial (CPSA 1, ENVR 59), Topological (MOLC 8, MOMI 1)	49	0.863	-	-	-
Present work	GFA-spline (equation 5)	Extended topochemical atom indices ($\Sigma\alpha/N_{v,}$ $\Sigma\beta'_{s,}$, η')	42	0.809	0.778	11	0.813

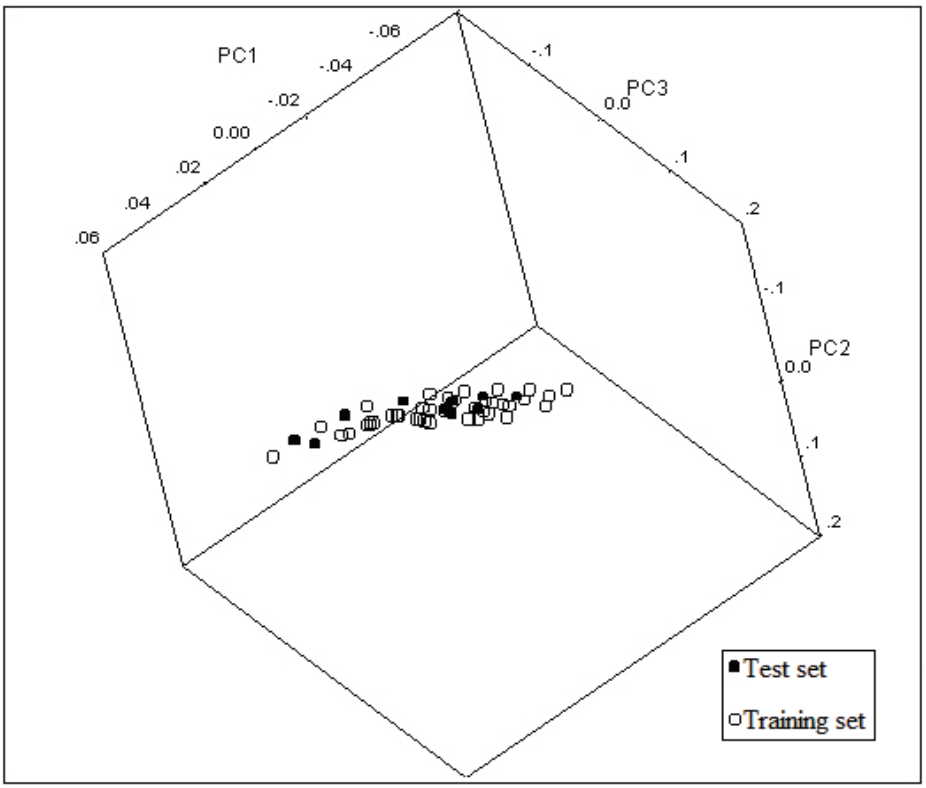

Figure S1. Principle component analysis plot with the first three principle components generated by factor analysis