

## QSTR with Extended Topochemical Atom Indices. 2. Fish Toxicity of Substituted Benzenes

Kunal Roy\* and Gopinath Ghosh

Drug Theoretics and Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry,  
Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

Received September 15, 2003

Considering the importance of quantitative structure-toxicity relationship (QSTR) studies in the field of aquatic toxicology from the viewpoint of ecological safety assessment, fish toxicity of various benzene derivatives has been modeled by the multiple regression technique using recently introduced extended topochemical atom (ETA) indices. The toxicity data have also been modeled using other selected topological descriptors and physicochemical variables, and the best ETA model has been compared to the non-ETA ones. Principal component factor analysis was used as the data preprocessing step to reduce the dimensionality of the data matrix and identify the important variables that are devoid of collinearities. All-possible-subsets regression was also applied on the parameters to cross-check the variable selection for the best model. Multiple linear regression analyses show that the best non-ETA model involves  $^1\chi$ , ALogP98, and LUMO (energy) as predictor variables and the quality of the relation is as follows:  $n = 92$ ,  $Q^2 = 0.718$ ,  $R_a^2 = 0.730$ ,  $R^2 = 0.738$ ,  $R = 0.859$ ,  $F = 82.8$  ( $df\ 3, 88$ ),  $s = 0.340$ . On the other hand, the best ETA model has the following quality:  $n = 92$ ,  $Q^2 = 0.865$ ,  $R_a^2 = 0.876$ ,  $R^2 = 0.885$ ,  $R = 0.941$ ,  $F = 92.6$  ( $df\ 7, 84$ ),  $s = 0.230$ . The ETA relations showed positive contributions of molecular bulk (size), chloro and hydroxy substitutions in the benzene ring, and the simultaneous presence of methyl and nitro substitutions to the toxicity. Further, the presence of fluoro and ether functionality, amino or nitro functionality in an otherwise unsubstituted ring, and nitro functionality that is ortho to a chloro substituent decreases toxicity. An attempt to use non-ETA descriptors along with ETA ones did not improve the quality in comparison to the best ETA model. Interestingly, the ETA model developed presently for the fish toxicity is better than the previously reported models on the same data set. Thus, it appears that ETA descriptors have significant potential in QSAR/QSPR/QSTR studies, which warrants extensive evaluation.

### INTRODUCTION

Assessment of the adverse effects of pollutants on an exposed ecosystem is a field of contemporary interest from the viewpoint of environmental safety. Research in environmental science should provide the scientific basis for risk assessment procedures and precautionary measures and finally provide a decision support for policy and management.<sup>1</sup> Faced with the need to predict physical and chemical properties, environmental fate, ecological effects, and health effects of organic chemicals in the absence of experimental data, several government organizations have been applying Structure–Activity Relationships (SARs) and Quantitative Structure–Activity Relationships (QSARs) to develop those predictions.<sup>2</sup> In the field of aquatic toxicology, quantitative structure–activity relationships (QSARs) have developed as scientifically credible tools for predicting the toxicity of chemicals when little or no empirical data are available.<sup>3</sup> A fundamental understanding of toxicological principles has been considered an important component to the acceptance and application of QSAR approaches as biologically relevant in ecological risk assessment.<sup>3</sup> As a consequence, there has been an evaluation of QSAR development and application from that of a chemical-class perspective to one that is more

consistent with assumptions regarding modes of toxic action.<sup>3</sup> Application of QSAR helps to reduce expense and the time consumption to predict the toxicity levels (i.e., data) of chemical agents. Without a biological experiment, one can also predict the toxicity value of a chemical agent with the help of a QSAR study. Toxicity to fish and other aquatic organisms has been used in a great deal of literature to indicate some of the hazardous effects on the ecosystem. To evaluate environmentally safe levels of dangerous chemicals, there is the need for a set of biological toxicity data on organisms representative of the ecosystems, which is often unavailable or inadequate.<sup>4</sup> Different topological descriptors and physicochemical variables are used to develop QSTR models for the correlation of toxicity of environmental pollutants for better ecotoxicological management. More than 2100 chemically defined organic chemicals are listed in the Research Institute of Fragrance Materials/Flavor and Extract Manufacturer's Association (RIFM/FEMA) database that are used as ingredients of fragrances for consumer products.<sup>5</sup> An approach was developed for prioritizing these fragrance materials for aquatic risk assessment by first estimating the predicted environmental concentration (PEC) of these fragrance materials in the aquatic environment based upon their physicochemical properties and annual value of use.<sup>5</sup> Subsequently, a toxicity level was predicted with a general quantitative structure–activity

\* Corresponding author phone: +91-33-2414 6676; e-mail: kunalroy\_in@yahoo.com.

relationship (QSAR) for aquatic toxicity, and a predicted no-effect concentration (PNEC) was calculated from this toxicity level by using an assessment factor (AF) that accounts for uncertainty in the toxicity QSAR prediction.<sup>5</sup> Katritzky et al. developed QSTR for the prediction of aqueous toxicities for *Poecilia reticulata* (guppy) using the CODESSA treatment.<sup>6</sup> Estrada et al. has applied the topological substructural molecular design (TOPS-MODE) approach to the study of toxicological properties.<sup>7</sup> After a comparative study of four QSAR models of aromatic compounds to aquatic organisms on the basis of octanol/water partition coefficient, linear solvation energy relationship (LSER), molecular connectivity index, and group contribution, Yu et al. showed that LSER was the best one.<sup>8</sup> Bioconcentration factors (BCFs) have also been used to develop QSAR models to predict the toxicity data of organic compounds for fish.<sup>9</sup> Kulkarni et al. have modeled and classified organic chemicals using physicochemical parameters to assist designing ecofriendly molecules.<sup>10</sup> Recently, we have modeled the toxicity of substituted phenols against *Tetrahymena pyriformis* to explore the suitability of newly developed extended topochemical atom (ETA) indices<sup>11</sup> in modeling studies. We established a good correlation between the toxicity of substituted phenols and ETA parameters.<sup>11</sup> In our present work, fish toxicity data of 92 diverse aromatic compounds against *Poecilia reticulata*<sup>12</sup> have been modeled with ETA parameters using the multiple regression technique, and the best relation obtained has been compared to that with some selected topological and physicochemical descriptors and also with models reported previously.<sup>12,13</sup>

## MATERIALS AND METHODS

In late 1980s, TAU descriptors were reported<sup>14,15</sup> and claimed to have diagnostic power to unveil specific contributions of functionality, branching, shape, and size factors to biological activity or physicochemical parameters. Later, a number of papers were published in support of the claim.<sup>16–23</sup> To accomplish further refinement over the TAU formalism in the valence electron mobile (VEM) environment, some of the basic parameters introduced in the TAU scheme were redefined in the ETA scheme.<sup>11</sup>

The core count of a non-hydrogen vertex  $[\alpha]$  is defined as

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

In eq 1, PN stands for period number. With the hydrogen atom being considered as the reference,  $\alpha$  for hydrogen is taken to be zero. One may note that  $\alpha$  values of different atoms (which are commonly found in organic compounds) have a high correlation ( $r = 0.946$ )<sup>11</sup> with (uncorrected) van der Waals volume.<sup>24</sup> Thus,  $\sum \alpha$  values of all non-hydrogen atoms of a molecule (instead of vertex count  $N_v$ ) may be taken as a gross measurement of molecular bulk.

Again, another term  $\epsilon$ , as a measure of electronegativity, has been defined<sup>11</sup> in the following manner:

$$\epsilon = -\alpha + 0.3Z^v \quad (2)$$

It is interesting to note that  $\epsilon$  has a good correlation ( $r = 0.937$ ) with Pauling's electronegativity scale (EN).<sup>11</sup>

The terms such as  $(\sum \alpha)_p / \sum \alpha$ ,  $(\sum \alpha)_v / \sum \alpha$ , and  $(\sum \alpha)_x / \sum \alpha$  can be used as shape parameters.  $(\sum \alpha)_p$ ,  $(\sum \alpha)_v$ , and  $(\sum \alpha)_x$  stand for summation of  $\alpha$  values of the vertices that are joined to one, three, and four other vertices, respectively, in the molecular graph.

For calculation of VEM count  $\beta$ , contribution of a sigma bond ( $x$ ) between two atoms of similar electronegativity ( $\Delta \epsilon \leq 0.3$ ) is considered to be 0.5, and for a sigma bond between two atoms of different electronegativity ( $\Delta \epsilon > 0.3$ ) it is considered to be 0.75. In the TAU scheme, contribution of all sigma bonds to VEM count was 0.5. Again, in case of pi bonds, contributions ( $y$ ) are considered depending on the type of double bond: (i) for the pi bond between two atoms of similar electronegativity ( $\Delta \epsilon \leq 0.3$ ),  $y$  is taken to be 1; (ii) for the pi bond between two atoms of different electronegativity ( $\Delta \epsilon > 0.3$ ) or for the conjugated (nonaromatic) pi system,  $y$  is considered to be 1.5; (iii) for the aromatic pi system,  $y$  is taken as 2. In the TAU scheme, the contribution of all kinds of pi bonds was 2. Thus  $\beta$  of the ETA scheme is defined as

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

In the above equation,  $\delta$  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.).

$\beta$  can be split into two parts,  $\beta_s$  (sigma contribution to VEM count) and  $\beta_{ns}$  (nonsigma contribution to VEM count) which may be defined as below:

$$\beta_s = \sum x v \quad (4)$$

$$\beta_{ns} = \sum y \pi + \delta \quad (5)$$

For a given part (substructure) of a molecular graph,  $\sum \beta_s$  and  $\sum \beta_{ns}$  may be calculated considering all bonds (sigma bonds for the former and pi bonds and lone pair of electrons for the latter) in the substructure.  $\sum \beta'_s$  (defined as  $\sum \beta_s / N_v$ ) may be taken as a relative measure of the number of electronegative atoms in the substructure, while  $\sum \beta'_{ns}$  (defined as  $\sum \beta_{ns} / N_v$ ) may be taken as a relative measure of electron-richness (unsaturation) of the substructure.

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

$$\gamma_i = \frac{\alpha_i}{\beta_i} \quad (6)$$

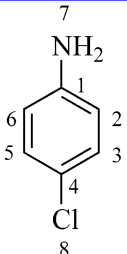
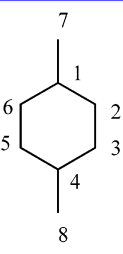
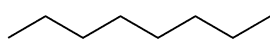
In the above equation,  $\alpha_i$  stands for the  $\alpha$  value for the  $i$ th vertex and  $\beta_i$  stands for the VEM count considering all bonds connected to the atom and lone pair of electrons (if any).

Finally, the composite index  $\eta$  is defined in the following manner:

$$\eta = \sum_{i < j} \left[ \frac{\gamma_i \gamma_j}{r_{ij}^2} \right]^{0.5} \quad (7)$$

In eq 7, both bonded and nonbonded interactions have been considered.  $r_{ij}$  stands for the topological distance between

**Table 1.** Calculations of ETA Parameters: Example of 4-Chloroaniline

<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>4-Chloroaniline</p> </div> <div style="text-align: center;">  <p>Reference alkane</p> </div> <div style="text-align: center;">  <p>Normal alkane</p> </div> </div>																	
vertex no.	4-chloroaniline								reference alkane								
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	
$\alpha_i$	0.5	0.5	0.5	0.5	0.5	0.5	0.4	0.72	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
$[\beta_s]_i$	1.75	1.00	1.00	1.75	1.00	1.00	0.75	0.75	1.50	1.00	1.00	1.50	1.00	1.00	0.50	0.50	
$[\beta_{ns}]_i$	2.00	2.00	2.00	2.00	2.00	2.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\beta_i$	3.75	3.00	3.00	3.75	3.00	3.00	1.25	1.25	1.50	1.00	1.00	1.50	1.00	1.00	0.50	0.50	
$\gamma_i$	0.13	0.17	0.17	0.13	0.17	0.17	0.32	0.58	0.33	0.50	0.50	0.33	0.50	0.50	1.00	1.00	
$[\eta]_i$	0.77	.075	0.76	0.82	0.76	0.75	0.73	0.95									
$[\eta_R]_i$									2.05	2.12	2.12	2.05	2.12	2.12	2.10	2.10	
$[\eta'_F]_i$	0.16	0.17	0.17	0.16	0.17	0.17	0.17	0.14									
$\eta$				3.141													
$\eta_R$												8.392					
$\eta'_F$				0.656													
$\eta^{\text{local}}$				1.413													
$\eta_R^{\text{local}}$												3.786					
$\eta_F^{\text{local}}$				2.373													
$\eta_N^{\text{local}}$												3.914					
$\eta'_B$				0.016 <sup>a</sup>													
$\Sigma\alpha$				4.115													
$[\Sigma\alpha]_p$				1.115													

<sup>a</sup> Without ring correction.

the  $i$ th atom and the  $j$ th atom. Thus, in addition to the local topology, global topology is also included in the formalism. Again, when all heteroatoms in the molecular graph are replaced by carbon and multiple bonds are replaced by a single bond, a corresponding molecular graph may be considered as the reference alkane, and the corresponding composite index value is designated as  $\eta_R$ . Considering functionality as the presence of heteroatoms (atoms other than carbon or hydrogen) and multiple bonds, functionality index  $\eta_F$  may be calculated as  $\eta_R - \eta$ . To avoid dependence of functionality on the vertex count or bulk, we have defined<sup>11</sup> another term  $\eta'_F$  as  $\eta_F/N_V$ . Again, one can determine the contribution of a particular position or vertex (within the common substructure in the congeneric series) to functionality in the following manner:

$$[\eta]_i = \sum_{j \neq i} \left[ \frac{\gamma_i \gamma_j}{r_{ij}^2} \right]^{0.5} \quad (8)$$

In eq 12,  $[\eta]_i$  stands for the contribution of the  $i$ th vertex to  $\eta$ . Similarly, the contribution of the  $i$ th vertex  $[\eta_R]_i$  to  $\eta_R$  can be computed. Contribution of the  $i$ th 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<sup>11</sup> as  $[\eta_F]_i/N_V$ .

Again, when only bonded interactions are considered ( $r_{ij} = 1$ ), the corresponding composite index is written as  $\eta^{\text{local}}$ .

$$\eta^{\text{local}} = \sum_{i < j, r_{ij}=1} (\gamma_i \gamma_j)^{0.5} \quad (9)$$

In the similar way,  $\eta_R^{\text{local}}$  for the corresponding reference alkane may also be calculated. The  $\eta_R^{\text{local}}$  value is similar to  $T_R$  of the reference alkane in the TAU scheme.

The local functionality contribution (without considering global topology),  $\eta_F^{\text{local}}$ , may be calculated as  $\eta_R^{\text{local}} - \eta^{\text{local}}$ .

For the calculation of branching, consideration of the local topology is sufficient. Branching is calculated with respect to the  $\eta$  value of the corresponding normal alkane (the straight chain compound of the same vertex count obtained from the reference alkane),  $\eta_N^{\text{local}}$ , which may be conveniently calculated as (when  $N_V \geq 3$ ):

$$\eta_N^{\text{local}} = 1.414 + (N_V - 3)0.5 \quad (10)$$

The branching index  $\eta_B$  can be calculated as  $\eta_N^{\text{local}} - \eta_R^{\text{local}} + 0.086N_R$ , 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. To calculate the branching contribution in comparison to the molecular size, another term  $\eta'_B$  is defined as  $\eta_B/N_V$ .

The calculation of different indices is illustrated using an example of 4-chloroaniline in Table 1.

In the present communication, the utility of ETA parameters has been demonstrated through a QSTR study using fish toxicity (pC) of a set of 92 substituted benzenes (taken from ref 12) as the model data set (Table 2). Different ETA

**Table 2.** Observed, Calculated, and Residual Fish Toxicity Values of Substituted benzenes

Sl. no.	compound name	obs <sup>a</sup>	calc <sup>b</sup>	res <sup>c</sup>	pred <sup>d</sup>	pres <sup>e</sup>	Sl. no.	compound name	obs <sup>a</sup>	calc <sup>b</sup>	res <sup>c</sup>	pred <sup>d</sup>	pres <sup>e</sup>
1	phenol	3.45	3.54	-0.09	3.55	-0.10	47	nitrobenzene	2.97	3.08	-0.11	3.15	-0.18
2	2-methylphenol	3.77	3.75	0.02	3.75	0.02	48	2-nitrotoluene	3.59	3.83	-0.24	3.86	-0.27
3	3-methylphenol	3.48	3.74	-0.26	3.76	-0.28	49	3-nitrotoluene	3.65	3.82	-0.17	3.83	-0.18
4	4-methylphenol	3.74	3.74	0.00	3.74	0.00	50	4-nitrotoluene	3.67	3.81	-0.14	3.82	-0.15
5	2,4-dimethylphenol	3.86	3.95	-0.09	3.96	-0.10	51	2,3-dimethylnitrobenzene	4.39	4.16	0.23	4.07	0.32
6	2,6-dimethylphenol	3.75	3.96	-0.21	3.97	-0.22	52	3,4-dimethylnitrobenzene	4.21	4.15	0.06	4.13	0.08
7	3,4-dimethylphenol	3.92	3.95	-0.03	3.95	-0.03	53	2-chloronitrobenzene	3.72	3.87	-0.15	3.93	-0.21
8	2,3,6-trimethylphenol	4.21	4.18	0.03	4.17	0.04	54	3-chloronitrobenzene	4.01	4.10	-0.09	4.11	-0.10
9	4-Ethylphenol	4.07	3.94	0.13	3.93	0.14	55	4-chloronitrobenzene	4.42	4.09	0.33	4.08	0.34
10	4-propylphenol	4.09	4.15	-0.06	4.15	-0.06	56	2,3-dichloronitrobenzene	4.66	4.53	0.13	4.49	0.17
11	4-butylphenol	4.47	4.36	0.11	4.35	0.12	57	2,4-dichloronitrobenzene	4.46	4.51	-0.05	4.52	-0.06
12	4-tert-butylphenol	4.46	4.37	0.09	4.36	0.10	58	2,5-dichloronitrobenzene	4.59	4.51	0.08	4.49	0.10
13	2-tert-butyl-4-methylphenol	4.90	4.61	0.29	4.58	0.32	59	3,5-dichloronitrobenzene	4.58	4.74	-0.16	4.75	-0.17
14	4-pentylphenol	5.12	4.57	0.55	4.53	0.59	60	2-chloro-6-nitrotoluene	4.52	4.46	0.06	4.45	0.07
15	4-tert-pentylphenol	4.81	4.58	0.23	4.56	0.25	61	4-chloro-2-nitrotoluene	4.44	4.46	-0.02	4.46	-0.02
16	2-allylphenol	3.96	4.18	-0.22	4.19	-0.23	62	aniline	2.91	2.79	0.12	2.71	0.20
17	2-phenylphenol	4.76	4.84	-0.08	4.85	-0.09	63	2-methylaniline	3.12	3.36	-0.24	3.38	-0.26
18	1-naphthol	4.50	4.40	0.10	4.39	0.11	64	3-methylaniline	3.47	3.36	0.11	3.36	0.11
19	4-chlorophenol	4.18	4.19	-0.01	4.19	-0.01	65	4-methylaniline	3.72	3.36	0.36	3.35	0.37
20	4-chloro-3-methylphenol	4.33	4.37	-0.04	4.37	-0.04	66	N,N-dimethylaniline	3.33	3.35	-0.02	3.35	-0.02
21	4-chloro-3,5-dimethylphenol	4.66	4.55	0.11	4.55	0.11	67	2-ethylaniline	3.21	3.60	-0.39	3.61	-0.40
22	3-methoxyphenol	3.22	3.35	-0.13	3.40	-0.18	68	3-ethylaniline	3.65	3.60	0.05	3.60	0.05
23	4-methoxyphenol	3.05	3.35	-0.30	3.47	-0.42	69	4-ethylaniline	3.52	3.60	-0.08	3.60	-0.08
24	4-phenoxyphenol	4.58	4.46	0.12	4.43	0.15	70	4-butylaniline	4.16	4.06	0.10	4.06	0.10
25	quinoline	3.63	3.83	-0.20	3.84	-0.21	71	2,6-diisopropylaniline	4.06	4.53	-0.47	4.59	-0.53
26	chlorobenzene	3.77	3.64	0.13	3.63	0.14	72	2-chloroaniline	4.31	3.82	0.49	3.81	0.50
27	1,2-dichlorobenzene	4.40	4.29	0.11	4.29	0.11	73	3-chloroaniline	3.98	3.81	0.17	3.81	0.17
28	1,3-dichlorobenzene	4.28	4.28	0.00	4.28	-0.00	74	4-chloroaniline	3.67	3.80	-0.13	3.81	-0.14
29	1,4-dichlorobenzene	4.56	4.27	0.29	4.26	0.30	75	2,4-dichloroaniline	4.41	4.45	-0.04	4.46	-0.05
30	1,2,3-trichlorobenzene	4.89	4.92	-0.03	4.92	-0.03	76	2,5-dichloroaniline	4.99	4.45	0.54	4.44	0.55
31	1,2,4-trichlorobenzene	4.83	4.90	-0.07	4.90	-0.07	77	3,4-dichloroaniline	4.39	4.46	-0.07	4.46	-0.07
32	1,3,5-trichlorobenzene	4.74	4.90	-0.16	4.90	-0.16	78	3,5-dichloroaniline	4.62	4.45	0.17	4.45	0.17
33	1,2,3,4-tetrachlorobenzene	5.35	5.51	-0.16	5.53	-0.18	79	2,3,4-trichloroaniline	5.15	5.09	0.06	5.08	0.07
34	1,2,3,5-tetrachlorobenzene	5.43	5.50	-0.07	5.51	-0.08	80	2,3,6-trichloroaniline	4.73	5.08	-0.35	5.10	-0.37
35	1,2,4,5-tetrachlorobenzene	5.85	5.50	0.35	5.46	0.39	81	2,4,5-trichloroaniline	4.92	5.07	-0.15	5.08	-0.16
36	3-chlorotoluene	3.84	3.83	0.01	3.83	0.01	82	ααα-4-tetrafluoro-3-methylaniline	3.77	3.91	-0.14	3.91	-0.14
37	4-chlorotoluene	4.33	3.83	0.50	3.82	0.51	83	ααα-4-tetrafluoro-2-methylaniline	3.78	3.91	-0.13	3.91	-0.13
38	2,4-dichlorotoluene	4.54	4.45	0.09	4.45	0.09	84	pentafluoroaniline	3.69	3.81	-0.12	3.81	-0.12
39	2,4,5-trichlorotoluene	5.06	5.06	0.00	5.06	0.00	85	2-nitroaniline	4.15	3.63	0.52	3.61	0.54
40	3,4,5-trichlorotoluene	4.60	5.07	-0.47	5.09	-0.49	86	3-nitroaniline	3.24	3.63	-0.39	3.64	-0.40
41	pentachlorotoluene	6.15	6.23	-0.08	6.24	-0.09	87	4-nitroaniline	3.23	3.63	-0.40	3.64	-0.41
42	benzene	3.09	2.95	0.14	2.93	0.16	88	2-chloro-4-nitraniline	3.93	4.29	-0.36	4.30	-0.37
43	toluene	3.13	3.18	-0.05	3.18	-0.05	89	4-bromoaniline	3.56	3.75	-0.19	3.76	-0.20
44	2-xylene	3.48	3.41	0.07	3.41	0.07	90	3-benzyloxyaniline	4.34	4.48	-0.14	4.52	-0.18
45	3-xylene	3.45	3.41	0.04	3.41	0.04	91	4-hexyloxyaniline	4.78	4.28	0.50	4.18	0.60
46	4-xylene	3.48	3.41	0.07	3.41	0.07	92	4-ethoxy-2-nitroaniline	3.85	3.79	0.06	3.77	0.08

<sup>a</sup> obs = observed [ref 12]. <sup>b</sup> calc = calculated (from eq 14). <sup>c</sup> res = residual = obs - calc. <sup>d</sup> pred = predicted (from eq 14). <sup>e</sup> pres = obs - pred.

descriptors calculated for the benzene derivatives are defined in Table 3. Factor analysis has been performed as a data preprocessing step for identification of important descriptors for the subsequent multiple regression analysis.<sup>25,26</sup> For this purpose, the data matrix consisting of the descriptors has been subjected to the principal component factor analysis using STATISTICA software.<sup>27</sup> The principal objectives of factor analysis are to display multidimensional data in a space of lower dimensionality with a 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 factors describing  $\geq 5\%$  of the total variance were considered. The analyses were carried out based on the following postulates: (a) only variables with nonzero loadings in such factors where biological activity also has nonzero loadings are important in explaining the variance of the activity; (b) only variables with nonzero loadings in different

factors may be combined in regression equations; and (c) the factor pattern indicates whether in the parameter space the biological activity can be explained in a satisfactory manner; if not, a different set of variables are to be chosen.

The calculations of  $\eta$ ,  $\eta_R$ ,  $\eta_F$ ,  $\eta_B$  and contributions of different vertices to  $\eta_F$  were done, using distance matrix and VEM vertex counts as inputs, by the GW-BASIC programs *KRETA1* and *KRETA2* developed by one of the authors.<sup>28</sup> We have also modeled the toxicity data using other topological descriptors and various physicochemical variables and compared the ETA models with non-ETA ones. The values for the topological descriptors and physicochemical variables for the compounds have been generated by QSAR+ and Descriptor+ modules of the Cerius 2 version 4.6 software.<sup>29</sup> The various topological indices calculated are Balaban J, connectivity indices ( $^0\chi$ ,  $^1\chi$ ,  $^2\chi$ ,  $^3\chi_p$ ,  $^3\chi_c$ ,  $^0\chi^v$ ,  $^1\chi^v$ ,  $^2\chi^v$ ,  $^3\chi^v_p$ ,  $^3\chi^v_c$ ), kappa shape indices ( $^1\kappa$ ,  $^2\kappa$ ,  $^3\kappa$ ,  $^1\kappa_\alpha$ ,  $^2\kappa_\alpha$ ,  $^3\kappa_\alpha$ ), flexibility index ( $\phi$ ), Wiener index (W), Zagreb index, and Hosoya index (log Z). Among the physicochemical variables, the



**Table 3.** Definitions of Different ETA Parameters Used in Exploring QSAR of Fish Toxicity of Substituted Benzenes

variables	definition
$\Sigma\alpha$	sum of $\alpha$ values of all non-hydrogen vertices of a molecule
$[\Sigma\alpha]_p$	sum of $\alpha$ values of all non-hydrogen vertices each of which is joined to only one other vertex of the molecule
$\Sigma\beta'_s$	sum of $\Sigma\beta'_s$ values, considering all bonds; $\Sigma\beta'_s$ is defined as $[\Sigma\beta_s]/N_v$
$\Sigma\beta'_{ns}$	sum of $\Sigma\beta'_{ns}$ values, considering all bonds; $\Sigma\beta'_{ns}$ is defined as $[\Sigma\beta_{ns}]/N_v$
$\eta$	the composite eta index
$\eta_R$	the composite index for the reference alkane
$[\eta'_F]_{OH}$	functionality for the hydroxy group
$[\eta'_F]_{NH_2}$	functionality for the amino group
$[\eta'_F]_{Cl}$	functionality for the chlorine atom
$[\eta'_F]_F$	functionality for the fluorine atom
$[\eta'_F]_{NO_2}$	functionality for the nitro group
$[\eta'_F]_{C-SP_3}$	functionality for the $sp_3$ hybridized carbon
$[\eta'_F]_{OEt}$	functionality for the ethoxy group
$[\eta'_F]_{NH_2-o-Cl}$	functionality for such amino group having ortho-chloro substitution
$[\eta'_F]_{NO_2-o-Cl}$	functionality for such nitro group having ortho-chloro substitution
$[\eta'_F]_{N-EXP}$	functionality for such amino or nitro group without having ortho-chloro substitution
$\eta_F^{local}$	local functionality
$[\eta'_F]_{C-SP_3-NO_2}$	functionality of such methyl substituent which is simultaneously present with a nitro substituent in the benzene ring
$[\eta'_F]_{N\_UNS}$	functionality of such nitro or amino group which is present in otherwise unsubstituted benzene ring
$\eta_B$	$= \eta_B/N_v$

following were considered: molar refractivity (MolRef), hydrophobicity (AlogP, AlogP98), HOMO and LUMO energies (HOMO, LUMO), and the number of hydrogen bond donor and acceptors (H\_bond\_acc, H\_bond\_don). The definitions of different ETA parameters used in this paper are given in Table 3.

The regression analyses were carried out using a program RRR98.<sup>28</sup> The statistical quality of the equations<sup>30</sup> was judged by the parameters such as *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* ( $df$ ). PRESS (leave-one-out) statistics<sup>31,32</sup> were calculated using the programs KRPRES1 and KRPRES2,<sup>27</sup> and *leave-one-out cross-validation*  $R^2$  ( $Q^2$ ), *predicted residual sum of squares* (PRESS), *standard deviation based on PRESS* ( $S_{PRESS}$ ), *standard deviation of error of prediction* (SDEP), and *average absolute predicted residual* ( $Pres_{av}$ ) were reported. Finally, “leave-many-out” cross-validation was applied on the final equations. All the accepted equations have regression constants and  $F$  ratios significant at 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-possible-subsets regression was also applied on the variables using the program AUTOREG<sup>28</sup> to cross-check appropriateness of selection of variables in the best models. The cutoff intercorrelation ( $|r|$ ) among the predictor variables in an equation was set to 0.4.

## RESULTS AND DISCUSSION

The results of the principal component factor analyses are given in Tables 4–8.

**Table 4.** Factor Loadings<sup>a</sup> of the Variables (Topological Parameters) after VARIMAX Rotation

variables	factor 1	factor 2	factor 3	factor 4	communality
pC	0.153	0.882 <sup>a</sup>	0.059	0.155	0.829
Balaban J	−0.123	0.080	−0.820 <sup>a</sup>	0.353	0.818
$^1\chi$	0.812 <sup>a</sup>	0.299	0.402	0.263	0.979
$^2\chi$	0.612	0.145	0.755 <sup>a</sup>	−0.099	0.975
$^3\chi$	0.368	0.023	0.916 <sup>a</sup>	0.098	0.985
$^1\chi_\alpha$	0.668	0.551	0.350	0.306	0.965
$^2\chi_\alpha$	0.463	0.492	0.722 <sup>a</sup>	−0.036	0.979
$^3\chi_\alpha$	0.224	0.371	0.875 <sup>a</sup>	0.163	0.979
$\phi$	0.364	0.632	0.632	0.068	0.936
$^0\chi$	0.845 <sup>a</sup>	0.319	0.322	0.271	0.992
$^1\chi$	0.850 <sup>a</sup>	0.254	0.450	0.060	0.992
$^2\chi$	0.834 <sup>a</sup>	0.273	0.253	0.390	0.986
$^3\chi_p$	0.880 <sup>a</sup>	0.447	0.047	0.055	0.979
$^3\chi_c$	0.465	0.181	−0.092	0.840 <sup>a</sup>	0.963
$^0\chi^v$	0.468	0.787 <sup>a</sup>	0.296	0.231	0.980
$^1\chi^v$	0.469	0.721 <sup>a</sup>	0.465	0.095	0.965
$^2\chi^v$	0.322	0.802 <sup>a</sup>	0.227	0.392	0.952
$^3\chi^v_p$	0.364	0.908 <sup>a</sup>	0.017	0.081	0.964
$^3\chi^v_c$	0.076	0.569	−0.129	0.750 <sup>a</sup>	0.909
Zagreb	0.886 <sup>a</sup>	0.311	0.241	0.220	0.989
Log Z	0.856 <sup>a</sup>	0.253	0.435	0.015	0.985
Wiener	0.757 <sup>a</sup>	0.214	0.589	0.047	0.967
% variance	0.361	0.254	0.245	0.097	0.958

<sup>a</sup> Factor loadings more than 0.7.

**Table 5.** Factor Loadings<sup>a</sup> of the Variables (Physicochemical Parameters) after VARIMAX Rotation

variables	factor 1	factor 2	factor 3	factor 4	factor 5	communality
pC	0.855 <sup>a</sup>	0.337	0.109	0.080	0.237	0.920
MolRef	0.477	0.050	0.137	−0.076	0.859 <sup>a</sup>	0.992
HOMO	−0.093	−0.123	0.195	−0.966 <sup>a</sup>	0.051	0.997
LUMO	0.010	−0.943 <sup>a</sup>	0.265	0.012	−0.110	0.972
H_bond_acc	0.335	0.879 <sup>a</sup>	0.747 <sup>a</sup>	0.223	−0.063	0.944
H_bond_don	−0.270	−0.166	0.897 <sup>a</sup>	−0.245	0.135	0.983
AlogP	0.905 <sup>a</sup>	0.019	−0.335	0.067	0.214	0.982
AlogP98	0.937 <sup>a</sup>	0.096	−0.248	0.087	0.162	0.982
% Variance	0.356	0.229	0.140	0.133	0.113	0.971

<sup>a</sup> Factor loadings more than 0.7.

**Table 6.** Factor Loadings<sup>a</sup> of the Variables (Topological and Physicochemical Parameters) after VARIMAX Rotation

variables	factor 1	factor 2	factor 3	factor 4	factor 5	communality
pC	0.820 <sup>a</sup>	0.256	0.305	0.149	−0.048	0.856
$^0\chi$	0.188	0.923 <sup>a</sup>	0.189	0.249	0.048	0.988
$^1\chi$	0.141	0.978 <sup>a</sup>	0.066	0.039	0.003	0.981
$^2\chi$	0.157	0.877 <sup>a</sup>	0.196	0.384	0.068	0.984
$^3\chi_p$	0.291	0.859 <sup>a</sup>	0.226	0.134	0.102	0.902
$^3\chi_c$	0.112	0.390	0.244	0.850 <sup>a</sup>	0.179	0.978
$^0\chi^v$	0.674	0.660	0.094	0.224	−0.185	0.983
$^1\chi^v$	0.631	0.731 <sup>a</sup>	−0.070	0.075	−0.199	0.983
$^2\chi^v$	0.716 <sup>a</sup>	0.511	−0.003	0.407	−0.232	0.992
$^3\chi^v_p$	0.793 <sup>a</sup>	0.462	0.158	0.142	−0.127	0.904
$^3\chi^v_c$	0.503	0.109	0.028	0.817 <sup>a</sup>	−0.204	0.975
MolRef	0.495	0.800 <sup>a</sup>	−0.064	0.079	−0.311	0.992
LUMO	−0.049	−0.085	−0.958 <sup>a</sup>	−0.049	0.172	0.960
H_bond_acc	0.249	0.217	0.843 <sup>a</sup>	0.192	0.315	0.955
AlogP	0.933 <sup>a</sup>	0.146	0.021	0.174	0.087	0.931
AlogP98	0.963 <sup>a</sup>	0.129	0.101	0.115	0.106	0.979
% Variance	0.325	0.358	0.123	0.123	0.030	0.959

<sup>a</sup> Factor loadings more than 0.7.

**Table 7.** Factor Loadings<sup>a</sup> of the Variables (ETA Parameters) after VARIMAX Rotation

variables	factor 1	factor 2	factor 3	factor 4	factor 5	factor 6	factor 7	factor 8	factor 9	factor 10	factor 11	communality
pC	0.649	0.145	0.524	0.154	-0.058	-0.332	0.162	-0.099	-0.058	0.159	0.059	0.923
$\alpha$	0.958 <sup>a</sup>	0.005	0.232	0.000	-0.002	-0.036	0.056	0.033	-0.018	-0.102	0.066	0.991
$[\Sigma\alpha]^2$	0.963 <sup>a</sup>	0.009	0.186	0.016	-0.025	-0.053	0.042	0.018	-0.006	-0.128	0.082	0.991
$\Sigma\beta'_s$	0.177	0.230	0.189	-0.911 <sup>a</sup>	0.019	0.137	-0.014	0.050	-0.031	-0.051	0.051	0.978
$\Sigma\beta'_{ns}$	-0.269	0.778 <sup>a</sup>	-0.144	-0.386	0.201	0.106	-0.041	0.141	-0.007	0.013	0.140	0.941
$[\Sigma\alpha]_p$	0.319	-0.026	0.920 <sup>a</sup>	-0.037	0.006	-0.068	0.039	0.039	-0.059	0.078	0.108	0.980
$[\Sigma\alpha]_p/\Sigma\alpha$	0.102	-0.028	0.971 <sup>a</sup>	-0.061	0.023	-0.048	0.062	0.039	-0.083	0.089	0.083	0.987
$\eta$	0.804 <sup>a</sup>	-0.503	0.181	-0.079	-0.057	-0.015	0.044	-0.053	0.040	0.002	-0.107	0.960
$\eta_R$	0.883 <sup>a</sup>	-0.093	0.120	-0.307	0.167	0.156	0.031	0.126	0.001	-0.120	-0.114	0.994
$[\eta/F]_{OH}$	-0.003	-0.155	-0.214	0.075	-0.133	-0.318	0.071	-0.102	0.089	-0.055	-0.857 <sup>a</sup>	0.956
$[\eta/F]_{NH_2}$	-0.020	-0.032	-0.092	-0.219	-0.306	0.659	-0.016	-0.116	-0.501	0.006	0.318	0.952
$[\eta/F]_{Cl}$	0.167	0.377	0.727 <sup>a</sup>	0.158	-0.168	-0.295	0.038	-0.002	-0.089	0.050	0.339	0.965
$[\eta/F]_F$	0.034	0.003	0.026	-0.985 <sup>a</sup>	-0.058	0.055	0.022	-0.051	0.032	0.048	0.019	0.985
$[\eta/F]_{NO_2}$	0.053	0.275	0.088	0.078	0.651	0.425	-0.087	0.476	0.076	-0.001	-0.050	0.940
$[\eta/F]_{C-SP_3}$	0.029	-0.894 <sup>a</sup>	-0.125	0.005	0.140	0.017	0.010	-0.042	0.105	-0.021	-0.052	0.851
$[\eta/F]_{OEI}$	0.251	-0.007	-0.234	0.016	-0.049	0.009	0.026	-0.035	0.020	-0.931 <sup>a</sup>	-0.046	0.991
$[\eta/F]_{NH_2-O-Cl}$	0.016	0.121	0.189	0.036	-0.025	-0.058	0.015	-0.041	-0.963 <sup>a</sup>	0.017	0.043	0.987
$[\eta/F]_{NO_2-O-Cl}$	0.067	0.106	0.077	-0.007	-0.021	-0.112	0.020	0.961 <sup>a</sup>	0.047	0.029	0.077	0.967
$[\eta/F]_{N-EXP}$	-0.013	0.071	-0.128	-0.129	0.253	0.900 <sup>a</sup>	-0.092	-0.084	0.149	-0.014	0.184	0.984
$[\eta/F]_{local}$	0.549	0.505	-0.145	-0.256	0.345	0.220	-0.019	0.236	-0.030	-0.263	-0.053	0.939
$[\eta/F]_{C-SP_3-NO_2}$	0.030	-0.090	0.017	0.008	0.946 <sup>a</sup>	0.024	0.045	-0.091	0.034	0.046	0.104	0.929
$[\eta/F]_{N-UNS}$	-0.126	0.035	-0.118	0.012	-0.019	0.083	-0.977 <sup>a</sup>	-0.010	0.012	0.022	0.051	0.997
$\eta_B$	0.232	-0.119	0.835 <sup>a</sup>	-0.341	0.144	0.134	0.043	0.088	-0.053	0.103	-0.155	0.968
% variance	0.190	0.100	0.161	0.103	0.077	0.082	0.045	0.057	0.055	0.045	0.049	0.963

<sup>a</sup> Factor loadings more than 0.7.

Table 4 shows the results of factor analysis of the data matrix involving topological parameters. Four factors could explain 95.8% of the variance of the data matrix. The results show that the biological activity is highly loaded with factor 2, which is in turn highly loaded in the variable  $^3\chi_p^v$ . The best equation obtained from the regression analysis is as follows:

$$pC = 1.074(\pm 0.153) ^3\chi_p^v + 2.474(\pm 0.250)$$

$$n = 92, Q^2 = 0.664, R_a^2 = 0.679, R^2 = 0.683,$$

$$r = 0.826, F = 193.5(df\ 1,90), s = 0.370,$$

$$AVRES = 0.295, SDEP = 0.377,$$

$$S_{PRESS} = 0.381, PRESS = 13.1, Pres_{av} = 0.303 \quad (11)$$

The 95% confidence intervals of the regression coefficients are shown within parentheses. In eq 11, the variable  $^3\chi_p^v$  could singularly predict 66.4% and explain 67.9% of the variance of the fish toxicity. The standard error and predicted standard error of the equation are 0.370 and 0.377, respectively. The average values of absolute residuals and absolute predicted residuals are 0.295 and 0.303, respectively. Eq 11 suggests the importance of branching and substitutions in the benzene ring for the toxicity.

Table 5 shows the results of factor analysis of the data matrix involving physicochemical parameters. Five factors could explain 97.1% of the variance of the data matrix. The results show that the biological activity is highly loaded with factor 1, which is in turn highly loaded in the variable ALOGP98 and ALOGP. Again, the activity is moderately loaded with factor 2, which is in turn highly loaded in LUMO and H\_bond\_acc. However, as H\_bond\_acc has considerable loading in factor 1 also, LUMO is considered as a better predictor variable than H\_bond\_acc. The best equation obtained from the regression analysis is as follows:

$$pC = 0.599(\pm 0.086)A \log P98 -$$

$$0.091(\pm 0.053)LUMO + 2.772(\pm 0.285)$$

$$n = 92, Q^2 = 0.699, R_a^2 = 0.711, R^2 = 0.718,$$

$$R = 0.847, F = 113.1(df\ 2,89), s = 0.351,$$

$$AVRES = 0.269, SDEP = 0.357,$$

$$S_{PRESS} = 0.363, PRESS = 11.7, Pres_{av} = 0.278 \quad (12)$$

Eq 12 could predict and explain 69.9% and 71.1% of the variance of fish toxicity. The standard error and predicted standard error of the equation are 0.351 and 0.357, respectively, while the average values of absolute residuals and absolute predicted residuals are 0.269 and 0.278, respectively. The positive coefficient of AlogP98 indicates the positive contribution of lipophilicity to the fish toxicity. The negative coefficient of LUMO suggests that the toxicity rises with an increase in electron donating capacity of the compounds.

Using important topological and physicochemical variables (as evident from Tables 4 and 5) in combination, factor analysis was performed. The results of which are presented in Table 6. The results show that the biological activity is highly loaded with factor 1, which is in turn highly loaded in the variables AlogP98 and AlogP. Again, the activity is moderately loaded with factor 3, which is in turn highly loaded in LUMO and H\_bond\_acc. Another important factor with which the biological activity shows high loading is factor 2, which is highly loaded in  $^1\chi$ . The best equation obtained from the regression analysis is as follows:

$$pC = 0.123(\pm 0.093) ^1\chi + 0.569(\pm 0.086)A \log P98 -$$

$$0.083(\pm 0.052)LUMO + 2.264(\pm 0.508)$$

$$n = 92, Q^2 = 0.718, R_a^2 = 0.730, R^2 = 0.738,$$

$$R = 0.859, F = 82.8(df\ 3,88), s = 0.340,$$

$$AVRES = 0.254, SDEP = 0.345,$$

$$S_{PRESS} = 0.353, PRESS = 11.0, Pres_{av} = 0.265 \quad (13)$$

**Table 8.** Factor Loadings<sup>a</sup> of the Variables (ETA and non-ETA Parameters) after VARIMAX Rotation

variables	factor 1	factor 2	factor 3	factor 4	factor 5	factor 6	factor 7	factor 8	factor 9	factor 10	factor 11	communality
pC	0.578	0.077	0.548	-0.162	0.148	-0.390	0.163	0.108	-0.083	0.161	0.051	0.914
$\alpha$	0.943 <sup>a</sup>	0.021	0.274	-0.030	0.028	-0.103	0.054	-0.011	-0.021	-0.070	0.087	0.993
$[\Sigma\alpha]^2$	0.947	0.046	0.227	-0.045	0.029	-0.122	0.041	0.004	-0.011	-0.097	0.103	0.990
$\Sigma\beta'_s$	0.206	-0.015	0.193	0.902 <sup>a</sup>	0.244	0.137	-0.014	-0.043	-0.026	-0.043	0.055	0.978
$\Sigma\beta'_{ns}$	-0.256	-0.196	-0.165	0.393	0.780 <sup>a</sup>	0.109	-0.040	-0.126	-0.009	0.005	0.132	0.941
$[\Sigma\alpha]_p$	0.266	-0.011	0.934 <sup>a</sup>	0.036	-0.013	-0.110	0.038	-0.036	-0.066	0.072	0.107	0.981
$[\Sigma\alpha]_p/\Sigma\alpha$	0.052	-0.032	0.973 <sup>a</sup>	0.067	-0.019	-0.075	0.061	-0.040	-0.089	0.074	0.077	0.986
$\eta$	0.804 <sup>a</sup>	0.068	0.222	0.053	-0.488	-0.056	0.041	0.060	0.042	0.032	-0.084	0.960
$\eta_R$	0.911 <sup>a</sup>	-0.150	0.157	0.274	-0.066	0.111	0.029	-0.100	0.007	-0.081	-0.085	0.994
$[\eta'_F]_{OH}$	0.007	0.146	-0.219	-0.076	-0.171	-0.261	0.071	0.092	0.075	-0.052	-0.875 <sup>a</sup>	0.960
$[\eta'_F]_{NH_2}$	0.011	0.305	-0.075	0.206	-0.030	0.678	-0.014	0.118	-0.469	0.019	0.353	0.961
$[\eta'_F]_{Cl}$	0.084	0.171	0.727 <sup>a</sup>	-0.149	0.374	-0.344	0.039	0.002	-0.109	0.031	0.322	0.963
$[\eta'_F]_F$	0.066	0.059	0.024	0.982 <sup>a</sup>	0.005	0.066	0.021	0.047	0.035	0.051	0.021	0.984
$[\eta'_F]_{NO_2}$	0.099	-0.664	0.090	-0.091	0.313	0.405	-0.089	-0.446	0.096	0.019	-0.028	0.946
$[\eta'_F]_{C-SP_3}$	0.059	-0.155	-0.112	-0.011	-0.885 <sup>a</sup>	0.036	0.009	0.026	0.115	-0.016	-0.048	0.842
$[\eta'_F]_{OEI}$	0.298	0.061	-0.234	-0.029	-0.005	0.027	0.025	0.039	0.020	-0.915 <sup>a</sup>	-0.049	0.992
$[\eta'_F]_{NH_2-O-Cl}$	-0.000	0.030	0.187	-0.036	0.117	-0.023	0.016	0.039	-0.963 <sup>a</sup>	0.016	0.043	0.982
$[\eta'_F]_{NO_2-O-Cl}$	0.072	0.011	0.076	0.004	0.124	-0.105	0.019	-0.960 <sup>a</sup>	0.040	0.027	0.065	0.965
$[\eta'_F]_{N-EXP}$	0.049	-0.262	-0.111	0.111	0.095	0.877 <sup>a</sup>	-0.091	0.109	0.190	0.011	0.232	0.984
$[\eta'_F]_{local}$	0.590	-0.324	-0.130	0.232	0.528	0.188	-0.020	-0.200	-0.026	-0.231	-0.033	0.934
$[\eta'_F]_{C-SP_3-NO_2}$	0.044	-0.940 <sup>a</sup>	0.007	-0.003	-0.082	0.004	0.046	0.106	0.021	0.040	0.094	0.917
$[\eta'_F]_{N-UNS}$	-0.122	0.018	-0.122	-0.011	0.033	0.089	-0.977 <sup>a</sup>	0.011	0.014	0.021	0.054	0.997
$\eta'_B$	0.221	-0.149	0.843 <sup>a</sup>	0.338	-0.105	0.117	0.041	-0.081	-0.052	0.104	-0.147	0.964
$^1\chi$	0.955 <sup>a</sup>	-0.120	-0.095	0.141	-0.030	0.093	0.030	-0.079	0.017	-0.137	-0.061	0.995
AlogP98	0.462	0.075	0.498	-0.129	-0.081	-0.636	0.065	0.084	0.183	0.157	0.154	0.988
LUMO	-0.165	0.464	-0.488	0.005	-0.567	-0.159	0.006	0.272	-0.047	-0.071	-0.172	0.938
% variance	0.210	0.077	0.167	0.090	0.102	0.090	0.040	0.052	0.050	0.039	0.047	0.963

<sup>a</sup> Factor loading more than 0.7.**Table 9.** Results of All-Possible-Subsets Regression Applied on Different Parameters<sup>a</sup>

type of variables	variables	no. of predictor variables	variables	statistics			
				<i>r</i> or <i>R</i>	$R_a^2$	<i>F</i>	<i>s</i>
topological	as listed in Table 4	1	$^3\chi'_p$	0.826	0.679	193.5	0.370
physicochemical	as listed in Table 5	1	AlogP98	0.825	0.677	192.1	0.371
		2	AlogP98, H_bond_acc	0.858	0.729	123.7	0.340
topological and physicochemical	as listed in Table 6	1	$^3\chi'_p$	0.826	0.679	193.5	0.370
		2	AlogP98, H_bond_acc	0.858	0.729	123.7	0.340
		3	AlogP98, H_bond_acc, $^1\chi$	0.866	0.741	87.9	0.333
ETA/ETA and non-ETA	as listed in Table 7/8, except $[\Sigma\alpha]^2$	1	$\Sigma\alpha^b$	0.738	0.541	108.0	0.443
		2	$\eta$ , $[\eta'_F]_{Cl}$	0.888	0.783	165.3	0.304
		3	$\Sigma\alpha$ , $[\eta'_F]_{Cl}$ , $[\eta'_F]_{OH}$	0.915	0.832	150.9	0.268
		4	$\Sigma\alpha$ , $[\eta'_F]_{Cl}$ , $[\eta'_F]_{OH}$ , $[\eta'_F]_{OEI}$	0.931	0.860	140.4	0.245
		5	$\Sigma\alpha$ , $[\eta'_F]_{Cl}$ , $[\eta'_F]_{OH}$ , $[\eta'_F]_{OEI}$ , $[\eta'_F]_{N-UNS}$	0.935	0.867	119.5	0.239
		6	$\Sigma\alpha$ , $[\eta'_F]_{Cl}$ , $[\eta'_F]_{OH}$ , $[\eta'_F]_{OEI}$ , $[\eta'_F]_{N-UNS}$ , $[\eta'_F]_{C-SP_3-NO_2}$	0.938	0.871	103.7	0.235
		7	$\Sigma\alpha$ , $[\eta'_F]_{Cl}$ , $[\eta'_F]_{OH}$ , $[\eta'_F]_{OEI}$ , $[\eta'_F]_{N-UNS}$ , $[\eta'_F]_{C-SP_3-NO_2}$ , $[\eta'_F]_{NO_2-O-Cl}$	0.941	0.875	0.926	0.230

<sup>a</sup> Cutoff intercorrelation among predictor variables = 0.4. <sup>b</sup> Considering only ETA parameters.

Eq 13 is statistically superior to eq 12: the predicted variance and explained variance have increased to 71.8% and 73.0%, respectively.

Table 7 shows factor analysis of the data matrix involving ETA parameters. Relative importance of different factors (and corresponding variables) in explaining the biological activity data are evident from the table. The best equation involving ETA parameters obtained from regression analysis is as follows:

$$pC = 0.466(\pm 0.069)\Sigma\alpha + 2.352(\pm 0.373)[\eta'_F]_{Cl} - 1.885(\pm 1.632)[\eta'_F]_{N-UNS} - 1.379(\pm 1.367)[\eta'_F]_{NO_2-O-Cl} - 3.635(\pm 1.655)[\eta'_F]_{OEI} + 2.302(\pm 0.760)[\eta'_F]_{OH} + 1.357(\pm 1.423)*[\eta'_F]_{C-SP_3-NO_2} + 1.548(\pm 0.343)$$

$$n = 92, Q^2 = 0.865, R_a^2 = 0.876, R^2 = 0.885,$$

$$R = 0.941, F = 92.6(df\ 7,84), s = 0.230,$$

$$AVRES = 0.168, SDEP = 0.239, S_{PRESS} = 0.250,$$

$$PRESS = 5.3, Pres_{av} = 0.184 \quad (14)$$

The regression coefficient of  $[\eta'_F]_{C-SP_3-NO_2}$  in eq 14 is significant at the 90% level. Equation 14 is statistically much superior than the non-ETA relations. The predicted variance and explained variance of eq 14 are 86.5% and 87.6%, respectively. The calculated, predicted, residual, and predicted residual toxicity values according to eq 14 are given in Table 2.

Table 8 shows factor analysis of the data matrix involving important ETA and non-ETA parameters. The best equation

**Table 10.** Results of Leave-Many-Out Cross-Validation Applied on Eq 14<sup>c,d</sup>

type of cross-validation	number of cycles	average regression coefficients (standard deviations)	statistics
			$Q^2$ (average Pres)
leave-10%-out	10 <sup>a</sup>	0.465 (0.015) $\sum \alpha$ + 2.356 (0.071) $[\eta'_F]_{Cl}$ − 1.883 (0.172) $[\eta'_F]_{N\_UNS}$ − 1.376 (0.171) $[\eta'_F]_{NO_2-o-Cl}$ − 3.614 (0.394) $[\eta'_F]_{OEt}$ + 2.306 (0.098) $[\eta'_F]_{OH}$ + 1.345 (0.232) $[\eta'_F]_{C-SP_3-NO_2}$ + 1.550 (0.067)	0.864 (0.184)
leave-25%-out	4 <sup>b</sup>	0.465 (0.023) $\sum \alpha$ + 2.357 (0.122) $[\eta'_F]_{Cl}$ − 1.923 (0.405) $[\eta'_F]_{N\_UNS}$ − 1.332 (0.398) $[\eta'_F]_{NO_2-o-Cl}$ − 3.605 (0.512) $[\eta'_F]_{OEt}$ + 2.314 (0.159) $[\eta'_F]_{OH}$ + 1.351 (0.527) $[\eta'_F]_{C-SP_3-NO_2}$ + 1.549 (0.145)	0.865 (0.185)

<sup>a</sup> Compounds were deleted in 10 cycles in the following manner: (1, 11, 21,...91), (2, 12, 22,...92), ..... (10, 20, 30,...90). <sup>b</sup> Compounds were deleted in 4 cycles in the following manner: (1, 5, 9,...89), (2, 6, 10,...90), ..... (4, 8, 12,...92). <sup>c</sup>  $Q^2$  denotes cross-validated  $R^2$ . Average Pres means average of absolute values of predicted residuals. <sup>d</sup> Model equation,  $pC = \sum \beta_i x_i + \alpha$ .

involving ETA and non-ETA parameters obtained from the regression analysis is as follows:

$$pC = 0.474(\pm 0.073)\chi + 3.297(\pm 0.356)[\eta'_F]_{Cl} - 2.377(\pm 1.714)[\eta'_F]_{N\_UNS} - 2.271(\pm 1.456)[\eta'_F]_{NO_2-o-Cl} - 4.750(\pm 1.796)[\eta'_F]_{OEt} + 2.193(\pm 0.789)[\eta'_F]_{OH} - 0.538(\pm 0.404)[\eta'_F]_F + 1.571(\pm 0.345)$$

$$n = 92, Q^2 = 0.852, R_a^2 = 0.861, R^2 = 0.872, R = 0.934, F = 81.8(df\ 7,84), s = 0.243, AVRES = 0.178, SDEP = 0.250, S_{PRESS} = 0.262, PRESS = 5.8, Pres_{av} = 0.193 \quad (15)$$

The statistical quality of eq 15 is slightly inferior to that of eq 14.

To crosscheck the correctness of the selection of variables in multivariate relations, an all-possible-subsets regression technique (with a restriction that the predictor variables in a equation are less intercorrelated:  $|r| < 0.4$ ) was applied on different parameters as listed in Table 9. It is observed that in all the cases the same best models could be obtained after factor analysis and all-possible-subsets regression except that H\_bond\_acc replaces LUMO in case of all-possible-subsets regression (vide eqs 12 and 13). However, the reason of selection of LUMO instead of H\_bond\_acc in eqs 12 and 13 has been explained previously.

The ETA relations (eqs 14 and 15) show positive contributions of molecular bulk (size), chloro and hydroxy substitutions in the benzene ring, and the simultaneous presence of methyl and nitro substitutions to the toxicity. Further, the presence of fluoro and ether functionality, amino or nitro functionality in an otherwise unsubstituted ring, and nitro functionality which is ortho to a chloro substituent decreases toxicity. An attempt to use non-ETA descriptors along with ETA ones did not improve the quality in comparison to the best ETA model (vide eq 15 vs eq 14). Interestingly, the ETA model developed by us for the fish toxicity is better than the previously reported models on the same data set.<sup>12,13</sup>

In the present study, ETA indices could explore the important chemical information contributing to the fish toxicity of substituted benzenes and the relations generated could calculate the activity of the compounds to a satisfactory

extent (predicted variance up to 86.5%, explained variance up to 87.6%). "Leave-10%-out" and "leave-25%-out" cross-validations applied to eq 14 (Table 10) show robustness of the final relation. Thus, it appears that ETA descriptors have significant potential in QSAR/QSPR/QSTR studies, which warrants extensive evaluation.

#### ACKNOWLEDGMENT

A financial grant from J. U. Research Fund is thankfully acknowledged.

#### REFERENCES AND NOTES

- (1) Escher, B. I.; Hermens, J. L. Modes of Action in Ecotoxicology: Their Role in Body Burdens, Species Sensitivity, QSARs, and Mixture Effects. *Environ. Sci. Technol.* **2002**, *36*, 4201–4217.
- (2) Walker, J. D.; Carlsen, L.; Hulzebos, E.; Simon-Hettich, B. Global Government Applications of Analogues, SARs and QSARs to Predict Aquatic Toxicity, Chemical or Physical Properties, Environmental Fate Parameters and Health Effects of Organic Chemicals. *SAR QSAR Environ. Res.* **2002**, *13*, 607–616.
- (3) Bradbury, S. P. Quantitative Structure–Activity Relationships and Ecological Risk Assessment: an Overview of Predictive Aquatic Toxicology Research. *Toxicol. Lett.* **1995**, *79*, 229–237.
- (4) Vighi, M.; Gramatica, P.; Consolaro, F.; Todeschini, R. QSAR and Chemometric Approaches for Setting Water Quality Objectives for Dangerous Chemicals. *Ecotoxicol. Environ. Saf.* **2001**, *49*, 206–220.
- (5) Salvito, D. T.; Senna, R. J.; Federle, T. W. A Framework for Prioritizing Fragrance Materials for Aquatic Risk Assessment. *Environ. Toxicol. Chem.* **2002**, *21*, 1301–1308.
- (6) Katritzky, A. R.; Tatham, D. B.; Maran, U. Theoretical Descriptors for the Correlation of Aquatic Toxicity of Environmental Pollutants by Quantitative Structure–Toxicity Relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1162–1176.
- (7) Estrada, E.; Uriarte, E. Quantitative Structure–Toxicity Relationships Using TOPS-MODE.1. Nitrobenzene Toxicity to *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* **2001**, *12*, 309–324.
- (8) Yu, R. L.; Hu, G. R.; Zhao, Y. H. Comparative Study of Four QSAR Models of Aromatic Compounds to Aquatic Organisms. *J. Environ. Sci. (China)* **2002**, *14*, 552–557.
- (9) Dimitrov, S. D.; Mekenyan, O. G.; Walker, J. D. Nonlinear Modeling of Bioconcentration Using Partition Coefficients for Narcotic Chemicals. *SAR QSAR Environ. Res.* **2002**, *13*, 177–184.
- (10) Kulkarni, S. A.; Raje, D. V.; Chakrabarti, T. Quantitative Structure–Activity Relationships Based on Functional and Structural Characteristics of Organic Compounds. *SAR QSAR Environ. Res.* **2001**, *12*, 565–591.
- (11) Roy, K.; Ghosh, G. Introduction of Extended Topochemical Atoms (ETA) Indices in the Valence Electron Mobile (VEM) Environment as Tool for QSAR/QSPR Studies. *Internet Electron. J. Mol. Des.* **2003**, *2*, 599–620, <http://www.biochempress.com>.
- (12) Rose, K.; Hall, L. H. E-State Modeling of Fish Toxicity Independent of 3D Structure Information. *SAR QSAR Environ. Res.* **2003**, *14*, 113–129.
- (13) Robert, D. and Carbo-Dorca, R. Aromatic compounds aquatic toxicity QSAR using molecular quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401–422.



- (14) Pal, D. K.; Sengupta, C.; De, A. U. A New Topochemical Descriptor (TAU) in Molecular Connectivity Concept: Part I -- Aliphatic Compounds. *Indian J. Chem.* **1988**, *27B*, 734–739.
- (15) Pal, D. K.; Sengupta, C.; De, A. U. Introduction of A Novel Topochemical Index and Exploitation of Group Connectivity Concept to Achieve Predictability in QSAR and RDD. *Indian J. Chem.* **1989**, *28B*, 261–267.
- (16) Pal, D. K.; Sengupta, M.; Sengupta, C.; De, A. U. QSAR with TAU ( $\tau$ ) indices: Part I - - Polymethylene Primary Diamines as Amebicidal Agents. *Indian J. Chem.* **1990**, *29B*, 451–454.
- (17) Pal, D. K.; Purkayastha, S. K.; Sengupta, C.; De, A. U. Quantitative Structure–Property Relationships with TAU indices: Part I – Research Octane Numbers of Alkane Fuel Molecules. *Indian J. Chem.* **1992**, *31B*, 109–114.
- (18) Roy, K.; Pal, D. K.; De, A. U.; Sengupta, C. Comparative QSAR with Molecular Negentropy, Molecular Connectivity, STIMS and TAU Indices: Part I. Tadpole Narcosis of Diverse Functional Acyclic Compounds. *Indian J. Chem.* **1999**, *38B*, 664–671.
- (19) Roy, K.; Pal, D. K.; De, A. U.; Sengupta, C. Comparative QSAR Studies with Molecular Negentropy, Molecular Connectivity, STIMS and TAU Indices. Part II: General Anaesthetic Activity of Aliphatic Hydrocarbons, Halocarbons and Ethers. *Indian J. Chem.* **2001**, *40B*, 129–135.
- (20) Roy, K.; Saha, A. Comparative QSPR Studies with Molecular Connectivity, Molecular Negentropy and TAU Indices. Part I: Molecular Thermochemical Properties of Diverse Functional Acyclic Compounds. *J. Mol. Model.* **2003**, *9*, 259–270.
- (21) Roy, K.; Saha, A. Comparative QSPR Studies with Molecular Connectivity, Molecular Negentropy and TAU Indices. Part 2: Lipid-Water Partition Coefficient of Diverse Functional Acyclic Compounds. *Internet Electron. J. Mol. Des.* **2003**, *2*, 288–305, <http://www.biochempress.com>.
- (22) Roy, K.; Saha, A. QSPR with TAU Indices: Water Solubility of Diverse Functional Acyclic Compounds. *Internet Electron. J. Mol. Des.* **2003**, *2*, 475–491, <http://www.biochempress.com>.
- (23) Roy, K.; Chakraborty, S.; Ghosh, C. C.; Saha, A. QSPR with TAU Indices: Molar Thermochemical Properties of Diverse Functional Acyclic Compounds. *J. Indian Chem. Soc.* **2003**, *80*, accepted for publication.
- (24) Moriguchi, I.; Canada, Y.; Komatsu, K. van der Waals Volume and the Related Parameters for Hydrophobicity in Structure–Activity Studies. *Chem. Pharm. Bull.* **1976**, *24*, 1799–1806.
- (25) Lewi, P. J. In *Drug Design*; Ariens, E. J., Ed.; Academic Press: New York, 1980; Vol. 10, pp 307–342.
- (26) Franke, R.; Gruska, A. In *Chemometric Methods in Molecular Design*; Waterbeemd, H. van de, Ed.; VCH: Weinheim, 1995; Vol. 2, pp 113–163.
- (27) STATISTICA is a statistical software of Statsoft Inc., USA.
- (28) The GW-BASIC programs *RRR98*, *KRETA1*, *KRETA2*, *AUTOREG*, *KRPRES1*, and *KRPRES2* were developed by Kunal Roy and standardized using known data sets.
- (29) Cerius 2 version 4.6 is a product of Accelrys Inc., San Diego, CA.
- (30) Snedecor, G. W.; Cochran, W. G. *Statistical Methods*; Oxford & IBH Publishing Co. Pvt. Ltd.: New Delhi, 1967; pp 381–418.
- (31) Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*; Waterbeemd, H. van de, Ed.; VCH: Weinheim 1995; pp 309–318.
- (32) Debnath, A. K. In *Combinatorial Library Design and Evaluation*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001; pp 73–129.

CI0342066