Internet Electronic Journal of Molecular Design **2003**, *2*, 599–620 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

September 2003, Volume 2, Number 9, Pages 599–620

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday
Part 3

Guest Editor: Douglas J. Klein

Introduction of Extended Topochemical Atom (ETA) Indices in the Valence Electron Mobile (VEM) Environment as Tools for QSAR/QSPR Studies

Kunal Roy and Gopinath Ghosh

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

Received: April 24, 2003; Revised: May 22, 2003; Accepted: May 26, 2003; Published: September 30, 2003

Citation of the article:
K. Roy and G. Ghosh, Introduction of Extended Topochemical Atom (ETA) Indices in the Valence Electron Mobile (VEM) Environment as Tools for QSAR/QSPR Studies, *Internet Electron. J. Mol. Des.* **2003**, *2*, 599–620, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Introduction of Extended Topochemical Atom (ETA) Indices in the Valence Electron Mobile (VEM) Environment as Tools for QSAR/QSPR Studies[#]

Kunal Roy* and Gopinath Ghosh

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

Received: April 24, 2003; Revised: May 22, 2003; Accepted: May 26, 2003; Published: September 30, 2003

Internet Electron. J. Mol. Des. **2003**, *2* (9), 599–620

Abstract

In order 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 have been redefined and novel formalism of extended topochemical atom (ETA) indices has been presented. Apart from considering size, shape, branching and functionality contributions of a molecular graph, the ETA formalism also determines contributions of specific vertices or positions within common substructures of molecular graph towards total functionality. To explore utility of the newly developed ETA descriptors, toxicity of substituted phenols (n = 50) against *Tetrahymena pyriformis* was taken as the model data set. Principal component factor analysis was employed as the data–preprocessing step to select appropriate descriptors, which are devoid of collinearities, for the subsequent regression analyses. Statistical quality of the multivariate relations was judged by the parameters such as predicted variance (Q^2), explained variance (R_a^2), correlation coefficient (R), variance ratio (F), predicted residual sum of squares (*PRESS*), etc. Multiple regression of the response variable (toxicity) with combination of ETA descriptors (up to five or six independent variables) using least square method generated statistically robust equations that could predict (up to 94.5%) and explain (up to 95%) the variance to a significant extent. Final relations were subjected to leave–10%–out cross–validation, which shows Q^2 value of about 93%. The results suggest that the toxicity of substituted phenols increases with increase in molecular bulk (which showing parabolic relation, there exists an optimum size), branching and number of substitutions in the phenol nucleus. Again, presence of electronegative atoms in the substituent positions increases toxicity. Further, requirements of electronic features for the *meta* and *para* substituents were also found out. The study further shows the importance of the phenolic –OH group for the toxicity. Statistically robust relations generated in the study suggest that ETA indices merit further assessment to prove their utility in modeling studies.

Keywords. QSAR; quantitative structure–activity relationships; QSPR; quantitative structure–property relationships; structural descriptor; topological index; molecular graph; toxicity prediction; phenol toxicity; *Tetrahymena pyriformis*.

[#] Dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday.

* Correspondence author; phone: 91–33–2414 6676 (O), 91–33–2435 6547 (R); E–mail: kunalroy_in@yahoo.com, URL: http://www.geocities.com/kunalroy_in.

1 INTRODUCTION

The goal of any medicinal chemist has been to design and synthesize novel pharmacologically active molecules with reduced toxicity. Trial and error screening can no longer be relied upon as such process is costly and time consuming. Quantitative structure–activity/property/toxicity relationship (QSAR/QSPR/QSTR) techniques increase the probability of success and reduce time and cost involvement in drug discovery process [1,2].

Biological activity of drug molecules results from their specific interactions with receptor sites using intermolecular physicochemical forces [2–4]. The magnitude and type of such interactions are in turn dependent on structural features of drug compounds. Exploration of relation of biological activity with physicochemical properties or structural descriptors of drugs helps in identifying mechanistic features of drug action apart from optimization of drug structures for getting better analogues.

A long–standing goal in chemistry was to represent chemical structure in numerical form. Topological indices have been developed by scientists in pursuit of this goal [5]. These descriptors encode chemical information about structural environment of atoms, bonds, branching, unsaturation, heteroatom variation, cyclicity, aromatic nature, *etc* [6,7]. The descriptors are formulated in graph theoretic approach considering connection table of the vertices (*e.g*., molecular connectivity indices, Gordon–Scantelbury index, *etc*.) or distance matrix (Wiener index, kappa shape index, electrotopological state atom index, *etc*.) [8]. Apart from these, centric topological indices (based on the sequences of numbers obtained by pruning an acyclic graph towards its center, *e.g*., Balaban centric indices) and indices based on information theory (*e.g*., redundancy index, molecular negentropy, *etc*.) have been reported [8].

Despite simplicity of most of the topological descriptors, they have proved to contain a great quantity of chemical information [9,10]. Recent advances in the development and applications of graph–theoretic descriptors in physical chemistry and pharmaceutical sciences indicate that this approach is an important alternative to physicochemical parameters in QSPR/QSAR models, maintaining both statistical and physical meaning. These indices not only contain information on molecular topological features, but also on nature of atoms or bonds as well as electronic features of molecules as a whole. However, since most of the topological descriptors are global ones, they do not contain explicit information regarding number of functionality, pharmacophore, interatomic distance, charge distribution, stereochemistry or electrostatic potential [11]. Nevertheless, as different structural aspects are decoded by different indices, a rich database of topological indices may be helpful in exploring QSAR/QSPR [7,12–21] and classifying chemical compounds for a target property/activity [22–24].

In late eighties, TAU descriptors were reported [25,26] 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 have been published in support of the claim [27–33]. In the present paper, we are redefining some of the basic parameters introduced in TAU scheme and presenting a new formalism for calculation of Extended Topochemical Atom (ETA) indices to fill some of the lacunae present in the TAU scheme.

2 MATERIALS AND METHODS

TAU descriptors were developed in the valence electron mobile (VEM) environment [25,26], where a vertex in the molecular graph is considered to be composed of a core and a valence electronic environment. The core environment, identified as λ, is defined by the ratio of the number of core electrons to the number of valence electrons, *i.e.*:

$$\lambda = \frac{Z - Z^V}{Z^V} \tag{1}$$

In Eq. (1), Z and Z^v represent atomic number and valence electron number respectively. Obviously, 1/λ roughly corresponds to the strength of the positive field of the atomic core.

In TAU scheme, valence electronic environment is partitioned into two components, localized (VEL) and mobile (VEM). The bond with hydrogen in the graph theoretic method changes into a self–loop which implies that the pair of electrons forming a covalent bond with a hydrogen atom is predominantly enjoyed by the atom to which it is bonded. Again, in the TAU scheme it is assumed that an atom enjoys besides its own, fifty percent of the other electron in a sigma bond with a non–hydrogen atom. Further, π–bond and σ bond are given unequal weights. The mobile valence electronic environment, identified as θ, is defined as:

$$\theta = 8 - (2h + 1.5\nu + 2l) \tag{2}$$

When unsaturation is present, θ is defined as:

$$\theta = 0.5\nu + 2\pi \tag{3}$$

In Eqs. (2) and (3), the notations ν, h, l and π represent the numbers of sigma bonds (other than hydrogen), hydrogen atoms, loan pair of electrons and π bonds associated with the atom in that order.

VEM vertex weight V_i of the vertex i is defined as λ_i/θ_i. The composite topochemical index (T) for the molecular graph is defined as:

$$T = \sum_{i<j} E_{ij} = \sum_{i<j} \left(V_i V_j\right)^{0.5} \tag{4}$$

In Eq. (4), E_{ij} stands for VEM edge weight of the edge formed by i^{th} and j^{th} vertices. VEM edge weight of an edge incident on a heteroatom is assigned a negative value. Apart from defining

functionality and branching contributions, TAU scheme uses vertex count (N_V) as the bulk parameter and N_P, N_Y and N_X (number of primary, tertiary and quaternary carbons) as the shape parameters. However, lacunae are found in some aspects of the TAU formalism. It is observed that the λ values of different atoms increase disproportionately with increase in atomic number (Table 1). Again, VEM vertex count of a vertex in a sigma bond with another atom of similar electronegativity should have different value than that bonded to an atom of different electronegativity. Moreover, differentiation must be made among double bonds of different types (ordinary alkenes or alkynes, conjugated system, aromatic system, *etc*.) Further, functionality contribution should be found out at the atomic level or fragmental basis which may be used for determination of pharmacophore fragment. To implement all these, we are redefining the basic parameters of the TAU scheme and introducing the novel ETA formalism.

Table 1. Uncorrected van der Waals volume (V_w), λ and α values of common atoms in organic compounds

Atom	V_w (10^2 Å^2) [a]	λ	α
H	0.056	0.00	0.00
C	0.206	0.50	0.50
N	0.141	0.40	0.40
O	0.115	0.33	0.33
F	0.115	0.29	0.29
S	0.244	1.67	0.835
Cl (ali)	0.206	1.43	0.715
Cl (aro)	0.244	1.43	0.715
Br (ali)	0.244	4.00	1.333
Br (aro)	0.287	4.00	1.333
I (ali)	0.335	6.57	1.6425
I (aro)	0.388	6.57	1.6425

[a] Ref. [34]

We define core count of a non–hydrogen vertex [α] as:

$$\alpha = \frac{z - z^v}{z^v} . \frac{1}{PN - 1} \tag{5}$$

In Eq. (5), *PN* stands for period number. Hydrogen atom being considered as reference, α for hydrogen is taken to be zero. Table 1 shows that α values of different atoms (which are commonly found in organic compounds) have high correlation (r = 0.946) with (uncorrected) van der Waals volume [34] and the relation is statistically better than that shown by λ (r = 0.898). Thus, ∑α 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, in periodic table, electronegativity increases in a period as one goes to the right hand side (*i.e*., as number of valence electron increases) and it decreases as one goes to higher period (*i.e*., as size increases). Here, we define a term ε as a measure of electronegativity in the following manner:

$$\varepsilon = -\alpha + 0.3Z^V \tag{6}$$

Table 2 shows that ε has good correlation (r = 0.937) with Pauling's electronegativity scale (EN).

Table 2. Pauling electronegativity (EN) and ε values of common (neutral) atoms in organic compounds

Atom	EN	ε	PN
H	2.1	0.3	1
C	2.5	0.7	2
N	3.0	1.1	2
O	3.5	1.47	2
F	4.0	1.81	2
P	2.1	0.5	3
S	2.5	0.965	3
Cl	3.0	1.385	3
Br	2.8	0.767	4
I	2.5	0.4575	5

Again, instead of N_P, N_Y and N_X used in the TAU scheme, $(\Sigma\alpha)_p/\Sigma\alpha$, $(\Sigma\alpha)_Y/\Sigma\alpha$ and $(\Sigma\alpha)_X/\Sigma\alpha$ can be used as the shape parameters. $(\Sigma\alpha)_p$, $(\Sigma\alpha)_Y$ and $(\Sigma\alpha)_X$ stand for summation of α values of the vertices that are joined to one, three and four other non-hydrogen vertices respectively in the molecular graph.

For calculation of VEM count β, contribution of a sigma bond (x) between two atoms of similar electronegativity ($\Delta\varepsilon \leq 0.3$) is considered to be 0.5 and for sigma bond between two atoms of different electronegativity ($\Delta\varepsilon > 0.3$) 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 π bonds, contributions (y) are considered depending on the type of the double bond: (*a*) for π bond between two atoms of similar electronegativity ($\Delta\varepsilon \leq 0.3$), y is taken to be 1; (*b*) 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; (*c*) for aromatic π system, y is taken as 2. In the TAU scheme, contribution of all kinds of π bonds was 2. Thus β of the ETA scheme is defined as:

$$\beta = \sum x\nu + \sum y\pi + \delta \tag{7}$$

In the above equation, δ is a correction factor of value 0.5 per atom with loan pair of electrons capable of resonance with aromatic ring (e.g., nitrogen of aniline, oxygen of phenol, *etc*.).

β can be split into two parts, β_s (sigma contribution to VEM count) and β_{ns} (non–sigma contribution to VEM count) which may be defined as below:

$$\beta_s = \sum x\nu \tag{8}$$

$$\beta_{ns} = \sum y\pi + \delta \tag{9}$$

For a given part (substructure) of a molecular graph, $\Sigma\beta_s$ and $\Sigma\beta_{ns}$ may be calculated considering all bonds (sigma bonds for the former and π bonds and lone pair of electrons for the latter) in the substructure. $\Sigma\beta'_s$ (defined as $[\Sigma\beta_s]/N_V$) may be taken as a relative measure of number of

electronegative atoms in the substructure while $\Sigma\beta'_{ns}$ (defined as $[\Sigma\beta_{ns}]/N_V$) may be taken as a relative measure of electron–richness (unsaturation) of the substructure.

VEM vertex count γ_i of the i^{th} vertex in a molecular graph is defined as:

$$\gamma_i = \frac{\alpha_i}{\beta_i} \tag{10}$$

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

Finally, we define the composite index η in the following manner:

$$\eta = \sum_{i<j}\left[\frac{\gamma_i\gamma_j}{r_{ij}^{\,2}}\right]^{0.5} \tag{11}$$

In Eq. (11), both bonded and non–bonded interactions have been considered. r_{ij} stands for the topological distance between i^{th} atom and 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 single bond, corresponding molecular graph may be considered as the reference alkane and the corresponding composite index value is designated as η_R. Considering functionality as the presence of heteroatoms (atoms other than carbon or hydrogen) and multiple bonds, functionality index η_F may be calculated as $\eta_R - \eta$. To avoid dependence of functionality on vertex count or bulk, we define another term η'_F as η_F/N_V. Again, we can determine contribution of a particular position or vertex (within the common substructure in the congeneric series) to functionality in the following manner:

$$[\eta]_i = \sum_{j\neq i}\left[\frac{\gamma_i\gamma_j}{r_{ij}^{\,2}}\right]^{0.5} \tag{12}$$

In Eq. (12), $[\eta]_i$ stands for contribution of the i^{th} vertex to η. Similarly, contribution of the i^{th} vertex $[\eta_R]_i$ to η_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 $[\eta'_F]_i$ is defined as $[\eta_F]_i/N_V$. Again, when only bonded interactions are considered (r_{ij} = 1), the corresponding composite index may be written as η^{local}.

$$\eta^{local} = \sum_{i<j, r_{ij}=1}(\gamma_i\gamma_j)^{0.5} \tag{13}$$

In the similar way, η_R^{local} for the corresponding reference alkane may also be calculated. η_R^{local} value is similar to T_R of reference alkane in the TAU scheme. Local functionality contribution (without considering global topology), η_F^{local}, may be calculated as $\eta_R^{local} - \eta^{local}$.

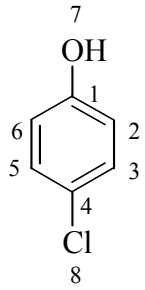
For calculation of branching, consideration of the local topology is sufficient. Branching is

calculated with respect to η value of the corresponding normal alkane (straight chain compound of same vertex count obtained from the reference alkane), η_N^{local}, which may be conveniently calculated (for compounds with $N_V \geq 3$) as:

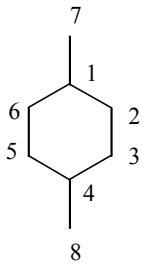
$$\eta_N^{local} = 1.414 + (N_V - 3)0.5 \quad (14)$$

The branching index η_B can be calculated as $\eta_N^{local} - \eta_R^{local} + 0.086N_R$, 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 branching contribution in comparison to the molecular size, another term η'_B is defined as η_B/N_V. Calculation of different indices is illustrated taking example of 4–chlorophenol in Table 3.

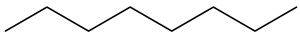
Table 3. Calculations of ETA parameters: example of 4–cholorophenol



4-Chlorophenol



Reference alkane



Normal alkane

	4–Chlorophenol								Reference alkane							
Vertex	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
α_i	0.5	0.5	0.5	0.5	0.5	0.5	0.33	0.72	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
$[\beta_s]_i$	1.75	1	1	1.75	1	1	0.75	0.75	1.5	1	1	1.5	1	1	0.5	0.5
$[\beta_{ns}]_i$	2	2	2	2	2	2	0.5	0.5	0	0	0	0	0	0	0	0
β_i	3.75	3	3	3.75	3	3	1.25	1.25	1.5	1	1	1.5	1	1	0.5	0.5
γ_i	0.13	0.17	0.17	0.13	0.17	0.17	0.26	0.57	0.33	0.5	0.5	0.33	0.5	0.5	1	1
$[\eta]_i$	0.75	0.74	0.75	0.82	0.75	0.74	0.66	0.94	–	–	–	–	–	–	–	–
$[\eta_R]_i$	–	–	–	–	–	–	–	–	2.06	2.12	2.12	2.06	2.12	2.12	2.10	2.10
$[\eta'_F]_i$	0.16	0.17	0.17	0.16	0.17	0.17	0.18	0.15	–	–	–	–	–	–	–	–
η	3.074								–							
η_R	–								8.392							
η'_F	0.665								–							
η^{local}	1.395								–							
η_R^{local}	–								3.786							
η_F^{local}	2.392								–							
η_N^{local}	–								3.914							
η'_B	0.027								–							
$\sum\alpha$	4.045								–							
$[\sum\alpha]_p$	1.045								–							

In the present communication, utility of the ETA parameters has been demonstrated through a QSAR study taking toxicity (*pC*) against *Tetrahymena pyriformis* for a set of 50 substituted phenols (taken from Ref. [35]) as model data set (Table 4). The common atoms of the molecules are numbered 1 through 7 (as shown in Table 3).

Table 4. Observed, calculated and predicted toxicity of substituted phenols against *Tetrahymena pyriformis*

No	Compound name	Toxicity against *Tetrahymena pyriformis*								
		Obs [a]	Calc [b]	Res [b]	Pred [b]	Pres [b]	Calc [c]	Res [c]	Pred [c]	Pres [c]
1	Phenol	–0.431	–0.464	0.033	–0.470	0.039	–0.430	–0.001	–0.430	–0.001
2	2,6–Difluorophenol	0.396	0.350	0.046	0.338	0.058	0.383	0.013	0.379	0.017
3	2–Fluorophenol	0.248	–0.052	0.300	–0.092	0.340	–0.019	0.267	–0.062	0.310
4	4–Fluorophenol	0.017	0.139	–0.122	0.164	–0.147	0.094	–0.077	0.110	–0.934
5	3–Fluorophenol	0.473	0.278	0.195	0.236	0.237	0.299	0.174	0.262	0.211
6	4–Methylphenol	–0.192	–0.092	–0.100	–0.084	–0.108	–0.075	–0.117	–0.066	–0.126
7	3–Methylphenol	–0.062	–0.092	0.030	–0.094	0.032	–0.075	0.013	–0.076	0.014
8	2–Chlorophenol	0.277	0.252	0.025	0.249	0.028	0.271	0.006	0.271	0.006
9	2–Bromophenol	0.504	0.463	0.041	0.461	0.043	0.456	0.048	0.453	0.051
10	4–Chlorophenol	0.545	0.442	0.103	0.430	0.115	0.384	0.161	0.363	0.182
11	3–Ethylphenol	0.229	0.251	–0.022	0.252	–0.023	0.252	–0.023	0.253	–0.024
12	2–Ethylphenol	0.176	0.251	–0.075	0.254	–0.078	0.252	–0.076	0.256	–0.080
13	4–Bromophenol	0.681	0.654	0.027	0.650	0.031	0.770	–0.089	0.778	–0.097
14	2,3–Dimethylphenol	0.122	0.251	–0.129	0.256	–0.134	0.252	–0.130	0.258	–0.136
15	2,4–Dimethylphenol	0.128	0.251	–0.123	0.256	–0.128	0.252	–0.124	0.258	–0.130
16	2,5–Dimethylphenol	0.009	0.251	–0.242	0.262	–0.253	0.252	–0.243	0.263	–0.254
17	3,4–Dimethylphenol	0.122	0.251	–0.129	0.256	–0.134	0.252	–0.130	0.258	–0.136
18	3,5–Dimethylphenol	0.113	0.251	–0.138	0.257	–0.144	0.252	–0.139	0.259	–0.146
19	3–Chloro–4–fluorophenol	0.842	1.160	–0.318	1.224	–0.382	1.090	–0.248	1.145	–0.303
20	2–Chloro–5–methylphenol	0.640	0.582	0.058	0.578	0.062	0.587	0.053	0.583	0.057
21	4–Iodophenol	0.854	0.840	0.014	0.838	0.016	0.948	–0.094	0.956	–0.102
22	3–Iodophenol	1.118	0.979	0.139	0.946	0.172	1.154	–0.036	1.159	–0.041
23	2–Isopropylphenol	0.803	0.565	0.238	0.553	0.250	0.553	0.250	0.541	0.262
24	3–Isopropylphenol	0.609	0.565	0.044	0.563	0.046	0.553	0.056	0.550	0.059
25	4–Isopropylphenol	0.473	0.565	–0.092	0.569	–0.096	0.553	–0.080	0.557	–0.084
26	2,5–Dichlorophenol	1.128	1.238	–0.110	1.255	–0.127	1.235	–0.107	1.252	–0.124
27	2,3–Dichlorophenol	1.271	1.238	0.033	1.232	0.038	1.235	0.036	1.229	0.042
28	4–Chloro–2–methylphenol	0.700	0.773	–0.073	0.778	–0.078	0.700	0.000	0.700	0.000
29	4–Chloro–3–methylphenol	0.795	0.773	0.022	0.771	0.024	0.700	0.095	0.691	0.104
30	2,4–Dichlorophenol	1.036	1.099	–0.063	1.106	–0.070	1.029	0.007	1.029	0.007
31	3–*tert*–Butylphenol	0.730	0.850	–0.120	0.858	–0.128	0.827	–0.097	0.833	–0.103
32	4–*tert*–Butylphenol	0.913	0.850	0.063	0.846	0.067	0.827	0.086	0.821	0.092
33	3,5–Dichlorophenol	1.562	1.568	–0.006	1.573	–0.106	1.553	0.009	1.546	0.016
34	2–Phenylphenol	1.094	1.334	–0.240	1.361	–0.267	1.292	–0.198	1.315	–0.221
35	2,4–Dibromophenol	1.403	1.376	0.027	1.373	0.030	1.463	–0.060	1.469	–0.066
36	2,3,6–Trimethylphenol	0.418	0.565	–0.147	0.572	–0.154	0.553	–0.135	0.560	–0.142
37	3,4,5–Trimethylphenol	0.930	0.565	0.365	0.547	0.383	0.553	0.377	0.534	0.396
38	2,4,6–Trimethylphenol	1.695	1.696	–0.001	1.697	–0.002	1.620	0.075	1.597	0.098
39	4–Chloro–3,5–dimethylphenol	1.203	1.074	0.129	1.066	0.137	0.989	0.214	0.972	0.231
40	4–Bromo–2,6–dichlorophenol	1.779	1.806	–0.027	1.810	–0.031	1.911	–0.132	1.954	–0.175
41	2,4,5–Trichlorophenol	2.100	2.026	0.074	2.008	0.092	1.938	0.162	1.904	0.196
42	4–Bromo–6–chloro–2–methylphenol	1.277	1.513	–0.236	1.532	–0.255	1.612	–0.335	1.649	–0.372
43	4–Bromo–2,6–dimethylphenol	1.278	1.215	0.063	1.207	0.071	1.308	–0.030	1.311	–0.033
44	2,4,6–Tribromophenol	2.050	1.894	0.156	1.870	0.180	1.964	0.086	1.953	0.097
45	2–*tert*–Butyl–4–methylphenol	1.297	1.107	0.190	1.089	0.208	1.073	0.224	1.052	0.245
46	4–Chloro–2–isopropyl–5–methylphenol	1.862	1.591	0.271	1.568	0.294	1.485	0.377	1.449	0.413
47	6–*tert*–Butyl–2,4–dimethylphenol	1.245	1.334	–0.089	1.344	–0.099	1.292	–0.047	1.298	–0.053
48	2,6–Diphenylphenol	2.113	2.097	0.016	2.079	0.034	2.040	0.073	1.962	0.151
49	2,4–Dibromo–6–phenylphenol	2.207	2.254	–0.047	2.285	–0.078	2.319	–0.112	2.390	–0.183
50	2,6–Di–*tert*–butyl–4–methylphenol	1.788	1.845	–0.057	1.856	–0.068	1.788	0.000	1.788	0.000

Obs = Observed, Calc = Calculated, Pred = Predicted, Res = Residual = Obs – Calc, Pres = Predicted residual = Obs – Pred

[a] Ref. [35], [b] according to Eq. (27), [c] according to Eq. (28)

Table 5. Definitions of different ETA parameters used in exploring QSAR of toxicity of substituted phenols

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 non–hydrogen vertex of the molecule
N_e	Number of atoms in the substituent positions of the phenol nucleus which have ε values larger than 1
N_V	Vertex count (excluding hydrogen)
N	Total number of atoms (including hydrogen)
$[\sum\varepsilon/N]_{sub}$	Sum of ε/N values of all atoms (including hydrogen) in substituent positions of phenol nucleus
$[N_e]_o$	Number of atoms in the *ortho* substituent positions of the phenol nucleus which have ε values larger than 1
$[N_e]_m$	Number of atoms in the *meta* substituent positions of the phenol nucleus which have ε values larger than 1
$[N_e]_p$	Number of atoms in the *para* substituent position of the phenol nucleus which have ε values larger than 1
$[\sum\beta'_{ns}]_o$	Sum of $\sum\beta'_{ns}$ values of two *ortho* positions; $\sum\beta'_{ns}$ is defined as $[\sum\beta_{ns}]/N_v$ for non–hydrogen substituent; in case, hydrogen is present in the substituent position, the value for that position is taken as zero
$[\sum\beta'_{ns}]_m$	Sum of $\sum\beta'_{ns}$ values of two *meta* positions; $\sum\beta'_{ns}$ is defined as $[\sum\beta_{ns}]/N_v$ for non–hydrogen substituent; in case, hydrogen is present in the substituent position, the value for that position is taken as zero
$[\sum\beta'_{ns}]_p$	Sum of $\sum\beta'_{ns}$ values of *para* position; $\sum\beta'_{ns}$ is defined as $[\sum\beta_{ns}]/N_v$ for non–hydrogen substituent; in case, hydrogen is present in the substituent position, the value for that position is taken as zero
$[\sum\beta'_s]_o$	Sum of $\sum\beta'_s$ values of two *ortho* positions; $\sum\beta'_s$ is defined as $[\sum\beta_s]/N_v$ for non–hydrogen substituent; in case, hydrogen is present in the substituent position, the value for that position is taken as zero
$[\sum\beta'_s]_m$	Sum of $\sum\beta'_s$ values of two *meta* positions; $\sum\beta'_s$ is defined as $[\sum\beta_s]/N_v$ for non–hydrogen substituent; in case, hydrogen is present in the substituent position, the value for that position is taken as zero
$[\sum\beta'_s]_p$	Sum of $\sum\beta'_s$ values of *para* position; $\sum\beta'_s$ is defined as $[\sum\beta_s]/N_v$ for non–hydrogen substituent; in case, hydrogen is present in the substituent position, the value for that position is taken as zero
η'_B	$= \eta'_B/N_V$
η'_F	$= \eta'_F/N_V$
$[\eta'_F]_i$	Contribution of the position *i* to η'_F value

Different ETA descriptors calculated for the phenol derivatives are defined in Table 5. Factor analysis has been performed as the data preprocessing step for identification of important descriptors for the subsequent multiple regression analysis [36,37]. For this purpose, the data matrix consisting of the descriptors has been divided into two parts, and each part, along with toxicity values, has been subjected to principal component factor analysis using STATISTICA software [38]. 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 ultimate goal of interpretation and / or prediction. The factors were extracted by principal component method and then rotated by VARIMAX rotation to obtain Thurston's simple structure. Only factors describing ≥ 5% of the total variance were considered. The analyses were carried out based on the following postulates: (*a*) only variables with non–zero loadings in such factors where biological activity also has non–zero loading are important in explaining variance of the activity; (*b*) only variables with non–zero loadings in different factors may be combined in regression equations; (*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 η, η_R, η_F, η_B and contributions of different vertices to η_F were done, using distance matrix and VEM vertex counts as inputs, by the GW–BASIC programs *KRETA*1 and *KRETA*2 developed by one of the authors [39]. The regression analyses were carried out using a

program *RRR*98 [39].

The statistical quality of the equations [40] 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), average of absolute values of the residuals (*AVRES*), variance ratio (F) at specified degrees of freedom (*df*), 95% confidence intervals of the regression coefficients, leave–one–out cross–validation R^2 (Q^2) [41], predicted residual sum of squares (*PRESS*) [41], standard deviation based on *PRESS* (S_{PRESS}) [6], standard deviation of error of prediction (*SDEP*) [6] and average absolute predicted residual ($Pres_{av}$). PRESS (leave–one–out) statistics [6,41] were calculated using the programs *KRPRES*1 and *KRPRES*2 [39]. 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. Finally, leave–many–out cross–validation was applied on the final equations.

3 RESULTS AND DISCUSSION

The calculated values of different ETA descriptors as defined in Table 5 are given in Tables 6–8. The results of the principal component factor analyses are given in Tables 9 and 10.

Table 9 shows that the data matrix [A] composed of 18 variables (including response variable *pC*) can be explained to the extent of 96.6% by 7 factors. Biological activity (*pC*) is highly loaded with factor 3 which is in turn loaded in the parameters $\sum\alpha$ and $[\sum\alpha]^2$. Other factors with which biological activity is moderately loaded are factor 6 (highly loaded in $[\sum\alpha]_P$ and $[\sum\alpha]_P/\sum\alpha$), factor 2 (highly loaded in $[N_e]_m$ and $[\sum\beta'_{ns}]_m$), factor 4 (highly loaded in $[N_e]_p$, $[\sum\beta'_s]_p$ and $[\sum\beta'_{ns}]_p$), factor 1 (highly loaded in η_B and η'_B) and factor 5 (highly loaded in N_e, $[N_e]_o$, $[\sum\beta'_s]_o$ and $[\sum\varepsilon/N]_{sub}$).

The size parameter $\sum\alpha$ can singularly explain 67.6% and predict 64.5% of the variance of toxicity of phenols. The relation is further improved when parabolic relation of $\sum\alpha$ is used.

$$pC = 0.440(\pm 0.087)\sum\alpha - 1.374(\pm 0.454)$$
$$n = 50, q^2 = 0.645, r_a^2 = 0.676, r^2 = 0.682, r = 0.826, s = 0.379, F = 103.0(df\,1, 48), \quad (15)$$
$$AVRES = 0.301, PRESS = 7.691, SDEP = 0.392, S_{PRESS} = 0.400, \Pr es_{av} = 0.317$$

$$pC = 1.398(\pm 0.503)\sum\alpha - 0.079(\pm 0.041)[\sum\alpha]^2 - 4.082(\pm 1.461)$$
$$n = 50, Q^2 = 0.737, R_a^2 = 0.749, R^2 = 0.759, R = 0.871, s = 0.333, F = 74.1(df\,2, 47), \quad (16)$$
$$AVRES = 0.263, PRESS = 5.703, SDEP = 0.338, S_{PRESS} = 0.348, \Pr es_{av} = 0.276$$

The 95% confidence intervals of the regression coefficients are shown within parentheses. Eq. (16) suggests that toxicity of phenols is highly dependent on bulk: toxicity increases as molecular size increases, however it decreases after some critical value of size.

Table 6. Different ETA descriptors for substituted phenols – part I

No	Compound name	Descriptors							
		$\sum\alpha$	$[\sum\alpha]_p$	$[\sum\beta'_s]_o$	$[\sum\beta'_s]_m$	$[\sum\beta'_s]_p$	$[\sum\beta'_{ns}]_o$	$[\sum\beta'_{ns}]_m$	$[\sum\beta'_{ns}]_p$
1	Phenol	3.3300	0.3300	0.00	0.00	0.00	0.00	0.00	0.00
2	2,6–Difluorophenol	3.9100	0.9100	1.50	0.00	0.00	1.00	0.00	0.00
3	2–Fluorophenol	3.6200	0.6200	0.75	0.00	0.00	0.50	0.00	0.00
4	4–Fluorophenol	3.6200	0.6200	0.00	0.00	0.75	0.00	0.00	0.50
5	3–Fluorophenol	3.6200	0.6200	0.00	0.75	0.00	0.00	0.50	0.00
6	4–Methylphenol	3.8300	0.8300	0.00	0.00	0.50	0.00	0.00	0.00
7	3–Methylphenol	3.8300	0.8300	0.00	0.50	0.00	0.00	0.00	0.00
8	2–Chlorophenol	4.0450	1.0450	0.75	0.00	0.00	0.50	0.00	0.00
9	2–Bromophenol	4.6630	1.6630	0.50	0.00	0.00	0.50	0.00	0.00
10	4–Chlorophenol	4.0450	1.0450	0.00	0.00	0.75	0.00	0.00	0.50
11	3–Ethylphenol	4.3300	0.8300	0.00	0.50	0.00	0.00	0.00	0.00
12	2–Ethylphenol	4.3300	0.8300	0.50	0.00	0.00	0.00	0.00	0.00
13	4–Bromophenol	4.6630	1.6630	0.00	0.00	0.50	0.00	0.00	0.50
14	2,3–Dimethylphenol	4.3300	1.3300	0.50	0.50	0.00	0.00	0.00	0.00
15	2,4–Dimethylphenol	4.3300	1.3300	0.50	0.00	0.50	0.00	0.00	0.00
16	2,5–Dimethylphenol	4.3300	1.3300	0.50	0.50	0.00	0.00	0.00	0.00
17	3,4–Dimethylphenol	4.3300	1.3300	0.00	0.50	0.50	0.00	0.00	0.00
18	3,5–Dimethylphenol	4.3300	1.3300	0.00	1.00	0.00	0.00	0.00	0.00
19	3–Chloro–4–fluorophenol	4.3350	1.3350	0.00	0.75	0.75	0.00	0.50	0.50
20	2–Chloro–5–methylphenol	4.5450	1.5450	0.75	0.50	0.00	0.50	0.00	0.00
21	4–Iodophenol	4.9725	1.9725	0.00	0.00	0.50	0.00	0.00	0.50
22	3–Iodophenol	4.9725	1.9725	0.00	0.50	0.00	0.00	0.50	0.00
23	2–*Iso*propylphenol	4.8300	1.3300	0.50	0.00	0.00	0.00	0.00	0.00
24	3–*Iso*propylphenol	4.8300	1.3300	0.00	0.50	0.00	0.00	0.00	0.00
25	4–*Iso*propylphenol	4.8300	1.3300	0.00	0.00	0.50	0.00	0.00	0.00
26	2,5–Dichlorophenol	4.7600	1.7600	0.75	0.75	0.00	0.50	0.50	0.00
27	2,3–Dichlorophenol	4.7600	1.7600	0.75	0.75	0.00	0.50	0.50	0.00
28	4–Chloro–2–methylphenol	4.5450	1.5450	0.50	0.00	0.75	0.00	0.00	0.50
29	4–Chloro–3–methylphenol	4.5450	1.5450	0.00	0.50	0.75	0.00	0.00	0.50
30	2,4–Dichlorophenol	4.7600	1.7600	0.75	0.00	0.75	0.50	0.00	0.50
31	3–*tert*–Butylphenol	5.3300	1.8300	0.00	0.50	0.00	0.00	0.00	0.00
32	4–*tert*–Butylphenol	5.3300	1.8300	0.00	0.00	0.50	0.00	0.00	0.00
33	3,5–Dichlorophenol	4.7600	1.7600	0.00	1.50	0.00	0.00	1.00	0.00
34	2–Phenylphenol	6.3300	0.3300	0.50	0.00	0.00	1.00	0.00	0.00
35	2,4–Dibromophenol	5.9960	2.9960	0.50	0.00	0.50	0.50	0.00	0.50
36	2,3,6–Trimethylphenol	4.8300	1.8300	1.00	0.50	0.00	0.00	0.00	0.00
37	3,4,5–Trimethylphenol	4.8300	1.8300	0.00	1.00	0.50	0.00	0.00	0.00
38	2,4,6–Trimethylphenol	5.4750	2.4750	1.50	0.00	0.75	1.00	0.00	0.50
39	4–Chloro–3,5–dimethylphenol	5.0450	2.0450	0.00	1.00	0.75	0.00	0.00	0.50
40	4–Bromo–2,6–dichlorophenol	6.0930	3.0930	1.50	0.00	0.50	1.00	0.00	0.50
41	2,4,5–Trichlorophenol	5.4750	2.4750	0.75	0.75	0.75	0.50	0.50	0.50
42	4–Bromo–6–chloro–2–methylphenol	5.8780	2.8780	1.25	0.00	0.50	0.50	0.00	0.50
43	4–Bromo–2,6–dimethylphenol	5.6630	2.6630	1.00	0.00	0.50	0.00	0.00	0.50
44	2,4,6–Tribromophenol	7.3290	4.3290	1.00	0.00	0.50	1.00	0.00	0.50
45	2–*tert*–Butyl–4–methylphenol	5.8300	2.3300	0.50	0.00	0.50	0.00	0.00	0.00
46	4–Chloro–2–*iso*propyl–5–methylphenol	6.0450	2.5450	0.50	0.50	0.75	0.00	0.00	0.50
47	6–*tert*–Butyl–2,4–dimethylphenol	6.3300	2.8300	1.00	0.00	0.50	0.00	0.00	0.00
48	2,6–Diphenylphenol	9.3300	0.3300	1.00	0.00	0.00	2.00	0.00	0.00
49	2,4–Dibromo–6–phenylphenol	8.9960	2.9960	1.00	0.00	0.50	1.50	0.00	0.50
50	2,6–Di–*tert*–butyl–4–methylphenol	7.8300	2.8300	1.00	0.00	0.50	0.00	0.00	0.00

Table 7. Different ETA descriptors for substituted phenols – part II

No	Compound name	Descriptors							
		N_e	$[\Sigma\varepsilon/N]_{sub}$	$[N_e]_o$	$[N_e]_m$	$[N_e]_p$	η	η_R	η'_B
1	Phenol	0	0.3000	0	0	0	2.2007	6.6258	0.0153
2	2,6–Difluorophenol	2	0.9040	2	0	0	3.3587	10.4897	0.0319
3	2–Fluorophenol	1	0.6020	1	0	0	2.7572	8.4719	0.0246
4	4–Fluorophenol	1	0.6020	0	0	1	2.7338	8.3923	0.0267
5	3–Fluorophenol	1	0.6020	0	1	0	2.7429	8.4234	0.0267
6	4–Methylphenol	0	0.3500	0	0	0	3.4090	8.3923	0.0267
7	3–Methylphenol	0	0.3500	0	0	0	3.4299	8.4234	0.0267
8	2–Chlorophenol	1	0.5170	1	0	0	2.6059	8.4719	0.0246
9	2–Bromophenol	0	0.3934	0	0	0	3.6662	8.4719	0.0246
10	4–Chlorophenol	1	0.5170	0	0	1	3.0744	8.3923	0.0267
11	3–Ethylphenol	0	0.3727	0	0	0	4.6030	9.9264	0.0195
12	2–Ethylphenol	0	0.3727	0	0	0	4.6482	9.9960	0.0176
13	4–Bromophenol	0	0.3934	0	0	0	3.6027	8.3923	0.0267
14	2,3–Dimethylphenol	0	0.3727	0	0	0	4.9966	10.4897	0.0319
15	2,4–Dimethylphenol	0	0.3727	0	0	0	4.9019	10.4101	0.0337
16	2,5–Dimethylphenol	0	0.3727	0	0	0	4.8779	10.4101	0.0337
17	3,4–Dimethylphenol	0	0.3727	0	0	0	4.9420	10.4101	0.0337
18	3,5–Dimethylphenol	0	0.3727	0	0	0	4.8893	10.3928	0.0356
19	3–Chloro–4–fluorophenol	2	0.8190	0	1	1	3.7190	10.4101	0.0337
20	2–Chloro–5–methylphenol	1	0.4856	1	0	0	4.4779	10.4101	0.0337
21	4–Iodophenol	0	0.3315	0	0	0	3.7619	8.3923	0.0267
22	3–Iodophenol	0	0.3315	0	0	0	3.7891	8.4234	0.0267
23	2–Isopropylphenol	0	0.3857	0	0	0	6.2863	12.1184	0.0287
24	3–Isopropylphenol	0	0.3857	0	0	0	6.2236	12.0177	0.0304
25	4–Isopropylphenol	0	0.3857	0	0	0	6.1820	11.9494	0.0304
26	2,5–Dichlorophenol	2	0.7340	1	1	0	4.1011	10.4101	0.0337
27	2,3–Dichlorophenol	2	0.7340	1	1	0	4.1621	10.4897	0.0319
28	4–Chloro–2–methylphenol	1	0.4856	0	0	1	4.5048	10.4101	0.0337
29	4–Chloro–3–methylphenol	1	0.4856	0	0	1	4.5238	10.4101	0.0337
30	2,4–Dichlorophenol	2	0.7340	1	0	1	4.1094	10.4101	0.0337
31	3–*tert*–Butylphenol	0	0.3941	0	0	0	8.3106	14.6211	0.0457
32	4–*tert*–Butylphenol	0	0.3941	0	0	0	8.2558	14.5291	0.0457
33	3,5–Dichlorophenol	2	0.7340	0	2	0	4.1004	10.3928	0.0356
34	2–Phenylphenol	0	0.4600	0	0	0	5.7317	16.5167	0.0162
35	2,4–Dibromophenol	0	0.4868	0	0	0	5.3784	10.4101	0.0337
36	2,3,6–Trimethylphenol	0	0.3857	0	0	0	6.6746	12.6481	0.0377
37	3,4,5–Trimethylphenol	0	0.3857	0	0	0	6.7051	12.5996	0.0394
38	2,4,6–Trimethylphenol	3	0.9510	2	0	1	5.2663	12.5996	0.0394
39	4–Chloro–3,5–dimethylphenol	1	0.4714	0	0	1	6.2033	12.5996	0.0394
40	4–Bromo–2,6–dichlorophenol	2	0.8274	2	0	0	5.9396	12.5996	0.0394
41	2,4,5–Trichlorophenol	3	0.9510	1	1	1	5.2564	12.5685	0.0394
42	4–Bromo–6–chloro–2–methylphenol	1	0.5440	1	0	0	6.4077	12.5996	0.0394
43	4–Bromo–2,6–dimethylphenol	0	0.4152	0	0	0	6.8929	12.5996	0.0394
44	2,4,6–Tribromophenol	0	0.5802	0	0	0	7.4643	12.5996	0.0394
45	2–*tert*–Butyl–4–methylphenol	0	0.4000	0	0	0	10.2882	17.0936	0.0494
46	4–Chloro–2–isopropyl–5–methylphenol	1	0.4579	0	0	1	9.4086	16.6730	0.0403
47	6–*tert*–Butyl–2,4–dimethylphenol	0	0.4043	0	0	0	12.4685	19.6823	0.0525
48	2,6–Diphenylphenol	0	0.4920	0	0	0	10.0956	28.6451	0.0166
49	2,4–Dibromo–6–phenylphenol	0	0.5223	0	0	0	9.8976	21.5126	0.0272
50	2,6–Di–*tert*–butyl–4–methylphenol	0	0.4125	0	0	0	19.2745	27.7274	0.0607

Table 8. Different ETA descriptors for substituted phenols – part III

No	Compound name	Descriptors							
		η'_F	$[\eta'_F]_1$	$[\eta'_F]_2$	$[\eta'_F]_3$	$[\eta'_F]_4$	$[\eta'_F]_5$	$[\eta'_F]_6$	$[\eta'_F]_7$
1	Phenol	0.6322	0.1789	0.1835	0.1769	0.1728	0.1769	0.1835	0.1918
2	2,6–Difluorophenol	0.7923	0.1706	0.1671	0.1692	0.1640	0.1692	0.1671	0.1929
3	2–Fluorophenol	0.7143	0.1742	0.1750	0.1775	0.1678	0.1677	0.1772	0.1924
4	4–Fluorophenol	0.7073	0.1666	0.1772	0.1775	0.1677	0.1775	0.1772	0.1832
5	3–Fluorophenol	0.7101	0.1696	0.1832	0.1705	0.1738	0.1715	0.1734	0.1868
6	4–Methylphenol	0.6229	0.1605	0.1680	0.1635	0.1400	0.1635	0.1680	0.1764
7	3–Methylphenol	0.6242	0.1614	0.1693	0.1428	0.1599	0.1623	0.1666	0.1782
8	2–Chlorophenol	0.7332	0.1769	0.1803	0.1804	0.1698	0.1692	0.1792	0.1949
9	2–Bromophenol	0.6007	0.1583	0.1398	0.1596	0.1560	0.1589	0.1654	0.1776
10	4–Chlorophenol	0.6647	0.1635	0.1725	0.1704	0.1552	0.1704	0.1725	0.1797
11	3–Ethylphenol	0.5915	0.1470	0.1567	0.1315	0.1483	0.1493	0.1523	0.1653
12	2–Ethylphenol	0.5942	0.1482	0.1353	0.1516	0.1461	0.1472	0.1544	0.1690
13	4–Bromophenol	0.5987	0.1587	0.1654	0.1596	0.1327	0.1596	0.1654	0.1744
14	2,3–Dimethylphenol	0.6103	0.1481	0.1347	0.1309	0.1488	0.1496	0.1547	0.1701
15	2,4–Dimethylphenol	0.6120	0.1473	0.1346	0.1531	0.1283	0.1507	0.1560	0.1684
16	2,5–Dimethylphenol	0.6147	0.1481	0.1339	0.1520	0.1488	0.1301	0.1571	0.1701
17	3,4–Dimethylphenol	0.6076	0.1470	0.1571	0.1309	0.1285	0.1520	0.1547	0.1660
18	3,5–Dimethylphenol	0.6115	0.1477	0.1558	0.1308	0.1499	0.1308	0.1558	0.1677
19	3–Chloro–4–fluorophenol	0.7435	0.1560	0.1714	0.1562	0.1592	0.1684	0.1658	0.1759
20	2–Chloro–5–methylphenol	0.6591	0.1536	0.1479	0.1581	0.1528	0.1328	0.1611	0.1751
21	4–Iodophenol	0.5788	0.1573	0.1632	0.1564	0.1267	0.1564	0.1632	0.1728
22	3–Iodophenol	0.5793	0.1571	0.1621	0.1295	0.1527	0.1575	0.1630	0.1737
23	2–Isopropylphenol	0.5832	0.1391	0.1292	0.1444	0.1376	0.1375	0.1451	0.1622
24	3–Isopropylphenol	0.5794	0.1367	0.1490	0.1257	0.1415	0.1406	0.1420	0.1569
25	4–Isopropylphenol	0.5767	0.1346	0.1451	0.1444	0.1235	0.1444	0.1451	0.1532
26	2,5–Dichlorophenol	0.7010	0.1571	0.1505	0.1621	0.1589	0.1466	0.1672	0.1788
27	2,3–Dichlorophenol	0.7031	0.1571	0.1545	0.1506	0.1589	0.1566	0.1616	0.1788
28	4–Chloro–2–methylphenol	0.6561	0.1500	0.1383	0.1592	0.1422	0.1568	0.1600	0.1714
29	4–Chloro–3–methylphenol	0.6540	0.1496	0.1611	0.1366	0.1426	0.1581	0.1587	0.1690
30	2,4–Dichlorophenol	0.7001	0.1555	0.1522	0.1654	0.1458	0.1598	0.1640	0.1764
31	3–*tert*–Butylphenol	0.5737	0.1286	0.1435	0.1219	0.1366	0.1339	0.1340	0.1506
32	4–*tert*–Butylphenol	0.5703	0.1259	0.1380	0.1393	0.1199	0.1393	0.1380	0.1459
33	3,5–Dichlorophenol	0.6992	0.1549	0.1649	0.1483	0.1621	0.1483	0.1649	0.1751
34	2–Phenylphenol	0.8296	0.1261	0.1272	0.1339	0.1238	0.1206	0.1296	0.1495
35	2,4–Dibromophenol	0.5591	0.1426	0.1259	0.1461	0.1196	0.1455	0.1513	0.1637
36	2,3,6–Trimethylphenol	0.5973	0.1375	0.1247	0.1207	0.1399	0.1416	0.1240	0.1635
37	3,4,5–Trimethylphenol	0.5895	0.1361	0.1462	0.1213	0.1192	0.1213	0.1462	0.1578
38	2,4,6–Trimethylphenol	0.7333	0.1490	0.1441	0.1563	0.1383	0.1563	0.1441	0.1738
39	4–Chloro–3,5–dimethylphenol	0.6396	0.1385	0.1498	0.1264	0.1326	0.1264	0.1598	0.1605
40	4–Bromo–2,6–dichlorophenol	0.6660	0.1453	0.1390	0.1476	0.1197	0.1476	0.1390	0.1695
41	2,4,5–Trichlorophenol	0.7312	0.1470	0.1426	0.1584	0.1403	0.1411	0.1600	0.1705
42	4–Bromo–6–chloro–2–methylphenol	0.6192	0.1403	0.1260	0.1421	0.1164	0.1450	0.1358	0.1650
43	4–Bromo–2,6–dimethylphenol	0.5707	0.1354	0.1226	0.1395	0.1131	0.1395	0.1226	0.1605
44	2,4,6–Tribromophenol	0.5135	0.1298	0.1148	0.1347	0.1092	0.1347	0.1148	0.1552
45	2–*tert*–Butyl–4–methylphenol	0.5671	0.1238	0.1175	0.1335	0.1059	0.1249	0.1315	0.1500
46	4–Chloro–2–isopropyl–5–methylphenol	0.6054	0.1238	0.1157	0.1357	0.1165	0.1097	0.1347	0.1500
47	6–*tert*–Butyl–2,4–dimethylphenol	0.5549	0.1175	0.1048	0.1206	0.1004	0.1269	0.1111	0.1465
48	2,6–Diphenylphenol	0.9763	0.1067	0.1029	0.1089	0.1058	0.1089	0.1029	0.1339
49	2,4–Dibromo–6–phenylphenol	0.7743	0.1114	0.0957	0.1096	0.0919	0.1208	0.1123	0.1383
50	2,6–Di–*tert*–butyl–4–methylphenol	0.5283	0.1054	0.0975	0.1120	0.0888	0.1120	0.0975	0.1367

Table 9. Factor loadings of the variables (data matrix [A]) after VARIMAX rotation

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Communality
*p*C	0.219	–0.248	**0.780**	0.237	0.197	0.365	–0.029	0.947
$\Sigma\alpha$	0.215	0.082	**0.952**	0.009	–0.041	0.173	0.065	0.995
$[\Sigma\alpha]^2$	0.163	0.087	**0.971**	–0.005	–0.047	0.079	0.076	0.991
$[\Sigma\ \alpha\ ]_p$	0.334	0.004	0.365	0.150	0.088	**0.840**	0.034	0.982
$[\Sigma\ \alpha\ ]_p/\Sigma\alpha$	0.388	–0.051	0.089	0.175	0.123	**0.881**	–0.041	0.985
N_e	0.025	–0.463	–0.081	0.415	**0.764**	–0.002	–0.113	0.991
$[N_e]_o$	–0.008	0.015	–0.043	–0.008	**0.972**	0.096	0.019	0.957
$[N_e]_m$	0.008	**–0.972**	–0.027	–0.012	0.124	–0.012	–0.108	0.973
$[N_e]_p$	0.059	–0.055	–0.089	**0.927**	0.158	–0.129	–0.166	0.942
$[\Sigma\beta'_{ns}]_o$	–0.286	0.091	**0.751**	–0.067	0.540	–0.084	0.053	0.960
$[\Sigma\beta'_{ns}]_m$	–0.041	**–0.971**	–0.027	–0.053	0.071	0.042	–0.123	0.970
$[\Sigma\beta'_{ns}]_p$	–0.179	0.054	0.176	**0.823**	0.120	0.429	0.084	0.948
$[\Sigma\varepsilon/N]_{sub}$	0.005	–0.444	0.120	0.328	**0.801**	0.024	–0.008	0.962
$[\Sigma\beta'_s]_o$	0.170	0.188	0.425	–0.135	**0.731**	0.222	0.223	0.895
$[\Sigma\beta'_s]_m$	0.075	–0.564	–0.168	–0.061	–0.104	0.029	**–0.786**	0.985
$[\Sigma\beta'_s]_p$	0.258	0.088	0.061	**0.868**	–0.036	0.245	0.184	0.926
η_B	**0.900**	0.045	0.360	0.032	–0.017	0.216	0.015	0.990
η'_B	**0.881**	–0.021	0.047	0.106	0.066	0.427	–0.071	0.982
%Variance	0.122	0.153	0.198	0.151	0.174	0.123	0.044	0.966

Table 10. Factor loadings of the variables (data matrix [B]) after VARIMAX rotation

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communality
*p*C	0.389	–0.449	**–0.745**	–0.166	–0.033	0.937
$[\Sigma\ \alpha\ ]_p$	0.316	0.208	**–0.892**	–0.191	0.062	0.979
$[\Sigma\ \alpha\ ]_p/\Sigma\alpha$	0.122	0.362	**–0.871**	–0.271	–0.045	0.980
η	**0.785**	–0.053	–0.167	–0.557	0.119	0.971
η_R	**0.770**	–0.458	–0.085	–0.409	0.095	0.986
η_F	0.498	**–0.858**	0.054	–0.075	0.031	0.993
η'_F	–0.221	**–0.949**	0.149	0.086	–0.027	0.979
$[\eta'_F]_1$	**–0.934**	0.135	0.197	0.227	0.016	0.982
$[\eta'_F]_2$	**–0.847**	0.112	0.337	0.133	–0.205	0.904
$[\eta'_F]_3$	**–0.880**	0.009	0.096	0.097	0.385	0.941
$[\eta'_F]_4$	**–0.838**	0.023	0.375	0.153	–0.222	0.916
$[\eta'_F]_5$	**–0.857**	0.144	0.175	0.183	0.201	0.860
$[\eta'_F]_6$	**–0.884**	0.150	0.255	0.223	–0.113	0.931
$[\eta'_F]_7$	**–0.950**	0.092	0.131	0.160	0.040	0.955
η_B	.0537	0.056	–0.316	**–0.778**	0.038	0.998
η'_B	0.270	0.281	–0.454	**–0.775**	–0.087	0.966
η_F^{local}	0.320	**–0.917**	0.098	0.175	0.011	0.983
%Variance	0.454	0.190	0.171	0.122	0.020	0.957

Again, when $\Sigma\alpha$ is used along with the shape parameter $[\Sigma\alpha]_P/\Sigma\alpha$, an equation with 77.2% explained variance and 75.4% predicted variance is obtained.

$$pC = 0.385(\pm 0.077)\sum\alpha + 1.885(\pm 0.821)[\sum\alpha]_P / \sum\alpha - 1.712(\pm 0.408)$$
$$n = 50, Q^2 = 0.754, R_a{}^2 = 0.772, R^2 = 0.781, R = 0.884, s = 0.318, F = 84.0(df\,2,47), \quad (17)$$
$$AVRES = 0.250, PRESS = 5.326, SDEP = 0.326, S_{PRESS} = 0.337, \Pr es_{av} = 0.267$$

Positive coefficient of $[\sum\alpha]_P/\sum\alpha$ in Eq. (17) suggests that toxicity increases as branching and number of substitutions in the phenol nucleus increase.

On inclusion of the parameter N_e in Eq. (16), explained variance rises to 91.7% and predicted variance rises to 91.0%.

$$\begin{aligned} &pC = 1.237(\pm0.290)\sum\alpha - 0.063(\pm0.024)[\sum\alpha]^2 + 0.312(\pm0.064)N_e - 3.891(\pm1.607) \\ &n = 50, Q^2 = 0.910, R_a{}^2 = 0.917, R^2 = 0.923, R = 0.960, s = 0.191, F = 182.6(df\,3,46), \\ &AVRES = 0.155, PRESS = 1.959, SDEP = 0.198, S_{PRESS} = 0.206, \mathrm{Pr}es_{av} = 0.168 \end{aligned} \quad (18)$$

Positive coefficient of N_e in Eq. (18) suggests that presence of electronegative atoms in the substituent positions increase toxicity. When $[\sum\varepsilon/N]_{sub}$ is used instead of N_e in Eq. (18), statistical quality of the resultant relation is slightly inferior.

$$\begin{aligned} &pC = 1.305(\pm0.326)\sum\alpha - 0.072(\pm0.026)[\sum\alpha]^2 + 1.469(\pm0.362)[\sum\varepsilon/N]_{sub} \\ &-4.535(\pm1.811) \\ &n = 50, Q^2 = 0.885, R_a{}^2 = 0.895, R^2 = 0.902, R = 0.950, s = 0.215, F = 140.9(df\,3,46), \\ &AVRES = 0.168, PRESS = 2.486, SDEP = 0.223, S_{PRESS} = 0.232, \mathrm{Pr}es_{av} = 0.181 \end{aligned} \quad (19)$$

However, it reconfirms that presence of electronegative atoms in the substituent positions increases toxicity. If $[N_e]_o$ is used in Eq. (18) instead of N_e, considerable decrease in statistical quality occurs.

$$\begin{aligned} &pC = 1.315(\pm0.457)\sum\alpha - 0.071(\pm0.037)[\sum\alpha]^2 + 0.258(\pm0.152)[N_e]_o - 3.934(\pm2.529) \\ &n = 50, Q^2 = 0.783, R_a{}^2 = 0.795, R^2 = 0.808, R = 0.899, s = 0.301, F = 64.4(df\,3,46), \\ &AVRES = 0.232, PRESS = 4.713, SDEP = 0.307, S_{PRESS} = 0.320, \mathrm{Pr}es_{av} = 0.249 \end{aligned} \quad (20)$$

On inclusion of shape parameter $[\sum\alpha]_P/\sum\alpha$ in Eqs. (18) and (19), statistical quality of the relations improves further:

$$\begin{aligned} &pC = 0.952(\pm0.416)\sum\alpha - 0.041(\pm0.033)[\sum\alpha]^2 + 0.661(\pm0.708)[\sum\alpha]_p/\sum\alpha \\ &+0.298(\pm0.064)N_e - 3.246(\pm2.298) \\ &n = 50, Q^2 = 0.914, R_a{}^2 = 0.922, R^2 = 0.928, R = 0.963, s = 0.186, F = 145.4(df\,4,45), \\ &AVRES = 0.147, PRESS = 1.863, SDEP = 0.193, S_{PRESS} = 0.203, \mathrm{Pr}es_{av} = 0.162 \end{aligned} \quad (21)$$

$$\begin{aligned} &pC = 0.983(\pm0.470)\sum\alpha - 0.047(\pm0.037)[\sum\alpha]^2 + 0.740(\pm0.796)[\sum\alpha]_p/\sum\alpha \\ &+1.394(\pm0.362)[\sum\varepsilon/N]_{sub} - 3.780(\pm2.598) \\ &n = 50, Q^2 = 0.885, R_a{}^2 = 0.901, R^2 = 0.909, R = 0.953, s = 0.209, F = 112.3(df\,4,45), \\ &AVRES = 0.162, PRESS = 2.489, SDEP = 0.223, S_{PRESS} = 0.235, \mathrm{Pr}es_{av} = 0.182 \end{aligned} \quad (22)$$

The regression coefficients of $[\sum\alpha]_P/\sum\alpha$ in Eqs. (21) and (22) are significant at 90% level. When $[\sum\beta'_{ns}]_m$ is incorporated in Eqs. (21) and (22), further increase in statistical quality is observed.

$$pC = 0.959(\pm0.360)\sum\alpha - 0.042(\pm0.028)[\sum\alpha]^2 + 0.661(\pm0.612)[\sum\alpha]_p / \sum\alpha$$
$$+0.238(\pm0.063)N_e + 0.507(\pm0.253)[\sum\beta'_{ns}]_m - 3.278(\pm1.988)$$
$$n = 50, Q^2 = 0.936, R_a{}^2 = 0.942, R^2 = 0.948, R = 0.973, s = 0.161, F = 159.1(df\,5,44),$$
$$AVRES = 0.124, PRESS = 1.381, SDEP = 0.166, S_{PRESS} = 0.177, \Pr es_{av} = 0.138 \quad (23)$$

$$pC = 0.982(\pm0.403)\sum\alpha - 0.046(\pm0.032)[\sum\alpha]^2 + 0.726(\pm0.682)[\sum\alpha]_p / \sum\alpha$$
$$+1.068(\pm0.348)[\sum\varepsilon / N]_{sub} + 0.577(\pm0.279)[\sum\beta'_{ns}]_m - 3.688(\pm2.230)$$
$$n = 50, Q^2 = 0.918, R_a{}^2 = 0.927, R^2 = 0.935, R = 0.967, s = 0.179, F = 126.1(df\,5,44),$$
$$AVRES = 0.134, PRESS = 1.778, SDEP = 0.189, S_{PRESS} = 0.201, \Pr es_{av} = 0.153 \quad (24)$$

Positive coefficient of $[\sum\beta'_{ns}]_m$ in Eqs. (23) and (24) suggests that such *meta* substituents which have high $\sum\beta'_{ns}$ values are conducive to the toxicity.

When $[N_e]_o$ is used in Eq. (23) instead of N_e, considerable decrease in statistical quality occurs.

$$pC = 0.917(\pm0.469)\sum\alpha - 0.039(\pm0.037)[\sum\alpha]^2 + 0.828(\pm0.795)[\sum\alpha]_p / \sum\alpha$$
$$+0.218(\pm0.108)[N_e]_o + 0.948(\pm0.290)[\sum\beta'_{ns}]_m - 3.130(\pm2.587)$$
$$n = 50, Q^2 = 0.894, R_a{}^2 = 0.901, R^2 = 0.911, R = 0.955, s = 0.209, F = 90.2(df\,5,44),$$
$$AVRES = 0.159, PRESS = 2.304, SDEP = 0.215, S_{PRESS} = 0.229, \Pr es_{av} = 0.177 \quad (25)$$

Using $[N_e]_m$ instead of $[\sum\beta'_{ns}]_m$ in Eq. (25), a statistically comparable relation is generated:

$$pC = 0.925(\pm0.478)\sum\alpha - 0.040(\pm0.038)[\sum\alpha]^2 + 0.882(\pm0.810)[\sum\alpha]_p / \sum\alpha$$
$$+0.206(\pm0.111)[N_e]_o + 0.484(\pm0.154)[N_e]_m - 3.154(\pm2.640)$$
$$n = 50, Q^2 = 0.889, R_a{}^2 = 0.897, R^2 = 0.907, R = 0.953, s = 0.214, F = 86.3(df\,5,44),$$
$$AVRES = 0.163, PRESS = 2.409, SDEP = 0.219, S_{PRESS} = 0.234, \Pr es_{av} = 0.182 \quad (26)$$

Thus, electronegative atoms in *meta* substitutions are conductive to the toxicity. Again, using $[\sum\beta'_{ns}]_p$ or $[N_e]_p$ as descriptor for electronic property of the *para* substituents, the following relations are obtained:

$$pC = 1.155(\pm0.232)\sum\alpha - 0.058(\pm0.019)[\sum\alpha]^2 + 0.193(\pm0.066)N_e$$
$$+0.660(\pm0.250)[\sum\beta'_{ns}]_m + 0.381(\pm0.216)[\sum\beta'_{ns}]_p - 3.671(\pm1.284)$$
$$n = 50, Q^2 = 0.945, R_a{}^2 = 0.950, R^2 = 0.955, R = 0.977, s = 0.149, F = 186.3(df\,5,44),$$
$$AVRES = 0.108, PRESS = 1.199, SDEP = 0.155, S_{PRESS} = 0.165, \Pr es_{av} = 0.120 \quad (27)$$

$$pC = 1.097(\pm0.248)\sum\alpha - 0.054(\pm0.020)[\sum\alpha]^2 + 0.202(\pm0.081)[N_e]_o$$
$$+1.040(\pm0.220)[\sum\beta'_{ns}]_m + 0.629(\pm0.200)[\sum\beta'_{ns}]_p - 3.484(\pm1.369)$$
$$n = 50, Q^2 = 0.936, R_a{}^2 = 0.943, R^2 = 0.949, R = 0.974, s = 0.159, F = 163.3(df\,5,44),$$
$$AVRES = 0.115, PRESS = 1.392, SDEP = 0.167, S_{PRESS} = 0.178, \Pr es_{av} = 0.130 \quad (28)$$

$$\begin{aligned} pC &= 0.414(\pm 0.045)\sum\alpha + 0.930(\pm 0.532)[\sum\alpha]_p / \sum\alpha + 0.194(\pm 0.094)[N_e]_o \\ &+1.013(\pm 0.253)[\sum\beta'_{ns}]_m + 0.557(\pm 0.246)[\sum\beta'_{ns}]_p - 1.784(\pm 0.296) \\ n &= 50, Q^2 = 0.916, R_a{}^2 = 0.926, R^2 = 0.933, R = 0.966, s = 0.181, F = 123.2(df\,5,44), \\ AVRES &= 0.130, PRESS = 1.824, SDEP = 0.191, S_{PRESS} = 0.204, \Pr es_{av} = 0.146 \end{aligned} \quad (29)$$

$$\begin{aligned} pC &= 0.428(\pm 0.042)\sum\alpha + 1.103(\pm 0.485)[\sum\alpha]_p / \sum\alpha + 0.195(\pm 0.074)N_e \\ &+0.616(\pm 0.283)[\sum\beta'_{ns}]_m + 0.274(\pm 0.258)[\sum\beta'_{ns}]_p - 1.893(\pm 0.280) \\ n &= 50, Q^2 = 0.929, R_a{}^2 = 0.937, R^2 = 0.943, R = 0.971, s = 0.167, F = 145.9(df\,5,44), \\ AVRES &= 0.122, PRESS = 1.550, SDEP = 0.176, S_{PRESS} = 0.188, \Pr es_{av} = 0.137 \end{aligned} \quad (30)$$

$$\begin{aligned} pC &= 0.955(\pm 0.370)\sum\alpha - 0.041(\pm 0.029)[\sum\alpha]^2 + 0.708(\pm 0.630)[\sum\alpha]_p / \sum\alpha \\ &+0.191(\pm 0.086)[N_e]_o + 0.469(\pm 0.119)[N_e]_m + 0.330(\pm 0.120)[N_e]_p - 3.278(\pm 2.044) \\ n &= 50, Q^2 = 0.932, R_a{}^2 = 0.938, R^2 = 0.946, R = 0.973, s = 0.165, F = 125.4(df\,6,43), \\ AVRES &= 0.121, PRESS = 1.473, SDEP = 0.172, S_{PRESS} = 0.185, \Pr es_{av} = 0.138 \end{aligned} \quad (31)$$

Among the above relations, Eq. (27) shows maximum statistical quality: it can explain 95.0% and predict 94.5% of the variance of the toxicity. However, from Table 9 it is observed that factor 4 has high loading in the variables $[\sum\beta'_{ns}]_p$ and $[N_e]_p$. Again, N_e has also considerable loading in factor 4. Thus, Eq. (27) is not considered as the best equation (because it contains the terms $[\sum\beta'_{ns}]_p$ and N_e, both of which are considerably loaded with factor 4); instead of this, Eq. (28), which is next in quality, is considered as the best one. Optimum $\sum\alpha$ values calculated from Eqs. (27) and (28) are 9.957 and 10.157 respectively. 2–Flurophenol (**3**), 3–chloro–4–flurophenol (**19**) and 3,4,5–trimethylphenol (**37**) act as outliers for Eq. (27), while 3,4,5–trimethylphenol (**37**), 4–bromo–6–chloro–2–methylphenol (**42**) and 4–chloro–2–isopropyl–5–methylphenol (**46**) act as outliers for Eq. (28). Positive coefficients of $[\sum\beta'_{ns}]_p$ and $[N_e]_p$ in Eqs. (27)–(31) suggest that such *para* substituents which have high $\sum\beta'_{ns}$ values or electronegative atoms are conducive to the toxicity.

When η'_B is used instead of $\sum\alpha]_P/\sum\alpha$ as shape parameter, the following equations are obtained:

$$\begin{aligned} pC &= 0.430(\pm 0.048)\sum\alpha + 9.423(\pm 6.259)\eta'_B + 0.175(\pm 0.083)N_e \\ &+0.729(\pm 0.311)[\sum\beta'_{ns}]_m + 0.506(\pm 0.264)[\sum\beta'_{ns}]_p - 1.892(\pm 0.330) \\ n &= 50, Q^2 = 0.906, R_a{}^2 = 0.923, R^2 = 0.930, R = 0.965, s = 0.185, F = 117.8(df\,5,44), \\ AVRES &= 0.141, PRESS = 2.031, SDEP = 0.202, S_{PRESS} = 0.215, \Pr es_{av} = 0.163 \end{aligned} \quad (32)$$

$$\begin{aligned} pC &= 0.404(\pm 0.049)\sum\alpha + 10.612(\pm 6.579)\eta'_B + 0.691(\pm 0.422)[\sum\varepsilon / N]_{sub} \\ &+0.825(\pm 0.319)[\sum\beta'_{ns}]_m + 0.591(\pm 0.270)[\sum\beta'_{ns}]_p - 2.059(\pm 0.398) \\ n &= 50, Q^2 = 0.896, R_a{}^2 = 0.913, R^2 = 0.922, R = 0.960, s = 0.196, F = 103.6(df\,5,44), \\ AVRES &= 0.145, PRESS = 2.259, SDEP = 0.213, S_{PRESS} = 0.227, \Pr es_{av} = 0.167 \end{aligned} \quad (33)$$

$$pC = 0.458(\pm 0.054)\sum\alpha + 8.512(\pm 7.475)\eta'_B + 0.224(\pm 0.113)[N_e]_o$$
$$+0.476(\pm 0.160)[N_e]_m + 0.350(\pm 0.161)[N_e]_p - 1.945(\pm 0.378)$$
$$n = 50, Q^2 = 0.864, R_a{}^2 = 0.890, R^2 = 0.901, R = 0.949, s = 0.221, F = 79.9(df\,5, 44),$$
$$AVRES = 0.167, PRESS = 2.944, SDEP = 0.243, S_{PRESS} = 0.259, \Pr es_{av} = 0.193 \quad (34)$$

$$pC = 0.375(\pm 0.055)\sum\alpha + 10.286(\pm 6.905)\eta'_B + 0.183(\pm 0.144)[\sum\beta_s]_o$$
$$+1.144(\pm 0.284)[\sum\beta'_{ns}]_m + 0.777(\pm 0.251)[\sum\beta'_{ns}]_p - 1.699(\pm 0.375)$$
$$n = 50, Q^2 = 0.889, R_a{}^2 = 0.905, R^2 = 0.915, R = 0.957, s = 0.205, F = 94.7(df\,5, 44),$$
$$AVRES = 0.157, PRESS = 2.397, SDEP = 0.219, S_{PRESS} = 0.233, \Pr es_{av} = 0.178 \quad (35)$$

$$pC = 0.414(\pm 0.046)\sum\alpha + 10.259(\pm 6.134)\eta'_B + 0.205(\pm 0.094)[N_e]_o$$
$$+1.065(\pm 0.253)[\sum\beta'_{ns}]_m + 0.714(\pm 0.227)[\sum\beta'_{ns}]_p - 1.851(\pm 0.315)$$
$$n = 50, Q^2 = 0.912, R_a{}^2 = 0.924, R^2 = 0.932, R = 0.965, s = 0.183, F = 120.7(df\,5, 44),$$
$$AVRES = 0.136, PRESS = 1.911, SDEP = 0.196, S_{PRESS} = 0.208, \Pr es_{av} = 0.155 \quad (36)$$

Positive coefficients of η'_B in Eqs. (32)–(36) suggest that toxicity increase with increase in branching.

Table 10 shows that the data matrix [B] composed of 17 variables (including response variable *pC*) can be explained to the extent of 95.7% by 5 factors. Biological activity (*pC*) is highly loaded with factor 3 which has in turn high loading in $[\sum\alpha]_P$ and $[\sum\alpha]_P/\sum\alpha$. Apart from this, *pC* has significant loadings with factor 2 (highly loaded in η_F, η'_F and η_F^{local}) and factor 1 (highly loaded in η, η_R, $[\eta'_F]_1$, $[\eta'_F]_2$, $[\eta'_F]_3$, $[\eta'_F]_4$, $[\eta'_F]_5$, $[\eta'_F]_6$ and $[\eta'_F]_7$). When η_R, η'_F and $[\sum\alpha]_P/\sum\alpha$ are used as predictor variables, the resultant equation can explain 80.5% and predict 77.7% of the variance.

$$pC = 0.069(\pm 0.020)\eta_R + 3.857(\pm 1.218)\eta'_F + 4.335(\pm 0.873)[\sum\alpha]_p/\sum\alpha$$
$$-3.883(\pm 1.006)$$
$$n = 50, Q^2 = 0.777, R_a{}^2 = 0.805, R^2 = 0.817, R = 0.904, s = 0.294, F = 68.3(df\,3, 46),$$
$$AVRES = 0.234, PRESS = 4.836, SDEP = 0.311, S_{PRESS} = 0.324, \Pr es_{av} = 0.258 \quad (37)$$

Eq. (37) suggests that toxicity increases with increase in the value of composite index for reference alkane (η_R) and functionality contribution (η'_F) apart of branching. When η_R is replaced by functionality values of individual positions of phenol nucleus, the best two relations are obtained with $[\eta'_F]_1$ and $[\eta'_F]_7$:

$$pC = -18.415(\pm 4.884)[\eta'_F]_1 + 4.978(\pm 1.105)\eta'_F$$
$$+4.248(\pm 0.835)[\sum\alpha]_p/\sum\alpha - 1.060(\pm 1.138)$$
$$n = 50, Q^2 = 0.809, R_a{}^2 = 0.823, R^2 = 0.834, R = 0.913, s = 0.280, F = 76.9(df\,3, 46),$$
$$AVRES = 0.219, PRESS = 4.144, SDEP = 0.288, S_{PRESS} = 0.300, \Pr es_{av} = 0.237 \quad (38)$$

The intercept of Eq. (38) is significant at the 90% level.

$$pC = -23.309(\pm 3.708)[\eta'_F]_7 + 5.135(\pm 0.843)\eta'_F + 4.404(\pm 0.627)[\sum\alpha]_p / \sum\alpha$$
$$n = 50, Q^2 = 0.825, R_a{}^2 = 0.836, R^2 = 0.842, R = 0.918, s = 0.270, F = 249.6(df\,3,47), \quad (39)$$
$$AVRES = 0.214, PRESS = 3.795, SDEP = 0.276, S_{PRESS} = 0.287, \Pr es_{av} = 0.226$$

Defining a new term $[\eta'_F]_{1+7}$ as the sum of $[\eta'_F]_1$ and $[\eta'_F]_7$, the following relation is obtained:

$$pC = -11.189(\pm 1.839)[\eta'_F]_{1+7} + 4.700(\pm 0.803)\eta'_F + 4.064(\pm 0.618)[\sum\alpha]_p / \sum\alpha$$
$$n = 50, Q^2 = 0.817, R_a{}^2 = 0.827, R^2 = 0.834, R = 0.913, s = 0.277, F = 236.6(df\,3,47), \quad (40)$$
$$AVRES = 0.215, PRESS = 3.976, SDEP = 0.282, S_{PRESS} = 0.294, \Pr es_{av} = 0.227$$

The intercept terms of Eqs. (39) and (40) are insignificant and thus have been set to zero. Eq. (40) suggests that functionality contributions of atoms 1 and 7 are important which actually implies importance of the phenolic –OH group.

Table 11. Results of leave–10%–out cross–validation applied on Eqs. (27) and (28). Model equation, $pC = \sum\beta_i x_i + \alpha$. Q^2 denotes cross–validated R^2. Average Pres means average of absolute values of predicted residuals.

Eq.	Number of cycles	Average regression coefficients (standard deviations)	Statistics Q^2 (Average Pres)
(27)	10 [a]	1.153 (0.031) $\sum\alpha$ –0.057 (0.002) $[\sum\alpha]^2$ +0.193 (0.005) N_e + 0.662 (0.046) $[\sum\beta'_{ns}]_m$ +0.383 (0.034) $[\sum\beta'_{ns}]_p$ –3.669 (0.093)	0.939 (0.115)
(28)	10 [a]	1.097 (0.043) $\sum\alpha$ –0.054 (0.004) $[\sum\alpha]^2$ +0.203 (0.013) $[N_e]_o$ + 1.041 (0.035) $[\sum\beta'_{ns}]_m$ +0.632 (0.034) $[\sum\beta'_{ns}]_p$ –3.484 (0.122)	0.930 (0.125)

[a] Compounds were deleted in 10 cycles in the following manner: (**1**, **11**, **21**,....**41**), (**2**, **12**, **22**,....**42**),..., (**10**, **20**, **30**,....**50**)

In the QSAR study made by Hall and Vaughn [35] on the present biological activity data set using electrotopological state atom index and kappa shape index, the best equation involved four descriptor variables (with one insignificant coefficient at 95% level), and showed correlation coefficient (*r*) of 0.96 and cross–validation r (r_{PRESS}) of 0.90. However, in the present study we could get a relation Eq. (18) with correlation coefficient of 0.960 using only three predictor variables and all significant coefficients. Eqs. (21), (23), (24), (27)–(32) and (36) are statistically better than the equation reported in reference [35]. Considering statistical quality of the equations, Eqs. (27) and (28) are considered as the best two equations describing the toxicity of phenols. The calculated and predicted values according to Eqs. (27) and (28) are shown in Table 4. Leave–10%–out cross–validation (Table 11) applied on these two equations show robustness of the relations.

4 CONCLUSIONS

The study shows that toxicity of phenols increase with increase in bulk and branching. Again, presence of electronegative atoms in substituent positions increase toxicity. Further, *meta* and *para* substituents with higher $\sum\beta'_{ns}$ values are conducive to the toxicity. Toxicity values also increase with increase in η_R and η'_F values. Functionality contribution of phenolic O and adjacent aromatic carbon show specific importance of the phenolic –OH group to the toxicity. It appears that acidity of phenols is an important factor for toxicity: presence of electronegative atoms in the substituent positions actually enhances acidity and thereby increases the toxicity.

In the present study, ETA indices could explore the important chemical information contributing to the toxicity of phenols and the relations generated could predict the activity of the compounds to a satisfactory extent (explained variance up to 95.0%, predicted variance up to 94.5%). Leave–10%–out cross–validation applied on the final equations show robustness of the relations. Thus, these indices deserve more extensive work on diverse biological activities and physicochemical properties of diverse categories of compounds to prove their utility on QSAR/QSPR research.

Acknowledgment

One of the authors (KR) remembers stimulating discussions with Sri Dipak Kumar Pal regarding the TAU formalism. The authors thank Dr. C. Sengupta of Dept. Pharmaceutical Technology, J. U., for constant help and inspiration. Financial grant from J. U. Research Fund is also thankfully acknowledged.

5 REFERENCES

[1] F. Choplin, Computers and the medicinal chemist; in: Comprehensive Medicinal Chemistry, vol. 4, Eds. C. Hansch, P. G. Sammes, and J. B. Taylor, Pergamon Press, Oxford, 1990, pp. 33–58.
[2] R. Franke, Theoretical Drug Design Methods, Elsevier, Amsterdam, 1984.
[3] P. R. Andrews, and M. Tintelnot, Intermolecular forces and molecular binding; in: Comprehensive Medicinal Chemistry, vol. 4, Eds. C. Hansch, P. G. Sammes, and J. B. Taylor, Pergamon Press, Oxford, 1990, pp. 321–347.
[4] Y. C. Martin, Theoretical basis of medicinal chemistry: Structure–activity relationships and three–dimensional structures of small and macromolecules; in: *Modern Drug Research*. *Paths to Better and Safer Drugs*, Eds. Y. C. Martin, E. Kutter, and V. Anstel, Marcel Dekker, Inc., 1989, pp. 161–216.
[5] P. C. Jurs, S. L. Dixon, and L. M. Eglof, Representations of molecules; in: *Chemometric Methods in Molecular Design*, Ed. H. van de Waterbeemd, VCH, Weinheim, 1995, pp. 15–38.
[6] A. K. Debnath, Quantitative structure–activity relationship (QSAR): A versatile tool in drug design, in: *Combinatorial Library Design and Evaluation*, Eds. A. K. Ghose, and V. N. Viswanadhan, Marcel Dekker, Inc., NY, 2001, pp. 73–129.
[7] R. Gozalbes, J. P. Doucet, and F. Derouin, Applications of Topological Descriptors in QSAR and Drug Design: History and New Trends, *Curr*. *Drug Targets Infect*. *Disord*. **2002**, *2*, 93–102.
[8] C. Silipo, and A. Vittoria, Three–dimensional structure of drugs; in: *Comprehensive Medicinal Chemistry*, vol. 4, Eds. C. Hansch, P. G. Sammes and J. B. Taylor, Pergamon Press, Oxford, 1990, pp. 153–204.
[9] E. Estrada, and H. Gonzalez, What are the Limits of Applicability for Graph Theoretic Descriptors in QSPR/QSAR? Modeling Dipole Moments of Aromatic Compounds with TOPS–MODE Descriptors, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **2003**, *43*, 75–84.
[10] E. Estrada, E. Molina, and I. Perdomo–Lopez, Can 3D Structural Parameters be Predicted from 2D (Topological) Molecular Descriptors? *J*. *Chem*. *Inf*. *Comput*. *Sci*. **2001**, *41*, 1015–1021.
[11] O. Ivanciuc, T. Ivanciuc, and A. T. Balaban, Quantitative Structure–Property Relationship Evaluation of Structural Descriptors Derived from the Distance and Reverse Wiener Matrices, *Internet Electron*. *J*. *Mol*. *Des*. **2002**, *1*, 467–487, http://www.biochempress.com.

[12] S. C. Basak, and B. D. Gute, Characterization of Molecular Structures Using Topological Indices, *SAR QSAR Environ. Res.* **1997**, *7*, 1–21.
[13] J. Devillers and A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach Science Publishers, Netherlands, 1999.
[14] A. Mercader, E. A. Castro, and A. A. Toropov, Maximum Topological Distances Based Indices as Molecular Descriptors for QSPR. 4. Modeling the Enthalpy of Formation of Hydrocarbons from Elements, *Int. J. Mol. Sci.* **2001**, *2*, 121–132, http://www.mdpi.org/ijms.
[15] I. Rios–Santamarina, R. García–Domenech, J. Cortijo, P. Santamaria, E. J. Morcillo, and J. Gálvez, Natural Compounds with Bronchodilator Activity Selected by Molecular Topology, *Internet Electron. J. Mol. Des.* **2002**, *1*, 70–79, http://www.biochempress.com.
[16] A. A. Toropov and A. P. Toropova, QSAR Modeling of Mutagenicity Based on Graphs of Atomic Orbitals, *Internet Electron. J. Mol. Des.* **2002**, *1*, 108–114, http://www.biochempress.com.
[17] D. J. G. Marino, P. J. Peruzzo, E. A. Castro, and A. A. Toropov, QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants, *Internet Electron. J. Mol. Des.* **2002**, *1*, 115–133, http://www.biochempress.com.
[18] S.–S. Liu, H.–L. Liu, Y.–Y. Shi, and L.–S. Wang, QSAR of Cyclooxygenase–2 (COX–2) Inhibition by 2,3–Diarylcyclopentenones Based on MEDV–13, *Internet Electron. J. Mol. Des.* **2002**, *1*, 310–318, http://www.biochempress.com.
[19] R. García–Domenech, A. Catalá–Gregori, C. Calabuig, G. Antón–Fos, L. del Castillo, and J. Gálvez, Predicting Antifungal Activity: A Computational Screening Using Topological Descriptors, *Internet Electron. J. Mol. Des.* **2002**, *1*, 339–350, http://www.biochempress.com.
[20] S.–S. Liu, S.–H. Cui, Y.–Y. Shi, and L.–S. Wang, A Novel Variable Selection and Modeling Method based on the Prediction for QSAR of Cyclooxygenase–2 Inhibition by Thiazolone and Oxazolone Series, *Internet Electron. J. Mol. Des.* **2002**, *1*, 610–619, http://www.biochempress.com.
[21] B. S. Junkes, R. D. M. C. Amboni, R. A. Yunes, and V. E. F. Heinzen, Semiempirical Topological Index: A Novel Molecular Descriptor for Quantitative Structure–Retention Relationship Studies, *Internet Electron. J. Mol. Des.* **2003**, *2*, 33–49, http://www.biochempress.com.
[22] G. W. Kauffman and P. C. Jurs, QSAR and K–Nearest Neighbour Classification Analysis of Selective Cyclooxygenase–2 Inhibitors Using Topologically–Based Numerical Descriptors, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553–1560.
[23] R. Garcia–Domenech, J. V. de Julian–Ortiz, M. J. Duart, J. M. Garcia–Torrecillas, G. M. Anton–Fos, I. Rios–Santamarina, C. de Gregorio–Alapont, and J. Galvez, Search of A Topological Pattern to Evaluate Toxicity of Heterogeneous Compounds, *SAR QSAR Environ. Res.* **2001**, *12*, 237–254.
[24] G. A. Bakken and P. C. Jurs, Classification of Multidrug–Resistance Reversal Agents Using Structure–Based Descriptors and Linear Discriminant Analysis, *J. Med. Chem.* **2000**, *43*, 4534–4541.
[25] D. K. Pal, C. Sengupta, and A. U. De, A New Topochemical Descriptor (TAU) in Molecular Connectivity Concept: Part I — Aliphatic Compounds, *Indian J. Chem.* **1988**, *27B*, 734–739.
[26] D. K. Pal, C. Sengupta, and A. U. De, 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.
[27] D. K. Pal, M. Sengupta, C. Sengupta, and A. U. De, QSAR with TAU (τ) indices: Part I – Polymethylene Primary Diamines as Amebicidal Agents, *Indian J. Chem.* **1990**, *29B*, 451–454.
[28] D. K. Pal, S. K. Purkayastha, C. Sengupta, and A. U. De, Quantitative Structure–Property Relationships with TAU indices: Part I – Research Octane Numbers of Alkane Fuel Molecules, *Indian J. Chem.*, **1992**, *31B*, 109–114.
[29] K. Roy, D. K. Pal, A. U. De, and C. Sengupta, 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.
[30] K. Roy, D. K Pal, A. U. De, and C. Sengupta, 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.
[31] K. Roy and A. Saha, 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.
[32] K. Roy and A. Saha, 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.
[33] K. Roy and A. Saha, QSPR with TAU Indices: Water Solubility of Diverse Functional Acyclic Compounds, *Internet Electron. J. Mol. Des.* **2003**, *2*, 475–491, http://www.biochempress.com.
[34] I. Moriguchi, Y. Canada, and K. Komatsu, van der Waals Volume and the Related Parameters for Hydrophobicity

in Structure–Activity Studies, *Chem*. *Pharm*. *Bull*. **1976**, *24*, 1799–1806.
[35] L. H. Hall and T. A. Vaughn, QSAR of Phenol Toxicity Using Electrotopological State and Kappa Shape Indices, *Med*. *Chem*. *Res*. **1997**, *7*, 407–416.
[36] P. J. Lewi, Multivariate data analysis in structure–activity relationships; in: Drug Design, vol. 10, Ed. E. J. Ariens, Academic Press, New York, 1980, pp. 307–342.
[37] R. Franke and A. Gruska, Principal component and factor analysis; in: *Chemometric Methods in Molecular Design*, vol. 2, Ed. H. van de Waterbeemd, VCH, Weinheim, 1995, pp. 113–163.
[38] STATISTICA is a statistical software of Statsoft Inc., USA.
[39] The GW–BASIC programs *RRR*98, *KRETA*1, *KRETA*2, *KRPRES*1 *and KRPRES*2 were developed by Kunal Roy and standardized using known data sets.
[40] G. W. Snedecor and W. G. Cochran, *Statistical Methods*; Oxford & IBH Publishing Co. Pvt. Ltd., New Delhi, 1967, pp. 381 – 418.
[41] S. Wold and L. Eriksson, Statistical validation of QSAR results; in: *Chemometric Methods in Molecular Design*, Ed. H. van de Waterbeemd, VCH, Weinheim, 1995, pp. 309–318.