J Mol Model (2006) 12: 306–316
DOI 10.1007/s00894-005-0033-7

ORIGINAL PAPER

Kunal Roy · Gopinath Ghosh

QSTR with extended topochemical atom (ETA) indices. VI. Acute toxicity of benzene derivatives to tadpoles (*Rana japonica*)

Received: 19 April 2005 / Accepted: 25 July 2005 / Published online: 26 October 2005

Abstract Quantitative structure-toxicity relationship (QSTR) studies have proved to be a valuable approach in research on the toxicity of organic chemicals for ranking chemical substances with respect to their potential hazardous effects on living systems. With this background, we have modeled here the acute lethal toxicity of 51 benzene derivatives with recently introduced extended topochemical atom (ETA) indices [Roy and Ghosh, Internet Electron J Mol Des 2:599–620 (2003)]. We also compared the ETA relations with non-ETA models derived from different topological indices (Wiener *W*, Balaban *J*, flexibility index ϕ, Hosoya *Z*, Zagreb, molecular connectivity indices, E-state indices and kappa shape indices) and physicochemical parameters (*AlogP*98, *MolRef*,*H_bond_donor* and *H_bond_acceptor*). Genetic function approximation (GFA) and factor analysis (FA) were used as the data-preprocessing steps for the development of final multiple linear regression (MLR) equations. Principal-component regression analysis (PCRA) was also used to extract the total information from the ETA/non-ETA/combined matrices. All the models developed were cross-validated using leave-one-out (LOO) and leave-many-out techniques. The summary of the statistics of the best models is as follows: (1) FA-MLR: ETA model- Q^2 (LOO) = 0.852, R^2 = 0.894; non-ETA model- Q^2 = 0.782, R^2 = 0.835; ETA + non-ETA model-Q^2 = 0.815, R^2 = 0.859. (2) GFA-MLR: ETA model-Q^2 = 0.847, R^2 = 0.915; non-ETA model-Q^2 = 0.863, R^2 = 0.898; ETA + non-ETA model-Q^2 = 0.859, R^2 = 0.893. 3. PCRA: ETA model-Q^2 = 0.864, R^2 = 0.901; non-ETA model- Q^2 = 0.866, R^2 = 0.922; ETA + non-ETA model-Q^2 = 0.846, R^2 = 0.890. The statistical quality of the ETA models is comparable to that of non-ETA models. Again, use of non-ETA descriptors in addition to ETA descriptors does not increase the statistical acceptance of the relations significantly. The predictive potential of these models was better than that of the previously reported models using physicochemical parameters [Huang et al., Chemosphere 53:963–970 (2003)]. The relations from ETA descriptors suggest a parabolic dependence of the toxicity on molecular size. Furthermore, the toxicity increases with functionality contribution of chloro substituent and decreases with those of methoxy, hydroxy, carboxy and amino groups. This study suggests that ETA parameters are sufficiently rich in chemical information to encode the structural features that contribute significantly to the acute toxicity of benzene derivatives to *Rana japonica*.

Keywords QSTR · QSAR · ETA · TAU · VEM · Factor analysis · Genetic function approximation

K. Roy (✉) · G. Ghosh
Division of Medicinal and Pharmaceutical Chemistry,
Department of Pharmaceutical Technology,
Drug Theoretics and Cheminformatics Lab,
Jadavpur University, Kolkata 700 032, India
E-mail: kroy@pharma.jdvu.ac.in

Introduction

The effect of hazardous chemicals and pollutants on the ecosystem is a matter of great concern considering the fact that although large number of chemical compounds (in 10s of 1000s) are in commercial use, relatively few of these have been subjected to adequate assessment for their hazardous environmental properties. Accumulating evidence suggests that humans and domestic and wildlife species have suffered adverse health consequences from exposure to environmental chemicals [1]. Animal testing is still considered essential to the support of risk assessment, but is often too costly and time consuming to be applied to the full range of chemicals for which some level of toxicological screening is necessary and desired [2]. Currently, there are ecotoxicological data available for < 1% of compounds. Thus, the European Union Commission's Scientific committee on toxicity, ecotoxicity, and environment (CSTEE) has recommended the use of (Q)SAR models and precautions to prioritize further risk assessment of approximately 4,500 compounds and their adjuvants [2]. Faced with the task of screening a large number of chemicals, for an increasing array of toxicity endpoints, using

limited resources, quantitative structure-activity relationships (QSARs) have been used in many diverse problem settings as a complement to experimental data [2]. QSARs have emerged as an indispensible tool for predicting the ecotoxicological hazard of new chemicals. The regulatory agencies must rely on QSAR techniques, as these can predict potential ecotoxicological hazards rapidly at minimum cost [3]. The US Environmental Protection Agency (EPA) designed assessment tools for evaluation of risk (ASTER), which is an integration of AQUIRE (database of aquatic toxicity) and QSAR (database of physicochemical properties and QSAR models) to assist regulators in producing assessments [4]. Apart from prediction of ecological and human health effects, QSARs are being used to help industry design safer chemicals for commercial use [5]. QSARs have also been used in exploring the mechanisms of toxic actions of chemicals [6]. Different QSAR models, descriptors and statistical methods have been used to model toxicity data of diverse chemicals by different groups of workers. A paper on E-state modeling of fish toxicity independent of 3D structure information was published by Rose and Hall [7]. In this paper, the authors showed the utility of the E-state index in toxicity modeling with direct physicochemical significance. Mazzatorta et al. [8] have modeled the toxicity of 562 organic chemicals using neural and fuzzy-neural networks. Principal component analysis was used as a classification tool for the toxicity data of dangerous chemicals by Vighi et al. [9]. Bask et al. [10] have used H-QSAR for predicting the toxicity of chemicals. Devillers [11] has derived a general model for predicting acute toxicity of pesticides.

Quantitative structure-toxicity relationship (QSTR) studies, having proved to be a valuable approach in research on the toxicity of organic chemicals for ranking the chemical substances with respect to their potential hazardous effects on living systems, there is an urgent need to develop newer descriptors to encode molecular features and chemical information from different dimensions. With this background, we have recently introduced [12–16] extended topochemical atom (ETA) indices as an extension of the TAU concept in the valence electron mobile (VEM) environment [17–26], and modeled different toxicity data (phenol toxicity [12], fish toxicity [13], and nitrobenzene toxicity [14]) to establish the utility of ETA indices in modeling studies. Very recently, we reported modeling of the acute toxicity of 56 phenylsulfonyl carboxylates to *Vibrio fischeri* using factor analysis and principal-component regression analysis [15] and also using genetic function approximation (GFA) [16]. In our present work, we have modeled acute toxicity of 51 benzene derivatives to tadpoles (*Rana joponica*) with ETA indices using GFA and factor analysis as the data preprocessing steps for the development of final multiple linear regression (MLR) equations. The best model with ETA indices was compared with non-ETA models derived from different topological and selected physicochemical indices and also with models reported previously [27].

Materials and methods

Definitions of some of the basic parameters used in the ETA scheme are given below.

The core count of a non-hydrogen vertex [α] is defined as [12]:

$$\alpha = \frac{Z - Z^{v}}{Z^{v}} \frac{1}{\mathrm{PN} - 1}. \quad (1)$$

In Eq. 1, Z and Z^{v} represent atomic number and valence electron number respectively, while PN denotes period number. The hydrogen atom being considered as the reference, α for hydrogen is taken to be zero. Again, another term ε (a measure of electronegativity) is defined [12] in the following manner:

$$\varepsilon = -\alpha + 0.3Z^{V}. \quad (2)$$

It is interesting to note that α values of different atoms commonly found in organic compounds have a high correlation ($r = 0.946$) [12] with (uncorrected) van der Waals volume while ε correlates well ($r = 0.937$) with Pauling's electronegativity scale [12].

The VEM count β of the ETA scheme is defined as

$$\beta = \sum x\sigma + \sum y\pi + \delta. \quad (3)$$

In the above equation, δ is a correction factor of value 0.5 per atom with a lone pair of electrons capable of resonance with an aromatic ring (e.g., nitrogen of aniline, oxygen of phenol, etc.). For calculation of the VEM count, the contribution (x) of a sigma bond (σ) between two atoms of similar electronegativity ($\Delta\varepsilon \leq 0.3$) is considered to be 0.5, and for a sigma bond (σ) between two atoms of different electronegativity ($\Delta\varepsilon > 0.3$), it is considered to be 0.75. Again, in the case of π-bonds (π), contributions (y) are considered depending on the type of double bond: (1) for π-bond between two atoms of similar electronegativity ($\Delta\varepsilon \leq 0.3$), y is taken to be 1; (2) for π-bond between two atoms of different electronegativity ($\Delta\varepsilon > 0.3$) or for conjugated (non-aromatic) π-system, y is considered to be 1.5; (3) for aromatic pi system, y is taken as 2.

The VEM vertex count $\gamma_i$of the ith vertex in a molecular graph is defined as

$$\gamma_i = \frac{\alpha_i}{\beta_i}. \quad (4)$$

In the above equation, α_i stands for the α value for the ith vertex and β_i stands for VEM count considering all bonds connected to the atom i and its lone pair of electrons (if any).

Finally, the composite index η is defined in the following manner:

$$\eta = \sum_{i<j} \left[\frac{\gamma_i \gamma_j}{r_{ij}^2}\right]^{0.5}. \quad (5)$$

In Eq. 5, r_{ij} stands for the topological distance between the ith and jth atoms. Again, when all heteroatoms and multiple bonds in the molecular graph are replaced by carbon and single bonds, respectively, the corresponding molecular graph may be considered as the reference alkane and the corresponding composite index value is designated η_R. Considering functionality as the presence of heteroatoms (atoms other than carbon or hydrogen) and multiple bonds, the functionality index η_F may be calculated as $\eta_R - \eta$. To avoid dependence of functionality on vertex count or bulk, we have defined [12] another term η'_F as η_F / N_V, N_V being the count of non-hydrogen vertices. Again, one can determine the contribution of a particular position, vertex or substructure to functionality in the following manner:

$$[\eta]_i = \sum_{j \neq i} \left[\frac{\gamma_i \gamma_j}{r_{ij}^2} \right]^{0.5} . \quad (6)$$

In Eq. 6, $[\eta]_i$ stands for the contribution of the ith vertex to η. Similarly, the contribution of the ith vertex $[\eta_R]_i$ to η_R can be computed. Contribution of the ith vertex $[\eta_F]_i$ to functionality may be defined as $[\eta_R]_i - [\eta]_i$. To avoid dependence of this value on N_V, a related term $[\eta'_F]_i$ was defined [12] as $[\eta_F]_i / N_V$.

Again, considering only bonded interactions ($r_{ij} = 1$), the corresponding composite index is written as η^{local}.

$$\eta^{local} = \sum_{i<j, r_{ij}=1} (\gamma_i \gamma_j)^{0.5} \quad (7)$$

In a similar way, η_R^{local} for the corresponding reference alkane may also be calculated. The local functionality contribution (without considering global topology), η_F^{local}, may be calculated as $\eta_R^{local} - \eta^{local}$.

The branching index η_B can be calculated as $\eta_N^{local} - \eta_R^{local} + 0.086\ N_R$, where N_R stands for the number of rings in the molecular graph of the reference alkane. The N_R term in the branching index expression represents a correction factor for cyclicity. η_N^{local} indicates the η value of the corresponding normal alkane (straight chain compound of same vertex count obtained from the reference alkane), which may be conveniently calculated as (when $N_V \geq 3$):

$$\eta_N^{local} = 1.414 + (N_V - 3)0.5 \quad (8)$$

To calculate the branching contribution relative to the molecular size, another term η'_B has been defined as η_B / N_V.

In the present communication, the utility of ETA parameters is demonstrated through a QSTR study taking acute toxicity of benzene derivatives to the tadpole (*R. japonica*) [27] as the model dataset (Table 1). Definitions of important ETA parameters are given in Table 2. GFA [28, 29] and factor analysis (FA) [30, 31] were performed as the data preprocessing step for identifying important descriptors for the final multiple regression analysis.

The GFA technique [28, 29] was used to generate a population of equations rather than one single equation for correlation between the toxicity and descriptors. GFA involves the combination of the multivariate adaptive regression splines (MARS) algorithm with a genetic algorithm to evolve a population of equations that best fit the training set data. It provides an error measure, called the lack of fit (LOF) score that automatically penalizes models with too many features. This is done as follows: (1) an initial population of equations is generated by random choice of descriptors; (2) pairs from the population of equations are chosen at random and "crossovers" are performed and progeny equations are generated; (3) it is better at discovering combinations of features that take advantage of correlations between multiple features; (4) the fitness of each progeny equation is assessed by the lack-of-fit (LOF) measure; (5) it can use a larger variety of equation term types in construction of its models; (6) the fitness of a new progeny equation is preserved if it is better. The models with proper balance of all statistical terms are used to explain variance in the biological activity. A distinctive feature of GFA is that it produces a population of models (e.g., 100) instead of generating a single model, as do most other statistical methods. The range of variations in this population gives added information on the quality of fit and importance of the descriptors. The GFA study was done using the GFA module in the QSAR + environment of the Cerius2 software [32]. Apart from the number of crossovers (10,000), all other default settings were used for the analysis (linear terms, smoothness factor = 1, mutation probability for adding new term = 50%).

For the purpose of factor analysis, the data matrix consisting of the ETA/non-ETA/combined descriptors was subjected to principal-component factor analysis using the SPSS software [33]. The principal objectives of factor analysis are to display multidimensional data in a space of lower dimensionality with minimal loss of information and to extract basic features behind the data with the ultimate goal of interpretation and/or prediction. The factors were extracted by the principal-component method and then rotated by VARIMAX rotation to obtain Thurston's simple structure. Only variables with non-zero loadings in such factors where biological activity also has non-zero loading were considered important in explaining variance of the activity. Further, variables with non-zero loadings in different factors were combined in regression equations. An attempt was also made to perform PCRA [31] taking factor scores as predictor variables. In this case, the principal components serve as latent variables. PCRA has the advantage that colinearities among X variables are not a disturbing factor and that the number of variables included in the analysis may exceed the number of observations [31]. In PCRA, all descriptors are assumed to be important while the aim of factor analysis is to identify relevant descriptors.

Table 1 Observed, calculated and predicted toxicity of benzene derivatives to the tadpole (*R. japonica*)

S. no.	Compound	Obs[a]	Cal[b]	Pred[b] from LOO	Pred[b] from L-20%-O	Cal[c]	Pred[c] from LOO	Pred[c] from L-20%-O	Cal[d]	Pred[d] from LOO	Pred[d] from L-20%-O
1	1,2,3-Trichlorobenzene	4.431	4.334	4.311	4.328	4.317	4.291	4.326	4.291	4.267	4.288
2	1,2,4-Trichlorobenzene	4.500	4.330	4.292	4.273	4.303	4.258	4.226	4.265	4.227	4.201
3	1-Bromo-2,3-dichlorobenzene	4.560	4.474	4.465	4.473	4.583	4.588	4.580	4.469	4.448	4.498
4	1-Bromo-2,6-dichlorobenzene	4.481	4.468	4.467	4.470	4.584	4.605	4.596	4.471	4.469	4.459
5	*m*-Dichlorobenzene	3.679	3.853	3.873	3.903	3.785	3.794	3.803	3.729	3.733	3.764
6	*p*-Dichlorobenzene	3.850	3.850	3.850	3.826	3.781	3.774	3.772	3.721	3.712	3.710
7	*o*-Dichlorobenzene	3.790	3.856	3.864	3.804	3.795	3.795	3.728	3.747	3.744	3.687
8	Chlorobenzene	3.195	3.229	3.237	3.253	3.259	3.264	3.276	3.169	3.167	3.143
9	Phenol	2.769	2.478	2.407	2.490	2.795	2.800	2.846	2.649	2.635	2.718
10	2-Chlorophemol	3.011	3.197	3.212	3.232	3.326	3.350	3.341	3.227	3.237	3.259
11	4-Bromophenol	3.664	3.544	3.536	3.532	3.580	3.570	3.517	3.382	3.314	3.309
12	4-Chlorophenol	3.421	3.204	3.187	3.145	3.315	3.306	3.261	3.201	3.191	3.152
13	4-Fluorophenol	2.693	2.759	2.767	2.773	2.820	2.844	2.849	2.629	2.621	2.576
14	2-Methoxyphenol	2.654	2.621	2.586	2.588	2.508	2.372	2.385	2.471	2.300	2.319
15	2-Methylphenol	2.837	2.941	2.950	2.963	3.019	3.032	3.041	3.126	3.141	3.158
16	4-Methoxyphenol	2.624	2.658	2.689	2.687	2.764	2.919	2.916	2.799	2.994	2.999
17	4-Methylphenol	3.057	2.949	2.940	2.910	3.010	3.007	2.964	3.100	3.102	3.052
18	4- *tert*-Butylphenol	4.033	3.925	3.916	3.922	3.886	3.870	3.865	4.223	4.244	4.254
19	2,6-Dimethylphenol	3.324	3.333	3.334	3.341	3.339	3.340	3.395	3.569	3.579	3.602
20	1-Naphthalenol	3.807	3.903	3.911	3.909	3.684	3.675	3.709	3.948	3.956	3.947
21	2-Naphthalenol	3.886	3.914	3.916	3.913	3.667	3.653	3.648	3.904	3.905	3.906
22	2,4-Dichlorophenol	3.873	3.771	3.753	3.707	3.725	3.707	3.646	3.745	3.738	3.691
23	2-Bromo-4-methylphenol	3.717	3.820	3.829	3.817	3.791	3.806	3.752	3.827	3.847	3.822
24	Resorcinol	2.066	2.490	2.612	2.540	2.643	2.747	2.729	2.690	2.760	2.757
25	Diphenylol propane	4.201	4.028	3.767	3.682	4.196	4.191	4.128	4.123	4.060	3.993
26	Diphenylol ethane	3.914	4.167	4.284	4.300	3.952	3.971	4.015	3.881	3.864	3.914
27	2,4-Dichloroaniline	3.732	3.360	3.134	3.147	3.502	3.479	3.486	3.641	3.637	3.588
28	4-Chloro-benzoic acid	3.417	3.246	3.167	3.201	3.442	3.443	3.435	3.394	3.393	3.380
29	4-Bromo-benzoic acid	3.625	3.680	3.697	3.681	3.705	3.713	3.749	3.575	3.564	3.596
30	Salicylic acid	2.840	2.749	2.714	2.728	2.896	2.903	2.889	2.934	2.942	2.940
31	5-Chloro-salicylic acid	3.011	3.224	3.321	3.326	3.365	3.416	3.411	3.367	3.384	3.365
32	4-Hydroxybenzaldehyde	3.080	3.212	3.219	3.183	2.871	2.845	2.811	2.874	2.854	2.842
33	Nitrobenzene	3.286	3.413	3.433	3.439	3.262	3.260	3.274	3.215	3.212	3.183
34	2-NitroToluene	3.530	3.759	3.782	3.731	3.544	3.545	3.560	3.588	3.590	3.608
35	4-NitroToluene	3.624	3.759	3.773	3.792	3.533	3.527	3.560	3.551	3.549	3.568
36	2-Nitrophenol	3.502	3.355	3.347	3.342	3.339	3.328	3.354	3.413	3.409	3.428
37	3-Nitrophenol	3.510	3.371	3.364	3.347	3.274	3.260	3.232	3.307	3.297	3.275
38	4-Nitrophenol	3.657	3.382	3.369	3.386	3.236	3.212	3.212	3.241	3.219	3.208
39	1-Chloro-4-nitrobenzene	3.934	3.971	3.973	3.962	3.770	3.762	3.775	3.709	3.703	3.722
40	1-Bromomethyl-4-nitrobenzene	4.383	4.348	4.345	4.321	4.576	4.632	4.645	4.621	4.670	4.666
41	1-Chloromethyl-4-nitrobenzene	4.321	4.182	4.173	4.161	4.165	4.157	4.161	4.213	4.206	4.234
42	4-Chloro-2-nitrophenol	3.882	3.862	3.860	3.831	3.850	3.847	3.799	3.908	3.909	3.875
43	2-Nitroresorcinal	3.492	3.568	3.573	3.567	3.400	3.378	3.325	3.577	3.582	3.556
44	2-Chloro-5-nitroaniline	3.466	3.871	4.172	4.134	3.894	3.919	3.889	3.805	3.815	3.812
45	4-Nitro-naphthalen-1-ylamine	4.236	4.185	4.162	4.059	4.140	4.116	4.134	4.058	4.039	4.008
46	*o*-Dinitrobenzene	4.050	4.066	4.067	4.037	4.160	4.181	4.188	4.241	4.279	4.303
47	*m*-Dinitrobenzene	4.015	4.066	4.071	4.080	3.979	3.974	3.969	3.948	3.940	3.934
48	2,4-DinitroToluene	4.061	4.267	4.288	4.278	4.192	4.211	4.198	4.163	4.176	4.163
49	2,4-Dinitrophenol	4.306	3.941	3.903	3.943	4.060	3.994	4.023	4.097	4.051	4.055
50	2,4-Dinitrobromobenzene	4.461	4.451	4.449	4.371	4.700	4.750	4.754	4.552	4.580	4.582
51	2,4-Dinitrochlorobenzene	4.342	4.418	4.423	4.420	4.516	4.553	4.573	4.450	4.470	4.519

[a]*Obs* observed [27], *Cal* calculated, *Pred* predicted
[b]From Eq. 9; [c] From Eq. 12; [d] From Eq. 15

The calculations of η, η_F, η_B and contributions of different vertices to η_F were performed, using the distance matrix and VEM vertex counts as inputs by the GW-BASIC programs *KRETA1* and *KRETA2* developed by one of the authors [34]. We have also modeled the toxicity data using other selected topological variables and compared the ETA models with non-ETA ones. The values for the non-ETA topological descriptors for the compounds were generated by the QSAR + and Descriptor + modules of the Cerius 2 Version 4.8 software [32]. The various topological indices calculated are Wiener *W*, Balaban *J*, flexibility index (ϕ), Hosoya *Z*, Zagreb, connectivity indices ($^0\chi, ^1\chi, ^2\chi, ^3\chi_p, ^3\chi_c, ^3\chi_{CH}, ^0\chi^v, ^1\chi^v, ^2\chi^v, ^3\chi_p^v, ^3\chi_c^v, ^3\chi_{CH}^v$), kappa shape indices ($^1\kappa, ^2\kappa, ^3\kappa, ^1\kappa_\alpha, ^2\kappa_\alpha, ^3\kappa_\alpha$) and E-state parameters (S_sCH3, S_ssCH2, S_aaCH, S_dssC, S_aasC, S_aaaC, S_sNH2, S_ddsN, S_sOH, S_dO, S_ssO, S_sCl, S_sBr). Alongwith the topological

Table 2 Definitions of important ETA parameters used in exploring QSAR of toxicity of benzene derivatives to *R. japonica*

Parameter	Definition
$\sum\alpha$	Sum of α values of all non-hydrogen vertices of a molecule
$[\sum\alpha]_P$	Sum of α values of all non-hydrogen vertices each of which is joined to only one other vertex of the molecule
N_V	Vertex count (excluding hydrogen)
$[\eta'_F]_{OCH_3}$	Functionality for the methoxy group
$[\eta_F]_{Cl}$	Functionality for the chloro group
$[\eta_F]_{OH}$	Functionality for the hydroxyl group
$[\eta_F]_{COOH}$	Functionality for the carboxyl group
$[\eta_F]_{NH_2}$	Functionality for the amino group

descriptors, a few physicochemical descriptors like *AlogP*98, *MolRef*,*H_bond_donor* and *H_bond_acceptor* were also considered among non-ETA descriptors.

The statistical quality of the equations [35] was judged by the parameters like explained variance (R_a^2, i.e., adjusted R^2), correlation coefficient (r or R), standard error of estimate (s) and variance ratio (F) at specified degrees of freedom (df). All the accepted equations have regression constants and F ratios significant at the 95 and 99% levels respectively, if not stated otherwise. A compound was considered as an outlier if the residual is more than twice the standard error of estimate for a particular equation. All the developed models were cross-validated using "leave-one-out" (LOO) technique. PRESS (LOO) statistics [36, 37] were calculated using the programs *KRPRES1* and *KRPRES2* [34], and LOO *cross-validation* R^2 (Q^2), *predicted residual sum of squares* (*PRESS*) were reported. Some seleceted equations were also subjected to leave-20%-out (L-20%-O) crossvalidation.

Results and discussion

Results from GFA-MLR

Results with ETA indices

Three best equations using ETA descriptors selected from the population of generated equations based on the values of R^2 as well as R_a^2 and Q^2 are given below:

$$\begin{aligned}\log(1/LC_{50}) &= -1.928 + 1.871(\pm0.444)\sum\alpha \\ &- 0.137(\pm0.040)\left[\sum\alpha\right]^2 - 2.055(\pm0.982)\left[\eta'_F\right]_{OCH3} \\ &+ 0.621(\pm0.494)\left[\eta'_F\right]_{Cl} - 1.601(\pm0.694)\left[\eta'_F\right]_{OH} \\ &- 1.346(\pm0.393)\left[\eta'_F\right]_{COOH} - 1.860(\pm1.482)\left[\eta'_F\right]_{NH_2} \\ &n = 51,\ Q^2 = 0.847,\ R_a^2 = 0.901,\ R^2 = 0.915, \\ &R = 0.956, s = 0.183,\ F = 65.841(df\,7, 43), \\ &\text{PRESS} = 2.563, \end{aligned} \quad (9)$$

$$\begin{aligned}\log(1/LC_{50}) &= -1.954 + 2.019(\pm0.464)\sum\alpha \\ &- 0.141(\pm0.040)\left[\sum\alpha\right]^2 - 0.052(\pm0.054)N_V \\ &- 1.374(\pm0.403)\left[\eta'_F\right]_{COOH} - 1.750(\pm0.690)\left[\eta'_F\right]_{OH} \\ &- 2.119(\pm1.006)\left[\eta'_F\right]_{OCH_3} \\ &- 1.951(\pm1.521)\left[\eta'_F\right]_{NH_2} \\ &n = 51,\ Q^2 = 0.848,\ R_a^2 = 0.895,\ R^2 = 0.910, \\ &R = 0.954, s = 0.188,\ F = 61.894(df\,7, 43), \\ &\text{PRESS} = 2.544, \end{aligned} \quad (10)$$

$$\begin{aligned}\log(1/LC_{50}) &= -1.437 + 1.684(\pm0.451)\sum\alpha \\ &- 0.123(\pm0.039)\left[\sum\alpha\right]^2 + 0.111(\pm0.109)\left[\sum\alpha\right]_P \\ &- 1.369(\pm0.340)\left[\eta'_F\right]_{COOH} - 1.767(\pm0.578)\left[\eta'_F\right]_{OH} \\ &- 2.102(\pm0.847)\left[\eta'_F\right]_{OCH_3} - 1.999(\pm1.276)\left[\eta'_F\right]_{NH_2} \\ &n = 51,\ Q^2 = 0.846,\ R_a^2 = 0.893,\ R^2 = 0.908, \\ &R = 0.953, s = 0.189,\ F = 60.7(df\,7, 43), \\ &\text{PRESS} = 2.579, \end{aligned} \quad (11)$$

The 95% cofidence intervals are shown in parentheses. Among the above three equations, Eq. 9 was chosen as the best based on statistical significance of the regression coefficients and intercorrelation of predictor variables.

The intercorrelation matrix among the predictor variables (Eq. 9) is given in the Table 3. The predicted variances of Eqs. 9, 10 and 11 ranges from 84.6 to 84.8% while the explained variance ranges from 89.3 to 90.1%. Equations 9, 10 and 11 show parabolic relations of the toxicity with $\sum\alpha$, which indicates that the toxicity increases with increase of size up to a certain level, after which it decreases. The positive coefficient of $[\eta'_F]_{Cl}$ indicates that the toxicity increases with the increase of the functionality contribution of chloro substituents. Again, the negative coefficients of $[\eta'_F]_{COOH}$, $[\eta'_F]_{OH}$, $[\eta'_F]_{OCH_3}$and $[\eta'_F]_{NH_2}$ indicate negative contributions of functionalities like carboxy, hydroxy, methoxy and amino groups.

Results with non-ETA indices

While working with non-ETA descriptors, three best equations were selected from the population of generated models on the basis of the values of R^2 as well as R_a^2 and Q^2.

$$\log(1/LC_{50}) = 0.897 + 0.959(\pm 0.202)^{1}\chi^{V} - 0.004(\pm 0.002)\text{Wiener} - 0.672(\pm 0.189)\text{S\_ddsN} - 0.012(\pm 0.012)\text{S\_sOH} - 0.110(\pm 0.060)\text{S\_ssO} + 0.024(\pm 0.014)\text{S\_sCl},$$

$$n = 51,\ Q^2 = 0.863,\ R_a^2 = 0.884,\ R^2 = 0.898,$$

$$R = 0.948, s = 0.197,\ F = 64.661(df\,6, 44),$$

$$\text{PRESS} = 2.301, \qquad (12)$$

$$\log(1/LC_{50}) = 0.501 + 1.029(\pm 0.189)^{1}\chi^{V} - 0.005(\pm 0.002)\text{Wiener} + 0.034(\pm 0.032)\text{S\_aaCH} - 0.933(\pm 0.212)\text{S\_ddsN} - 0.102(\pm 0.060)\text{S\_ssO} + 0.031(\pm 0.016)\text{S\_sCl},$$

$$n = 51,\ Q^2 = 0.868,\ R_a^2 = 0.884,\ R^2 = 0.898,$$

$$R = 0.947, s = 0.198,\ F = 64.3401(df\,6, 44),$$

$$\text{PRESS} = 2.209, \qquad (13)$$

$$\log(1/LC_{50}) = 0.937 - 0.289(\pm 0.343)^{2}\kappa_{\alpha} + 1.139(\pm 0.230)^{1}\chi^{V} - 0.003(\pm 0.002)\text{Wiener} - 0.828(\pm 0.169)\text{S\_ddsN} - 0.089(\pm 0.066)\text{S\_ssO} + 0.033(\pm 0.016)\text{S\_sCl}$$

$$n = 51, Q^2 = 0.863, R_a^2 = 0.880, R^2 = 0.894,$$

$$R = 0.946, s = 0.201, F = 61.896(df\,6, 44),$$

$$\text{PRESS} = 2.303. \qquad (14)$$

From these equations, Eq. 12 was chosen as the best based on statistical significance of the regression coefficients and intercorrelation among predictor variables.

The intercorrelation (r) matrix among the predictor variables (Eq. 12) is given in Table 3. The predicted variance of Eqs. 12, 13 and 14 ranges from 86.3 to 86.8% and the explained variance ranges from 88.0 to 88.4%. Equations 12, 13 and 14 show the importance of connectivity, kappa shape, E-state and Wiener indices for the toxicity. When compared to Eqs. 9, 10 and 11, Eqs. 12, 13 and 14 show lower explained variance values and higher predicted variance values than those of the former ones.

Results with ETA and non-ETA indices

While working with both ETA and non-ETA descriptors, three best equations were selected from the population of models generated with GFA.

Table 3 Intercorrelation (r) matrix

	$[\eta'_F]_{CH_3}$	$[\eta'_F]_{Cl}$	$\sum\alpha$	$[\sum\alpha]^2$	$[\eta'_F]_{OH}$	$[\eta'_F]_{COOH}$	$[\eta'_F]_{NH_2}$	N_V	$^1\chi^V$	W	S_ddsN	S_sOH	S_ssO	S_sCl	S_aaCH	$^2\kappa_\alpha$	S_sBr	$[\sum\alpha]_P/\sum\alpha$
$[\eta'_F]_{CH_3}$	1.000	−0.130	−0.161	−0.149	0.171	−0.059	−0.050	−0.104	−0.093	−0.073	0.127	0.115	0.999	−0.130	−0.015	0.019	−0.081	−0.152
$[\eta'_F]_{Cl}$		1.000	0.012	−0.023	−0.373	−0.025	−0.057	−0.305	0.078	−0.310	0.239	−0.355	−0.130	0.970	−0.180	−0.011	0.064	0.573
$\sum\alpha$			1.000	0.990	−0.322	0.031	0.107	0.873	0.959	0.863	−0.249	0.160	−0.161	0.020	0.311	0.914	0.291	0.191
$[\sum\alpha]^2$				1.000	−0.267	0.007	0.094	0.873	0.951	0.882	−0.200	0.203	−0.149	−0.017	0.382	0.908	0.253	0.111
$[\eta'_F]_{OH}$					1.000	−0.015	−0.251	−0.171	−0.334	−0.148	0.261	0.669	0.170	−0.390	0.071	−0.315	−0.192	−0.412
$[\eta'_F]_{COOH}$						1.000	−0.073	0.025	−0.016	−0.014	0.183	0.343	−0.059	−0.013	−0.098	0.094	0.082	0.074
$[\eta'_F]_{NH_2}$							1.000	0.138	0.111	0.118	−0.003	−0.211	−0.050	−0.058	0.010	−0.029	−0.099	0.018
N_V								1.000	0.752	0.985	−0.509	0.285	−0.104	−0.290	0.238	0.866	−0.050	−0.101
$^1\chi^V$									1.000	0.751	−0.049	0.123	−0.090	0.089	0.398	0.853	0.346	0.217
Wiener										1.000	−0.468	0.293	−0.071	−0.299	0.285	0.875	−0.061	−0.151
S_ddsN											1.000	0.208	0.127	0.235	0.460	−0.357	0.052	−0.131
S_sOH												1.000	0.115	−0.359	0.234	0.210	−0.181	−0.343
S_ssO													1.000	−0.130	−0.017	0.021	−0.081	−0.152
S_sCl														1.000	−0.191	0.011	0.021	0.568
S_aaCH															1.000	0.131	−0.123	−0.619
$^2\kappa_\alpha$																1.000	0.193	0.184
S_sBr																	1.000	0.554
$[\sum\alpha]_P/\sum\alpha$																		1.000

$$\log(1/LC_{50}) = -0.222 + 1.513(\pm 0.226)^{1}\chi^{V} - 0.975(\pm 0.173)\text{S\_ddsN} - 0.126(\pm 0.060)\text{S\_ssO} - 0.009(\pm 0.002)\text{Wiener} - 0.139(\pm 0.070)\text{S\_sBr},$$
$$n = 51,\ Q^2 = 0.859,\ R_a^2 = 0.881,\ R^2 = 0.893,\ R = 0.945, s = 0.200,\ F = 74.839(df\,5, 45),\ \text{PRESS} = 2.375, \quad (15)$$

$$\log(1/LC_{50}) = 0.684 + 0.027(\pm 0.016)\text{S\_sCl} - 0.785(\pm 0.165)\text{S\_ddsN} + 1.034(\pm 0.197)^{1}\chi^{V} - 0.112(\pm 0.062)\text{S\_ssO} - 0.005(\pm 0.002)\text{Wiener},$$
$$n = 51,\ Q^2 = 0.858,\ R_a^2 = 0.875,\ R^2 = 0.887,\ R = 0.942, s = 0.205,\ F = 70.803(df\,5, 45),\ \text{PRESS} = 2.377, \quad (16)$$

$$\log(1/LC_{50}) = 0.685 - 0.785(\pm 0.165)\text{S\_ddsN} + 1.033(\pm 0.197)^{1}\chi^{V} - 0.005(\pm 0.002)\text{Wiener} + 0.027(\pm 0.016)\text{S\_sCl} - 1.956(\pm 1.104)[\eta'_F]_{OCH_3},$$
$$n = 51,\ Q^2 = 0.855,\ R_a^2 = 0.874,\ R^2 = 0.886,\ R = 0.941, s = 0.206,\ F = 70.223(df\,5, 45),\ \text{PRESS} = 2.432. \quad (17)$$

From these equations, Eq. 15 was chosen as the best one based on statistics of the regression coefficients and intercorrelation of the predictor variables. The intercorrelation matrix (Eq. 15) among the predictor variables is given in Table 3. When Eq. 15 is compared to Eq. 9, it is observed that there is marginal improvement in cross-validation statistics on using non-ETA parameters with ETA ones. This shows that the ETA parameters are sufficiently rich in chemical information to encode the structural features that contribute significantly to the acute toxicity of the benzene derivatives to tadpoles (*R. japonica*).

Results from FA-MLR

Results with ETA indices

Table 4 shows the results of factor analysis of the data matrix composed of ETA descriptors. It is observed that 11 factors could explain 97.7% of the variance of the data matrix. Based on the results of the factor analysis, the following relationship was derived:

$$\log(1/LC_{50}) = -1.477 + 1.648(\pm 0.556)\sum\alpha - 0.118(\pm 0.050)\left[\sum\alpha\right]^2 + 0.752(\pm 0.746)\left[\sum\alpha\right]_P/\sum\alpha - 1.557(\pm 0.715)[\eta'_F]_{OH} - 1.324(\pm 0.431)[\eta'_F]_{COOH} - 2.065(\pm 1.080)[\eta'_F]_{OCH_3}$$
$$n = 51,\ Q^2 = 0.852,\ R_a^2 = 0.879,\ R^2 = 0.894,\ R = 0.945, s = 0.201,\ F = 61.760(df\,6, 44),\ \text{PRESS} = 2.491. \quad (18)$$

Table 4 Factor loadings of the variables (ETA parameters) after VARIMAX rotation (ETA matrix)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11	Communalities
BA	0.554	0.152	0.472	−0.186	0.435	0.052	−0.280	−0.215	0.069	0.073	−0.041	0.916
η	0.166	−0.204	0.040	−0.095	0.055	0.036	−0.067	−0.059	−0.002	0.950	−0.024	0.996
η_R	0.872	0.309	−0.014	0.133	−0.025	0.091	0.037	−0.016	−0.017	−0.232	0.043	0.940
N_V	0.916	0.353	−0.030	0.125	−0.005	0.067	0.046	−0.006	−0.073	0.041	0.047	0.996
η_F	0.422	0.868	−0.017	0.094	−0.092	−0.024	0.112	−0.005	−0.110	−0.081	0.052	0.984
$[\eta_F]_{Cl}$	−0.154	−0.218	0.537	−0.644	0.217	−0.180	−0.138	−0.001	−0.210	0.185	−0.039	0.952
$[\eta_F]_{Br}$	0.063	−0.098	0.350	−0.030	−0.023	−0.080	−0.071	−0.010	0.906	−0.004	−0.045	0.973
$[\eta_F]_{OH}$	−0.146	−0.140	−0.302	0.060	−0.899	0.067	−0.017	0.045	0.032	−0.052	−0.150	0.977
$[\eta_F]_{NO_2}$	0.178	0.874	0.172	0.250	0.206	0.092	−0.107	−0.073	0.003	−0.098	−0.034	0.965
$[\eta_F]_{COOH}$	0.037	−0.119	0.064	−0.061	−0.006	−0.040	0.978	−0.032	−0.056	−0.061	−0.039	0.990
$[\eta_F]_{NH_2}$	0.089	0.011	0.016	0.064	0.107	−0.017	−0.039	−0.022	−0.036	−0.021	0.984	0.995
$[\eta_F]_{CH_3}$	0.051	−0.429	0.052	0.803	−0.019	0.088	−0.181	0.233	−0.149	−0.068	0.099	0.966
$[\eta_F]_{OCH_3}$	−0.093	−0.075	−0.124	0.123	−0.048	−0.019	−0.029	0.969	−0.007	−0.053	−0.027	0.992
$\sum\beta'_s$	0.198	0.875	0.157	−0.177	−0.066	−0.061	0.037	0.177	−0.198	−0.107	0.035	0.954
$\sum\beta'_{ns}$	−0.133	0.825	−0.253	−0.234	0.145	−0.220	−0.184	−0.191	0.138	−0.030	−0.024	0.978
$\sum\beta'$	−0.046	0.912	−0.153	−0.237	0.095	−0.192	−0.134	−0.097	0.050	−0.055	−0.008	0.992
$\Sigma\,\alpha$	0.958	0.040	0.159	−0.086	0.130	−0.009	−0.003	−0.028	0.109	0.113	0.016	0.994
$\Sigma\,\alpha]_P$	0.283	−0.053	0.879	−0.103	0.205	0.016	0.021	−0.021	0.260	0.050	−0.017	0.979
$[\Sigma\,\alpha]_Y$	0.810	0.405	0.273	0.092	−0.117	−0.175	0.092	−0.073	−0.047	−0.013	0.161	0.989
$[\Sigma\,\alpha]_X$	0.082	−0.252	0.129	0.110	−0.046	0.930	−0.051	−0.020	−0.073	0.036	−0.023	0.976
$[\Sigma\,\alpha]_P/\Sigma\,\alpha$	−0.010	−0.114	0.945	−0.075	0.170	−0.002	0.015	−0.024	0.195	0.042	−0.006	0.982
$[\Sigma\,\alpha]_Y/\Sigma\,\alpha$	0.416	0.561	0.347	0.227	−0.316	−0.249	0.156	−0.101	−0.197	−0.160	0.226	0.973
$[\Sigma\,\alpha]^2$	0.969	0.000	0.073	−0.076	0.099	−0.029	−0.022	−0.027	0.107	0.140	0.006	0.992
η^{local}	0.906	−0.318	−0.014	0.094	0.056	0.163	−0.007	−0.008	−0.050	0.125	−0.011	0.979
η'_B	0.087	0.300	0.820	0.221	−0.077	0.310	0.145	−0.158	−0.038	−0.097	0.096	0.986
Variance (%)	0.235	0.204	0.139	0.061	0.055	0.050	0.049	0.047	0.046	0.045	0.044	0.977

The 95% confidence intervals are shown in parentheses.The positive coefficient of $[\Sigma\ \alpha]_P/\Sigma\ \alpha$ in Eq. 18 indicates that the toxicity increases on increase of branching. Again, the parabolic relation of the toxicity with $\Sigma\ \alpha$ indicates that the toxcity increases with increase of size up to a certain level after which it decreases. The negative coefficients of $[\eta_F]_{COOH}$, $[\eta_F]_{OH}$ and $[\eta_F]_{OCH_3}$ indicate the reduction of the toxicity in presence of carboxy, hydroxy and methoxy groups.

Results with non-ETA indices

Based on the results (table not shown) of the factor analysis of the matrix consisting of non-ETA descriptors, the following equation was obtained:

$$\begin{aligned}\log(1/LC_{50}) &= 1.488 + 0.439(\pm 0.083)^1\chi \\ &+ 0.325(\pm 0.266)\text{S\_dssC} - 0.022(\pm 0.014)\text{S\_sOH} \\ &- 0.131(\pm 0.077)\text{S\_ssO} + 0.057(\pm 0.017)\text{S\_sCl} \\ &+ 0.165(\pm 0.066)\text{S\_sBr}, n = 51, \\ &Q^2 = 0.782,\ R_a^2 = 0.812,\ R^2 = 0.835,\ R = 0.914, \\ &s = 0.251,\ F = 37.048(df\,6, 44),\ \text{PRESS} = 3.655.\end{aligned} \tag{19}$$

The explained variance (0.812) and predicted variance (0.782) values of Eq. 19 are lower than those of Eq. 18 derived from ETA descriptors. Importance of E-state terms (S_dssC, S_sOH, S_ssO, S_sCl and S_sBr) and molecular connectivity parameter for the toxicity is relevant from Eq. 19.

Results with ETA and non-ETA indices

When the data matrix composed of both ETA and non-ETA parameters was considered, the following equation (Eq. 20) was obtained as best model after factor analysis:

$$\begin{aligned}\log(1/LC_{50}) &= -1.530 + 1.627(\pm 0.524)\sum \alpha \\ &- 0.108(\pm 0.048)\left[\sum \alpha\right]^2 - 0.029(\pm 0.012)\text{S\_sOH} \\ &- 0.096(\pm 0.070)\text{S\_ssO} + 0.019(\pm 0.014)\text{S\_sCl}, \\ &n = 51, \quad Q^2 = 0.815,\ R_a^2 = 0.843,\ R^2 = 0.859, \\ &R = 0.927, s = 0.229,\ F = 54.773(df\,5, 45), \\ &\text{PRESS} = 3.112.\end{aligned} \tag{20}$$

Equation 20 is inferior to both Eqs. 18 and 19 with respect to equation statistics; however, the crossvalidation statistics of the former are better than those of Eq. 19.

Results from PCRA

Results with ETA indices

An attempt was also made to use factor scores, as the predictor variables to avoid loss of information on selection of relevant molecular descriptors from the set of descriptors and a significant increase in statistical quality was obtained.

$$\begin{aligned}\log(1/LC_{50}) &= 3.643 + 0.321(\pm 0.056)f_1 \\ &+ 0.088(\pm 0.056)f_2 + 0.273(\pm 0.056)f_3 \\ &- 0.108(\pm 0.056)f_4 + 0.252(\pm 0.056)f_5 \\ &- 0.162(\pm 0.056)f_7 - 0.125(\pm 0.056)f_8, \\ &n = 51,\ Q^2 = 864,\ R_a^2 = 0.885,\ R^2 = 0.901, \\ &R = 0.949, s = 0.196,\ F = 56.065(df\,7, 43), \\ &\text{AVRES} = 0.163,\ \text{PRESS} = 2.291, \\ &\text{SDEP} = 0.212,\ S_{\text{PRESS}} = 0.231, \\ &\Pr\ es_{\text{av}} = 0.137.\end{aligned} \tag{21}$$

Equation 21 could predict and explain 86.4 and 88.5% respectively, of the variance of the acute toxicity.

Results with non-ETA indices

When factor scores (derived from non-ETA matrix) were used as predictor variables, a tangible rise in statistical quality was obtained with respect to Eq. 19.

$$\begin{aligned}\log(1/LC_{50}) &= 3.643 + 0.297(\pm 0.053)f_1 \\ &+ 0.336(\pm 0.053)f_2 - 0.098(\pm 0.053)f_3 \\ &+ 0.053(\pm 0.053)f_4 - 0.210(\pm 0.053)f_5 \\ &+ 0.112(\pm 0.053)f_6 + 0.068(\pm 0.053)f_7 \\ &+ 0.095(\pm 0.053)f_9 + 0.061(\pm 0.053)f_{10} \\ &- 0.148(\pm 0.053)f_{11}, \\ &n = 51,\ Q^2 = 0.866,\ R_a^2 = 0.903,\ R^2 = 0.922, \\ &R = 0.960, s = 0.181,\ F = 47.352(df\,10, 40), \\ &\text{AVRES} = 0.152,\ \text{PRESS} = 2.243, \\ &\text{SDEP} = 0.210,\ S_{\text{PRESS}} = 0.237, \\ &\Pr\ es_{\text{av}} = 0.115.\end{aligned} \tag{22}$$

Equation 22 based on the factor scores of the data matrix of non-ETA variables is statistically comparable to Eq. 21 based on the factor scores of the data matrix of ETA variables. This shows that the matrix composed of ETA descriptors is as rich in chemical information as the matrix composed of different and diverse non-ETA descriptors.

Results with ETA and non-ETA indices

Using factor scores derived from the combined matrix of ETA and non-ETA descriptors as the predictor variables, the following equation was obtained:

Table 5 Results of leave-20%-out cross-validation applied on selected equations *Model equation*, $pC = \sum \beta_i x_i + \alpha$

Name of Statistical Methods	Equation number	Number of cycles	Average regression coefficients (SD)	Statistics Q^2 (Average Pres)
GFA-MLR	(9)	5[a]	$-2.005(0.266) + 1.901(0.127)\sum\alpha - 0.140(0.014)\left[\sum\alpha\right]^2$ $+0.628(0.122)[\eta'_F]_{Cl} - 1.578(0.147)[\eta'_F]_{OH} - 1.342(0.122)[\eta'_F]_{COOH}$ $-1.854(1.079)[\eta'_F]_{NH2} - 2.056(0.135)[\eta'_F]_{OCH3}$	0.848 (0.168)
	(12)	5[a]	$0.894(0.088) + 0.961(0.041)^1\chi^V - 0.004(0.000)$Wiener $-0.672(0.043)$S_ddsN $- 1.263(0.003)$S_sOH $-0.109(0.025)$S_ssO $+ 2.367(0.004)$S_sCl	0.860 (0.173)
	(15)	5[a]	$-0.231(0.138) + 1.517(0.050)^1\chi^V$ $-0.982(0.045)$S_ddsN $- 0.124(0.030)$S_ssO $-0.009(0.000)$Wiener $- 0.139(0.016)$S_sBr	0.852 (0.175)
FA-MLR	(18)	5[a]	$-1.539(0.274) + 1.671(0.134)$ $\sum\alpha - 0.120(0.014)[\sum\alpha]^2 + 0.761(0.163)[\sum\alpha]_p/\sum\alpha$ $-1.536(0.130)[\eta'_F]$OH $- 1.323(0.136)[\eta'_F]_{COOH} - 2.064(0.149)[\eta'_F]_{OCH3}$	0.860 (0.164)
	(19)	5[a]	$1.475(0.156) + 0.442(0.029)^1\chi + 0.322(0.071)$S_dssC $-0.022(0.005)$S_sOH $- 0.130(0.018)$S_ssO $+0.057(0.002)$S_sCl $+ 0.163(0.012)$S_sBr	0.781 (0.204)
	(20)	5[a]	$-1.551(0.327) + 1.637(0.133)\sum\alpha - 0.109(0.014)[\sum\alpha]^2$ $-0.096(0.007)$S_ssO $- 0.029(0.005)$S_sOH $+ 0.018(0.003)$S_sCl	0.814 (0.184)
PCRA	(21)	5[a]	$3.643(0.011) + 0.320(0.017)f1 + 0.089(0.012)f2$ $+0.272(0.019)f3 - 0.108(0.009)f4$ $+0.251(0.014)f5 - 0.162(0.013)f7 - 0.125(0.009)f8$	0.871 (0.159)
	(22)	5[a]	$3.646(0.020) + 0.300(0.016)f1 + 0.335(0.004)f2$ $-0.098(0.005)f3 + 0.072(0.052)f4$ $-0.210(0.023)f5 + 0.110(0.014)f6$ $+0.069(0.011)f7 + 0.092(0.016)f9 + 0.065(0.005)f10$	0.839 (0.164)
	(23)	5[a]	$3.643(0.018) + 0.340(0.021)f1 + 0.077(0.007)f2$ $+0.269(0.006)f3 + 0.154(0.009)f5$ $-0.202(0.016)f6 - 0.134(0.014)f8 - 0.148(0.015)f10$	0.857 (0.172)

Q^2 denotes cross-validated R^2

Average *Pres* means average of absolute values of *predicted residuals*

[a]Compounds were deleted in five cycles in the following manner: (1, 6, 11, 16, 21, 26,...,46, 51), (2, 7, 12, 17, 22,...,47),..., (5, 10, 15, 20, 25, 30,...,50)

$$\log(1/LC_{50}) = 3.643 + 0.338(\pm0.058)f_1 + 0.079(\pm0.058)f_2 + 0.272(\pm0.058)f_3 + 0.155(\pm0.058)f_5 - 0.201(\pm0.058)f_6 - 0.134(\pm0.058)f_8 - 0.148(\pm0.058)f_{10},$$

$$n = 51,\ Q^2 = 0.846,\ R_a^2 = 0.872,\ R^2 = 0.890,$$

$$R = 0.943, s = 0.207,\ F = 49.647(df\,7, 43),$$

$$\text{AVRES} = 0.176,\ \text{PRESS} = 2.583,$$

$$\text{SDEP} = 0.225,\ S_{\text{PRESS}} = 0.245,\ \text{Pr}\,es_{\text{av}} = 0.147. \quad (23)$$

Equation 23 shows inferior cross-validation statistics than Eq. 21 derived from factor scores of the ETA matrix.

Overview of the results

Based on cross-validation and equation statistics, Eqs. 9, 12 and 15 are the best ones obtained from ETA, non-ETA and combined matrices, respectively. The calculated and predicted acute toxicity values according to Eqs. 9, 12 and 15 are given in the Table 1. The inter-correlation (r) among the predictor variables of different equations is given in the Table 3. Selected equations were also subjected to leave-20%-out (L-20%-O) cross-validation and the results are shown in Table 5. For

Table 6 Prediction of toxicity of test set compounds in different cross-validation cycles (leave-20%-out) based on the descriptor set of Eq. 18

Cycle	Test set compounds	Training set compounds	Regression coefficients							r^2_{pred} [a]
			Intercept	$\sum \alpha$	$[\sum \alpha]^2$	$[\sum \alpha]_P / \sum \alpha$	$[\eta'_F]_{OH}$	$[\eta'_F]_{COOH}$	$[\eta'_F]_{OCH3}$	
1	1, 6, 11, 16, 21, 26, 31, 36, 41, 46, 51	Rest of the compounds ($n=40$)	−1.121	1.468	−0.100	0.870	−1.382	−1.122	−1.882	0.887
2	2, 7, 12, 17, 22, 27, 32, 37, 42, 47	Rest of the compounds ($n=41$)	−1.577	1.698	−0.122	0.482	−1.685	−1.263	−1.959	0.801
3	3, 8, 13, 18, 23, 28, 33, 38, 43, 48	Rest of the compounds ($n=41$)	−1.645	1.697	−0.122	0.827	−1.439	−1.461	−2.083	0.889
4	4, 9, 14, 19, 24, 29, 34, 39, 44, 49	Rest of the compounds ($n=41$)	−1.483	1.652	−0.119	0.871	−1.525	−1.347	−2.264	0.830
5	5, 10, 15, 20, 25, 30, 35, 30, 45, 50	Rest of the compounds ($n=41$)	−1.871	1.839	−0.139	0.754	−1.647	−1.422	−2.132	0.910

[a] $r^2_{\text{pred}} = \sqrt{1 - \frac{(Y_{\text{obs}} - Y_{\text{pred}})^2}{(Y_{\text{obs}} - \overline{Y})^2}}$, where Y_{obs} and Y_{pred} indicate observed and predicted toxicity values, respectively, and $\overline{Y}$ indicate mean toxicity value of the corresponding training set

each equation, cross-validation was run in five cycles, in each of which 20% of the compounds were deleted from the original data set (*vide* footnote of Table 5) and the rest of the compounds were taken as the training set. Based on the equation developed from the reduced dataset (training set), the toxicity values for the deleted compounds (test set) were predicted and this procedure was continued until each of the compounds of the original data set was deleted once in five cycles of cross-validation. This is elaborated more in detail in Table 6 taking the example of leave-20%-out cross-validation applied on Eq. 18. In each cycle, toxicity values of 20% of the compounds were predicted based on the equation derived from the remaining 80% of compounds and the predictive r^2 (r^2_{pred}) value was calculated and reported. In all cases, r^2_{pred} values are found to be larger than 0.8.

The derived relations (Eqs. 9, 12, 15) from GFA are of excellent statistical quality (predicted variance 0.847, 0.863 and 0.859 with R^2 0.915, 0.898, 0.893 from ETA, non-ETA and combined matrices, respectively), which are comparable to those (predicted variance and R^2 of the best equation being 0.785 and 0.914, respectively) of the previously reported equations obtained from stepwise multiple regression analysis applied on the same data set using physicochemical descriptors [27].

Conclusion

This study on the current dataset suggests that ETA parameters are sufficiently rich in chemical information to encode the structural features that contribute significantly to the acute toxicity of benzene derivatives to tadpoles (*R. japonica*) [27]. This indicates that ETA indices merit further assessment to explore their potential in QSAR/QSPR/QSTR modeling.

References

1. Kavlock RJ, Daston GP, Derosa C, Fenner-Crisp P, Gray LE, Kaattari S, Lucier G, Luster M, Mac MJ, Maczka C, Miller R, Moore J, Rolland R, Csott G, Sheehan DM, Sinks T, Tilson HA (1996) Environ Health Perspect 104:715–740
2. Sanderson H, Jonson DJ, Reitsma T, Brain RA, Wilson CJ, Solomon KR (2004) Regul Toxicol Pharmacol 39:158–183
3. McKinney JD, Richard A, Waller C, Newman MC, Gerberick F (2000) Toxicol Sci 56:8–17
4. Comber MH, Walker JD, Watts C, Hermens J (2003) Environ Toxicol Chem 22:1822–1828
5. Russom CL, Anderson EB, Greenwood BE, Pilli A (1991) Sci Total Environ 109–110:667–670
6. Ren S (2002) Environ Toxicol 17:119–127
7. Rose K, Hall LH (2003) SAR QSAR Environ Res 14:113–129
8. Mazzatorta P, Benfenati E, Neagu CD, Gini G (2003) J Chem Inf Comput Sci 43:513–518
9. Vighi M, Gramatica P, Consolaro F, Todeschini R (2001) Ecotoxicol Environ Saf 49:206–220
10. Bask SC, Grunwald GD, Gute BD, Balasubramanian K, Opitz D (2000) J Chem Inf Comput Sci 40:885–890
11. Devillers J (2001) SAR QSAR Environ Res 11:397–417
12. Roy K, Ghosh G (2003) Internet Electron J Mol Des 2:599–620; http://www.biochempress.com
13. Roy K, Ghosh G (2004) J Chem Inf Comput Sci 44:559–567
14. Roy K, Ghosh G (2004) QSAR Comb Sci 23:99–108
15. Roy K, Ghosh G (2004) QSAR Comb Sci 23:526–535
16. Roy K, Ghosh G (2005) Bioorg Med Chem 13:1185–1194
17. Pal DK, Sengupta C, De AU (1988) Indian J Chem 27B:734–739
18. Pal DK, Sengupta C, De AU (1989) Indian J Chem 28B:261–267
19. Pal DK, Sengupta M, Sengupta C, De AU (1990) Indian J Chem 29B:451–454
20. Pal DK, Purkayastha SK, Sengupta C, De AU (1992) Indian J Chem 31B:109–114
21. Roy K, Pal DK, De AU, Sengupta C (1999) Indian J Chem 38B:664–671
22. Roy K, Pal DK, De AU, Sengupta C (2001) Indian J Chem 40B:129–135
23. Roy K, Saha A (2003) J Mol Model 9:259–270

24. Roy K, Saha A (2003) Internet Electron J Mol Des 2:288–305; http://www.biochempress.com
25. Roy K, Saha A (2003) Internet Electron J Mol Des 2:475–491 http://www.biochempress.com
26. Roy K, Chakroborty S, Ghosh CC, Saha A (2004) J Indian Chem Soc 81:115–125
27. Huang, H, Wang X, Ou W, Zhao J, Shao Y, Wang L (2003) Chemosphere 53:963–970
28. Rogers D, Hopfinger AJ (1994) J Chem Inf Comput Sci 34:854–866
29. Fan Y, Shi LM, Kohn KW, Pommier Y, Weinstein JN (2001) J Med Chem 44:3254–3263
30. Lewi PJ (1980) Multivariate data analysis in structure-activity relationships. In: Ariens EJ (ed) Drug design, vol 10. Academic Press, NY, pp 307–342
31. Franke R, Gruska A (1995) Principal component and factor analysis. In: van de Waterbeemd H (ed) Chemometric methods in molecular design, vol 2. VCH, Weinheim, pp 113–163
32. Cerius 2 Version 4.8 is a product of Accelrys Inc., San Diego, CA
33. SPSS is statistical software of SPSS Inc., IL
34. The GW-BASIC programs *RRR98, KRETA1, KRETA2, KRPRES1 and KRPRES2* were developed by Kunal Roy and standardized using known data sets
35. Snedecor GW, Cochran WG (1967) Statistical methods. Oxford and IBH Publishing Co Pvt Ltd, New Delhi, pp 381–418
36. Wold S, Eriksson L (1995) Statistical validation of QSAR results. In: van de Waterbeemd H (ed) Chemometric methods in molecular design. VCH, Weinheim, pp 312–317
37. Debnath AK (2001) Quantitative structure-activity relationship (QSAR): A versatile tool in drug design. In: Ghose AK, Viswanadhan VN (eds) Combinatorial library design and evaluation. Marcel Dekker, NY, pp 73–129