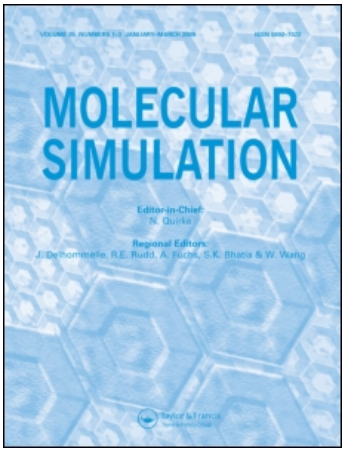

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

QSTR with extended topochemical atom (ETA) indices. 11. Comparative QSAR of acute NSAID cytotoxicity in rat hepatocytes using chemometric tools

Kunal Roy[a]; Gopinath Ghosh[a]
[a] Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India

To cite this Article Roy, Kunal and Ghosh, Gopinath(2009) 'QSTR with extended topochemical atom (ETA) indices. 11. Comparative QSAR of acute NSAID cytotoxicity in rat hepatocytes using chemometric tools', Molecular Simulation, 35: 8, 648 — 659
To link to this Article: DOI: 10.1080/08927020902744664
URL: http://dx.doi.org/10.1080/08927020902744664

Molecular Simulation
Vol. 35, No. 8, July 2009, 648–659

Taylor & Francis
Taylor & Francis Group

QSTR with extended topochemical atom (ETA) indices. 11. Comparative QSAR of acute NSAID cytotoxicity in rat hepatocytes using chemometric tools

Kunal Roy* and Gopinath Ghosh

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

(*Received 7 July 2008; final version received 11 January 2009*)

An attempt was made to develop quantitative structure–toxicity relationships (QSTRs) for acute rat hepatocyte cytotoxicity of a series of non-steroidal anti-inflammatory drugs (NSAIDs) with extended topochemical atom (ETA) indices. The ETA models were compared with those developed with a pool of other topological indices. Finally, attempt was made to develop models from the combined pool of topological (ETA and non-ETA) descriptors. The chemometric tools used for model development were multiple linear regression with factor analysis as the variable selection tool, stepwise regression, principal component regression analysis, partial least squares (PLS), genetic function approximation (GFA) and genetic PLS (G/PLS). Use of ETA descriptors along with non-ETA ones increased the statistical quality of the non-ETA models in different techniques except stepwise regression and GFA. The best three models came from GFA, G/PLS and PLS techniques (leave-one-out Q^2 values of 0.871, 0.854 and 0.834, respectively, using combined set of descriptors except for GFA). Use of the ETA parameters suggests that the toxicity increases with bulk and degree of branching. Moreover, heteroatom count and degree of unsaturation are also important for the toxicity.

Keywords: QSAR; QSTR; NSAID; ETA; chemometric tools

1. Introduction

Non-steroidal anti-inflammatory drugs (NSAIDs) are extensively used for the treatment of a number of pathological conditions such as headache, fever, arthritis and inflammation [1]. They act by the inhibition of prostaglandin synthase. However, their toxic effects (e.g. gastrointestinal lesions, bleeding, relapse of classic inflammatory bowel disease, suppression of renal functions, complications of diverticular disease and hepatotoxicity) limit their clinical usefulness [2–5]. These adverse drug reactions may cause significant health problems that may contribute to the morbidity and mortality of patients and these are often responsible for the discontinuation of NSAID therapy [6,7]. Drug-induced liver injury, carcinogenicity and/or cardiac liability associated with the blockade of hERG (human ether a-go-go) are the most frequent reasons for the withdrawal of some approved NSAIDs from market [8,9]. The pharmaceutical industry has endeavoured to discover, develop and market efficacious NSAIDs with reduced side effects so that treatment can be continued. This can be accelerated with the help of *in silico* tools, such as quantitative structure–activity relationships (QSARs), as an alternative rational approach for toxicity evaluation of drugs involving limited animal experimentation. With the help of QSAR, it is also possible to reduce expense, time and unnecessary animal sacrifice for the development of new medicinal agents with reduced toxicity. Moreover, reliability and accuracy of toxicity predictions may be achieved by identifying toxicophores with the help of QSAR. These predictions can guide the design of chemical libraries for hit and lead optimisation. In the lead optimisation phase of the synthetic chemistry project, various QSAR techniques with the aid of *in silico* and chemometric tools have been proposed and used according to the robustness and prediction capacity of the QSAR models. QSARs have emerged as an indispensable tool for predicting toxicity value of new chemicals. It has been demonstrated that the toxicity value of NSAIDs can be predicted with the help of QSAR. Siraki et al. [10] have applied quantitative structure–toxicity relationship (QSTR) for the calculation of acute NSAID cytotoxicity in rat hepatocytes using physicochemical descriptors. Lewis et al. [11] have performed QSAR analysis of the acute toxicities and anti-inflammatory activities of 16 analogues of benoxaprofen to identify a drug candidate likely to have similar anti-inflammatory activity to benoxaprofen but with lower toxicity. Continued research in this field led to the arrival of the cyclooxygenase-2 inhibitors, a new class of NSAID, prominently represented by Vioxx (Rofecoxib, Merck, NJ, USA), Celebrex (Celecoxib, Pfizer, NY, USA) [8] and Bextra (Valdecoxib,

*Corresponding author. Email: kunalroy_in@yahoo.com

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020902744664
http://www.informaworld.com

Pfizer, NY, USA). Although these drugs demonstrate analgesic and anti-inflammatory activity comparable with the traditional NSAIDs, they cause markedly less gastroduodenal irritation than aspirin and the traditional NSAIDs [5]. However, due to demonstration of increased risk of heart attacks and stroke in some individuals, Merck voluntarily withdrew Vioxx (Rofecoxib) from the market in 2004. The medical needs of patients suffering from chronic inflammatory disorders remain unmet and no current therapy constitutes a cure for this. A great amount of concerted efforts are being made worldwide to find a safe and acceptable anti-inflammatory drug.

An important advancement in *in silico* treatment of chemical structures and QSAR has been the application of a mathematical technique, namely 'graph theory', to chemistry [12]. In chemical graph theory, molecular structures are represented as hydrogen-suppressed graphs, commonly known as molecular graphs, in which the atoms are represented by vertices and the bonds by edges. The connections between the atoms can be described by various types of topological matrices (e.g. distance or adjacency matrices), which can be mathematically manipulated so as to derive a single number, usually known as graph invariant, graph-theoretical index or topological index (TI). In consequence, the TIs can be defined as two-dimensional descriptors that can be easily calculated from the molecular graphs, and do not depend on the way the graph is depicted or labelled. They offer a simple way of measuring molecular branching, shape and size [13], which is used to develop QSAR models to predict biological activity or toxicity.

In this present communication, we have developed QSTR models with extended topochemical atom (ETA) [14–24] indices for acute NSAID cytotoxicity [10] in rat hepatocytes using different chemometric tools. QSTR models developed with ETA descriptors have been compared with those with selected non-ETA (topological) and combined set of descriptors. Though Siraki et al. [10] used physicochemical parameters to model this data set, we have presently used only topological indices derived from graph theoretic approaches. The objectives of this paper have been to (i) develop predictive QSAR models for the hepatotoxicity of NSAIDs using topological descriptors and (ii) evaluate performance of ETA parameters in comparison with other topological parameters in deriving predictive QSAR models.

2. Materials and methods

2.1 The data set

In the present communication, utility and prediction capacity of ETA parameters has been demonstrated in comparison with other topological indices taking acute NSAID cytotoxicity in rat hepatocytes [10] as the model data set (Table 1). Toxicity data of 15 NSAIDs of classes of benzoic acid and phenylacetic acid derivatives (Figure 1) have been considered in this paper. Though the number of available data points is limited, an attempt was made to develop the QSTR models considering that limited number of reports have been published on such QSTR models involving drug-induced toxicity.

2.2 Types of descriptors

2.2.1 ETA descriptors

The ETA [14–24] indices developed recently in our laboratory, based on refinement of TAU descriptors [25–37], which were developed in late eighties in the valence electron mobile (VEM) environment, have been used to generate the QSTR models for acute NSAID cytotoxicity in rat hepatocytes [10]. Definitions of the some basic parameters used in the ETA scheme to develop QSTR models are given below.

i) The core count (α) [14]. The core count of a non-hydrogen vertex [α] is defined as

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

In Equation (1), Z and Z^v represent the atomic number and valence electron number, respectively, while PN denotes the period number. Hydrogen atom being considered as the reference, α for hydrogen is taken as zero value.

ii) The measure of electronegativity (EN; ε) [14]. It has been defined in the following manner:

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

In Equation (2), Z^v is the valence electron number of the vertex. It is interesting to note that the α values of different atoms (which are commonly found in organic compounds) have high correlation ($r = 0.946$) with (uncorrected) van der Waals volume, while ε has good correlation ($r = 0.937$) with Pauling's EN scale.

iii) The VEM count (β) [14]. The VEM count β of ETA scheme for a non-hydrogen vertex is defined as:

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

In Equation (3), δ is a correction factor of value 0.5 per atom with loan pair of electrons capable of resonance with an aromatic ring (e.g. nitrogen of aniline, oxygen of phenol, etc.). The parameters σ and π denote the number of sigma and pi bonds, respectively, associated with the vertex in the hydrogen-suppressed graphs. For the calculation of the VEM count, contribution of a sigma bond (x) between two atoms of similar EN ($\Delta\varepsilon \leq 0.3$) is considered to be 0.5, and for a sigma bond between two atoms of different EN ($\Delta\varepsilon > 0.3$),

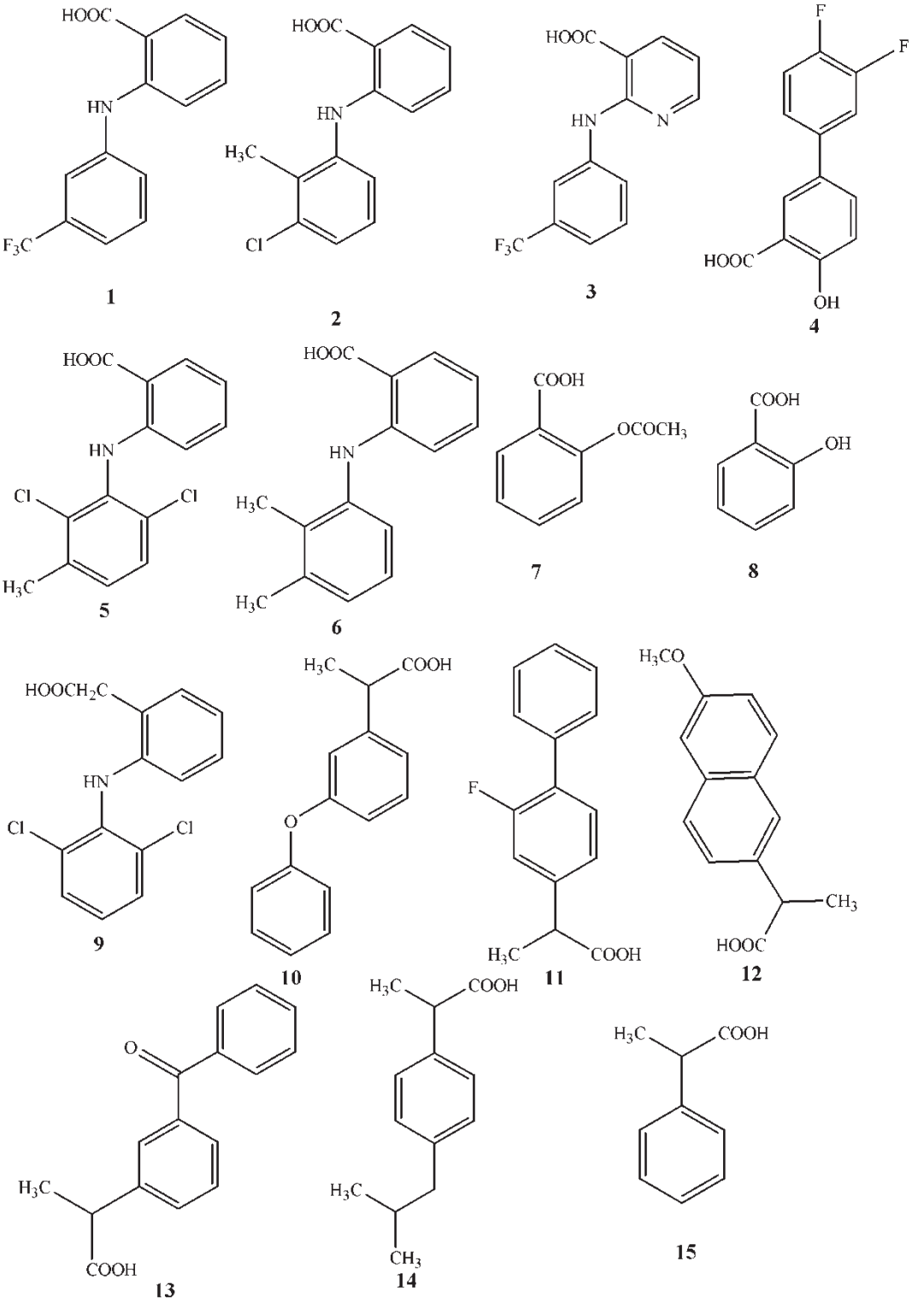

Figure 1. Structural features of 15 NSAIDs.

it is considered to be 0.75. Again, in case of pi bonds, contributions (y) are considered depending on the type of double bond: (i) for pi bond between two atoms of similar EN ($\Delta\varepsilon \leq 0.3$), y is taken to be 1; (ii) for pi bond between two atoms of different EN ($\Delta\varepsilon > 0.3$) or for conjugated (non-aromatic) pi system, y is considered to be 1.5; (iii) for aromatic pi system, y is taken as 2.

Table 1. Observed and calculated values of rat hepatotoxicity of 15 NSAIDs.

Sl No.	Name of the compounds	Cytotoxicity of NSAIDs in rat hepatocytes			
		Obs. [10]	Cal.[a]	Cal.[b]	Cal.[c]
1	Flufenamic acid	−2.08	−2.24	−2.14	−2.12
2	Tolfenamic acid	−2.23	−2.53	−2.63	−2.53
3	Niflumic acid	−2.23	−2.21	−2.14	−2.12
4	Diflunisal	−2.30	−2.28	−2.40	−2.43
5	Meclofenamic acid	−2.30	−2.49	−2.42	−2.44
6	Mefenamic acid	−2.60	−2.50	−2.68	−2.62
7	Acetylsalicylic acid	−3.48	−3.42	−3.54	−3.59
8	Salicylic acid	−3.85	−3.69	−3.77	−3.56
9	Diclofenac	−3.00	−2.66	−2.61	−2.62
10	Fenoprofen	−3.23	−3.43	−3.54	−3.49
11	Flurbiprofen	−3.48	−3.57	−3.48	−3.55
12	Naproxen	−3.60	−3.36	−3.40	−3.51
13	Ketoprofen	−3.60	−3.35	−3.31	−3.30
14	Ibuprofen	−3.70	−3.71	−3.39	−3.67
15	2-Phenylpropionic acid	−4.18	−4.41	−4.38	−4.32

[a] Calculated toxicity from Equation (18).
[b] Calculated toxicity from Equation (22).
[c] Calculated toxicity from Equation (23).

iv) The VEM vertex count (γ) [14]. The VEM vertex count γ_i of the ith vertex in a molecular graph is defined as:

$$\gamma_i = \frac{\alpha_i}{\beta_i}. \tag{4}$$

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

v) The composite index (η) [14]. The composite index η is defined in the following manner:

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

In Equation (5), both bonded and non-bonded interactions have been considered. The parameter r_{ij} stands for the topological distance between the ith and jth atoms. Again, when all heteroatoms and multiple bonds in the molecular graph are replaced by carbon and single bond, respectively, corresponding molecular graph may be considered as the reference alkane and the corresponding composite index value is designated as η_R.

vi) The functionality index (η'_F) [14]. 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 have defined another term η'_F as η_F/N_V.

vii) The atom level index [14]. One can determine contribution of a particular vertex or substructure to functionality in the following manner:

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

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

viii) The local index η^{local} [14]. When only bonded interactions are considered ($r_{ij} = 1$), the corresponding composite index is written as η^{local}.

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

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

ix) The branching index (η_B) [14]. It can be calculated as $\eta_N^{local} - \eta_R^{local} + 0.086 N_R$, where N_R stands for the number of rings in the molecular graph of the reference alkane. The N_R term in the branching index expression represents a correction factor for cyclicity. η_N^{local} indicates η value of the corresponding normal alkane (straight chain compound of same vertex count obtained from the reference alkane), which may be conveniently calculated as (when $N_V \geq 3$):

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

Table 2. Definitions of the different ETA parameters used in exploring QSARs of rat hepatoxicity of NSAIDs.

Variables	Definition
$\Sigma\alpha$	Sum of α values of all non-hydrogen vertices of a molecule
$[\Sigma\alpha]_p$	Sum of α values of all non-hydrogen vertices each of which is joined to only one other non-hydrogen vertex of the molecule
η	The composite ETA index
η_R	The composite index for the reference alkane
N_V	Vertex count (excluding hydrogen)
$\Sigma\beta'_s$	Sum of β'_s values of all non-hydrogen vertices of a molecule; $\sum\beta'_s$ is defined as $\left[\sum\beta_s\right]/N_V$
$\Sigma\beta'_{ns}$	Sum of β'_{ns} values of all non-hydrogen vertices of a molecule; $\sum\beta'_s$ is defined as $\left[\sum\beta_{ns}\right]/N_V$
$[\eta'_F]_{Cl}$	Functionality contribution for the chlorine atom
$[\eta'_F]_F$	Functionality contribution for the fluorine atom
$[\eta'_F]_{CH_3}$	Functionality contribution for the methyl group
$[\eta'_F]_{COOH}$	Functionality for the carboxylic group
$[\eta'_F]_{OH}$	Functionality for the hydroxyl group

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

x) The shape indices [14]. The terms like $(\Sigma\alpha)_p/\Sigma\alpha$, $(\Sigma\alpha)_Y/\Sigma\alpha$ and $(\Sigma\alpha)_X/\Sigma\alpha$ can be used as the shape parameters. The 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 of the reference alkane. The definitions of the important ETA parameters are given in Table 2.

2.2.2 *Non-ETA descriptors*

Apart from ETA descriptors, other selected topological descriptors also have been used in this paper. We have considered different non-ETA topological descriptors such as Wiener *W*, Balaban *J*, Zagreb, 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 ($^1\kappa$, $^2\kappa$, $^3\kappa$, $^1\kappa_\alpha$, $^2\kappa_\alpha$, $^3\kappa_\alpha$) and E-state parameters (S_sCH3, S_ssCH2, S_aaCH, S_dssC, S_aasC, S_ssssC, S_ssNH, S_sOH, S_dO, S_ssO, S_sCl, S_sF)

2.3 *Softwares used for descriptor generation*

The calculations of η, η_R, η_F, η_B and contributions of different vertices or fragments to η_F were done, using distance matrix and VEM vertex counts as inputs, using the GW-BASIC programs *KRETA1* and *KRETA2* developed by one of the authors [38]. We have also modelled the toxicity data using other selected non-ETA (topological) and combined set of variables, and compared the ETA models with models developed from other descriptors. The values for the non-ETA topological descriptors for the compounds of the current data set have been generated by QSAR+ and Descriptor+ modules of the Cerius 2 version 4.10 software [39].

2.4 *Model development*

2.4.1 *Statistical analysis performed*

For the development of models, six different statistical techniques were used: multiple linear regression with factor analysis (FA-MLR) as variable selection tool, stepwise regression, principal component regression analysis (PCRA), partial least squares (PLS), genetic function approximation (GFA) and genetic PLS (G/PLS).

In the case of FA-MLR, classical approach of multiple regression technique used as the final statistical tool for developing QSAR relations and FA [40,41] was used as the data preprocessing step to identify the important predictor variables contributing to the response variable (i.e. toxicity value) and to avoid collinearities among them.

In stepwise regression [42], a multiple-term linear equation was built step-by-step. The basic procedures involve (1) identifying an initial model, (2) iteratively 'stepping', i.e. repeatedly altering the model at the previous step by adding or removing a predictor variable in accordance with the 'stepping criteria' (in our case, $F = 4.0$ for inclusion; $F = 3.9$ for exclusion for the forward selection method) and (3) terminating the search when stepping is no longer possible given the stepping criteria, or when a specified maximum number of steps has been reached. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the equation; that variable is then included in the model, and the process starts again.

Attempt was made to perform PCRA [40] taking factor scores of the descriptor matrix (without toxicity value) as the predictor variables and adopting backward stepwise regression method. In this case, the principal components serve as latent variables. PCRA has an advantage that collinearities among independent variables are not a disturbing factor and that the number of variables included in the analysis may exceed the number of observations. In PCRA, all descriptors are assumed to be important, while the aim of FA is to identify relevant descriptors.

PLS is a generalisation of regression, which can handle data with strongly correlated and/or noisy or numerous independent variables [43]. It gives a reduced solution, which is statistically more robust than MLR. The linear PLS model finds 'new variables' (latent variables or independent scores), which are linear combinations of the original variables. To avoid over-fitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. Cross-validation is a practical and reliable method of testing this significance [44]. Application of PLS thus allows the construction of larger QSAR equations while still avoiding over-fitting and eliminating

most variables. PLS is normally used in combination with cross-validation to obtain the optimum number of components. This ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data [45]. In our study, based on the standardised regression coefficients, the variables with smaller coefficients were removed from the PLS regression, until there was no further improvement in Q^2 value, irrespective of the components.

GFA technique [46,47] was used to generate a population of equations rather than one single equation for correlation between the toxicity and topological descriptors. It provides an error measure, called the lack-of-fit (LOF) score that automatically penalises models with too many features. It also inspires the use of splines as a powerful tool for nonlinear modelling. GFA is done as follows: (i) an initial population of equations is generated by random choice of descriptors; (ii) pairs from the population of equations is chosen at random and 'crossovers' are performed and progeny equations are generated; (iii) it is better at discovering combinations of features that take advantage of correlations between multiple features; (iv) the fitness of each progeny equation is assessed by LOF measure; (v) it can use a larger variety of equation term types in construction of its models; and (vi) if the fitness of new progeny equation is better, then it is preserved. The model with proper balance of all statistical terms is used to explain the variance of the response variable. A distinctive feature of GFA is that it produces a population of models (e.g. 100), instead of generating a single model, as in most other statistical methods. The range of variations in this population gives added information on the quality of fit and importance of the descriptors.

The G/PLS algorithm [48,49] may be used as an alternative to a GFA calculation. G/PLS is derived from two QSAR calculation methods: GFA and PLS. The G/PLS algorithm uses GFA to select appropriate basis functions to be used in a model of the data and PLS regression as the fitting technique to weigh the basis functions' relative contributions in the final model.

GFA and G/PLS were performed using QSAR+ environment of Cerius2 software [39]. The regression analyses and PLS analyses were carried out using SPSS [50] and MINITAB 14 [51], respectively.

2.4.2 *Statistical parameters*

The statistical quality of the equations [52] was judged by the parameters such as explained variance (R^2_α, i.e. adjusted R^2), correlation coefficient (r or R) and variance ratio (F) at specified degrees of freedom (df). For validation of the developed models, predicted residual sum of squares (PRESS; leave-one-out) statistics [53] were calculated and leave-one-out (LOO) cross-validation R^2 (Q^2), PRESS, were reported. All the accepted MLR equations have regression coefficients 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 SE of estimate for a particular equation. As the number of data points is limited, attempt was not made to split the data set into a training set and a test set, and to perform an external validation. However, randomisation test was applied to selected models.

3. Results and discussion

The best QSTR models selected on the basis of predicted variance Q^2 obtained from different statistical analysis tools (i.e. FA-MLR, stepwise regression, PCRA, PLS, GFA and G/PLS) using different types of descriptors [ETA, non-ETA (topological) and combined] are described below for comparison. The significance of the equations is also described. The 95% confidence intervals of the regression coefficients for MLR equations are mentioned within parentheses.

3.1 *QSTR models with ETA descriptors*

The best QSTR models developed from FA-MLR, stepwise regression, PCRA, PLS and GFA using ETA descriptors for acute NSAID cytotoxicity in rat hepatocytes[10] and selected on the basis of predicted variance Q^2 of the models are shown below.

For the development of QSTR model with FA-MLR technique, important ETA descriptors were selected using FA technique followed by which MLR analysis was performed. FA of the descriptor matrix along with the hepatotoxicity data could resolve seven factors explaining 98.1% of the data. The best model derived is shown below:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -12.944 + 14.423(\pm 10.389) \sum \beta'_{\mathrm{s}} + 0.178(\pm 0.183) \sum \alpha \\ n &= 15, Q^2 = 0.571, R^2 = 0.691, R = 0.831, \\ R^2_{\mathrm{a}} &= 0.639, \quad F = 13.40(\mathrm{df}2, 12), \mathrm{PRESS} = 2.989. \end{aligned} \tag{9}$$

Equation (9) represents a two variable equation with 57.1% predicted variance and 63.9% explained variance. The regression coefficient of the bulk parameter $\Sigma\alpha$ is significant at the 90% level. Both variables used in Equation (9) have positive contributions to the acute NSAID cytotoxicity in rat hepatocytes. This indicates that hepatotoxicity of NSAIDs increases with an increase in the number of heteroatoms and bulk. This is corroborated by the observation that compound 3 (niflumic acid)

containing three fluorine atoms, two nitrogen atoms and two oxygen atoms is one of the most toxic compounds of the series. Again, acetylsalicylic acid being larger in bulk than salicylic acid is more toxic.

When we performed stepwise regression analysis ($F = 4.0$ for inclusion; $F = 3.9$ for exclusion for the forward selection method) to develop QSTR model with the same data matrix, we got the following best equation:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -11.215 + 0.068(\pm 0.068) \\ &\sum\beta + 11.289(\pm 12.132)\sum\beta'_{\mathrm{s}} \\ &n = 15, Q^2 = 0.540, R^2 = 0.695, \\ &R = 0.834, R^2_{\mathrm{a}} = 0.644, \\ &F = 13.65(\mathrm{df}2, 12),\ \mathrm{PRESS} = 3.206. \end{aligned} \quad (10)$$

Equation (10) is a two-descriptor model with inferior prediction variance ($Q^2 = 0.540$) compared with Equation (9), but with superior explained variance ($R^2 = 0.644$). Both $\Sigma\beta$ and $\Sigma\beta'_{\mathrm{s}}$ parameters have positive contributions to the hepatotoxicity of NSAID compounds. Presence of the parameter $\Sigma\beta$ in Equation (10) suggests that both heteroatom count and the degree of unsaturation of the compounds are important for the hepatotoxicity. Moreover, $\sum\beta$ is obtained by summing up the VEM contributions for different bonds and it is not corrected for the vertex count N_{V}. Thus, $\sum\beta$ is correlated with the volume parameter $\sum\alpha$ ($r^2 = 0.886$).

When FA was done on the descriptor matrix, excluding the hepatotoxicity values, 96.0% variance was explained by six factors. Corresponding factor scores were used as the predictor variables to perform PCRA. PCRA has an advantage that collinearities among independent variables are not a disturbing factor and that the number of variables included in the analysis may exceed the number of observations. The best equation obtained for ETA descriptors using PCRA has the following statistical quality:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -3.057 + 0.361(\pm 0.272) \\ &f1 + 0.423(\pm 0.272)f2 \quad n = 15, Q^2 = 0.482, \\ &R^2 = 0.623, R = 0.789, R^2_{\mathrm{a}} = 0.560, \\ &F = 9.916(\mathrm{df}2, 12), \mathrm{PRESS} = 3.610. \end{aligned} \quad (11)$$

Though Equation (11) also has two variables similar to Equations (9) and (10), its quality is inferior to the latter equations.

In case of PLS analysis on the present data set, the variables with smaller coefficients were removed from the PLS regression, until there was no further improvement in Q^2 value, irrespective of the components. In case of PLS with ETA descriptors, the following best equation was obtained:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -9.207 + 7.235\sum\beta'_{\mathrm{s}} + 0.113 \\ &\sum\alpha + 0.041\sum\beta - 1.183[\eta'_{\mathrm{F}}]_{\mathrm{F}} \\ &n = 15, Q^2 = 0.640, R^2 = 0.725, R = 0.851, \\ &R^2_{\mathrm{a}} = 0.704, \quad F = 34.27(\mathrm{df}1, 13), \mathrm{PRESS} = 2.506. \end{aligned} \quad (12)$$

Equation (12) is based on one PLS component and four independent variables. Intercorrelation is not a disturbing factor in the case of PLS. The cross-validation statistics of Equation (12) is better than those of Equations (9) and (10). In Equation (12), VEM count $\sum\beta$ and contribution of sigma electrons to VEM count $\sum\beta'_{\mathrm{s}}$ have positive contributions to rat hepatotoxicity of NSAIDs. The bulk parameter $\Sigma\alpha$ in both Equations (9) and (12) has positive contributions to the acute NSAID cytotoxicity in rat hepatocytes. Furthermore, negative coefficient of $[\eta'_{\mathrm{F}}]_{\mathrm{F}}$ in Equation (12) indicates that the contribution of fluoro to the toxicity is less compared with other heteroatoms (like chloro) as predicted by the parameter $\sum\beta'_{\mathrm{s}}$.

Another technique GFA was used on the present data set using ETA descriptors to get models on the basis of an error measure, called the LOF score that automatically penalises models with too many features. The best model obtained from 5000 iterations is shown below:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -11.929 + 0.330(\pm 0.122) \\ &\sum\beta - 17.834(\pm 9.411)[\eta'_{\mathrm{F}}]_{\mathrm{COOH}} - 6.718(\pm 4.140) \\ &\sum\beta'_{\mathrm{ns}} \quad n = 15, \\ &Q^2 = 0.643, R^2 = 0.843, R = 0.918, \\ &R^2_{\mathrm{a}} = 0.800, \quad F = 19.62(\mathrm{df}3, 11), \\ &\mathrm{PRESS} = 2.484, \mathrm{LOF} = 0.203. \end{aligned} \quad (13)$$

Equation (13) containing three variables show 80% explained variance and 64.3% prediction variance both being superior to other models developed from the same descriptor matrix using other different statistical techniques. The positive coefficient of VEM count $\sum\beta$ and negative coefficients of non-sigma VEM component $\sum\beta'_{\mathrm{ns}}$ and $[\eta'_{\mathrm{F}}]_{\mathrm{COOH}}$ are found in Equation (13). As mentioned earlier, the VEM count $\sum\beta$ is obtained by summing up the VEM contribution for different bonds and is not corrected for the vertex count N_{V}. Hence, $\sum\beta$ is correlated with the volume parameter $\sum\alpha$ ($r^2 = 0.886$). Though increase of volume increases toxicity, insertion of fragments with unsaturation and carboxylic acid fragments have less impact in enhancing the toxicity.

G/PLS was applied to the ETA descriptor matrix to model the hepatotoxicity data at 1000 iterations using scaled variables and options such as no fixed length of the

equations and initial equation length set at 4. The best model is given below:

$$-\log \mathrm{LD}_{50} = -10.888 + 0.165\sum\alpha + 0.627\sum[\alpha]_{\mathrm{Y}}/\sum\alpha - 1.721[\eta'_{\mathrm{F}}]_{\mathrm{F}} + 10.531\sum\beta'_{\mathrm{s}}$$
$$n = 15, Q^2 = 0.606, R^2 = 0.739, R^2_{\mathrm{a}} = 0.719,$$
$$F = 36.82(\mathrm{df}1, 13), \mathrm{PRESS} = 2.744. \quad (14)$$

The G/PLS model (Equation (14)) is not superior to the GFA (Equation (13)) and PLS (Equation (12)) models. The positive coefficient of the parameter $\sum[\alpha]_{\mathrm{Y}}/\sum\alpha$ indicates positive contribution of branching to the toxicity.

3.2 QSTR models with Non-ETA descriptors

To search for better QSTR models, we have calculated the values of non-ETA topological descriptors and developed regression models using different tools. The best QSTR models developed from FA-MLR, stepwise regression, PCRA, PLS, GFA and G/PLS analyses using non-ETA descriptors selected on the basis of predicted variance Q^2 are shown below.

When FA was done on the descriptor matrix, including hepatotoxicity values, 96.8% variance was explained by six factors. The best QSTR model with non-ETA (topological) descriptors from FA-MLR is shown below:

$$-\log \mathrm{LD}_{50} = -5.318 + 0.227(\pm0.117)\mathrm{SC\_3\_C} + 0.243(\pm0.141)\mathrm{S\_ssNH} + 0.063(\pm0.057)\mathrm{S\_sOH}$$
$$n = 15, Q^2 = 0.652, R^2 = 0.871, R = 0.933,$$
$$R^2_{\mathrm{a}} = 0.836, \quad F = 24.826(\mathrm{df}3, 11), \mathrm{PRESS} = 2.422. \quad (15)$$

Equation (15) developed with one subgraph count index and two E-state variables shows 83.6% explained variance and 65.2% predicted variance of the hepatotoxicity data. The predicted variance of Equation (15) is tangibly better than that of the FA-MLR model (Equation (9)) developed from the ETA descriptors. The positive values of the coefficients of the variables indicate that the toxicity increases with the values of subgraph count (cluster type) of third-order and E-state values of the secondary amino (—NH—) and hydroxyl (—OH) fragments.

But when stepwise regression (stepping criteria, $F = 4$ for inclusion; $F = 3.9$ for exclusion) was performed with the non-ETA descriptor matrix, the same equation as derived from FA-MLR (Equation (15)) was developed.

FA of the descriptor matrix excluding the toxicity values extracted six factors explaining 95.1% of variance. Corresponding factor scores were considered as the predictor variables for the PCRA model which is shown below:

$$-\log \mathrm{LD}_{50} = -3.057 + 0.460(\pm0.225)f1 + 0.241(\pm0.225)f2 + 0.338(\pm0.225)f4$$
$$n = 15, Q^2 = 0.629, R^2 = 0.771, R = 0.878,$$
$$R^2_{\mathrm{a}} = 0.709, \quad F = 12.356(\mathrm{df}3, 11), \mathrm{PRESS} = 2.582. \quad (16)$$

The statistical quality in Equation (16) is superior to Equation (11) involving factor scores derived from non-ETA descriptor matrix, but inferior to other QSTR models developed with non-ETA topological descriptors using other statistical tools.

To get an equation of improved quality, we have performed PLS analysis and developed the following best QSTR model:

$$-\log \mathrm{LD}_{50} = -4.460 + 0.204{}^3\chi_{\mathrm{p}} + 0.594\mathrm{S\_sssCH} + 0.196\mathrm{S\_ssNH} + 0.016\mathrm{S\_sF}$$
$$n = 15, Q^2 = 0.796, R^2 = 0.873, R = 0.934,$$
$$R^2_{\mathrm{a}} = 0.863, \quad F = 89.73(\mathrm{df}1, 13), \mathrm{PRESS} = 1.417. \quad (17)$$

Equation (17) involving only one PLS component has been developed from one connectivity ($^3\chi_{\mathrm{p}}$) parameter and three E-state variables. The positive regression coefficients suggest that the toxicity increases with the values of third-order connectivity (path type) and E-state indices of fragments > CH—, —NH— and —F. The statistical quality of Equation (17) is superior to Equation (12) involving ETA parameters and also Equations (15) and (16) involving non-ETA parameters.

The best model developed for the current data set using non-ETA descriptors is obtained with GFA (5000 iterations) and the model is shown below:

$$-\log \mathrm{LD}_{50} = -4.852 + 1.585(\pm0.471)\mathrm{S\_sssCH} - 1.641(\pm1.003)^3\chi^{\mathrm{V}}_{\mathrm{C}} + 0.098(\pm0.002)\mathrm{SC\_3\_P}$$
$$n = 15, Q^2 = 0.871, R^2 = 0.924, R = 0.961, \quad (18)$$
$$R^2_{\mathrm{a}} = 0.903, \quad F = 44.52(\mathrm{df}3, 11),$$
$$\mathrm{PRESS} = 0.896, \mathrm{LOF} = 0.098.$$

Equation (18) involving three descriptors shows 87.1% predicted variance and 90.3% explained variance. This equation contains one descriptor each of E-state, connectivity and subgraph count categories.

G/PLS was applied to the non-ETA descriptor matrix to model the hepatotoxicity data at 1000 iterations using scaled variables and options such as no fixed length of the equations and initial equation length set at 4.

The best model is given below:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -4.734 + 0.221\mathrm{S\_ssNH} \\ &\quad + 0.777\mathrm{S\_sssCH} - 0.146\mathrm{S\_aasC} + 0.308\,{}^3\chi_{\mathrm{p}} \\ n &= 15, Q^2 = 0.809, R^2 = 0.905, R_{\mathrm{a}}^2 = 0.889, \\ F &= 57.18(\mathrm{df2}, 12), \mathrm{PRESS} = 1.326. \end{aligned} \tag{19}$$

Equation (19) with two PLS components is comparable in statistical quality with the PLS model Equation (17), but inferior to the GFA model Equation (18).

3.3 QSTR models with combined set of descriptors

While working with both ETA and non-ETA descriptors, models were derived with FA-MLR, stepwise regression, PCRA, PLS, GFA and G/PLS, and the best models are shown below.

The modelling of hepatotoxicity of NSAID compounds adopting FA-MLR method using combined set of descriptors led to the following best equation:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -5.486 + 1.258(\pm 0.547)\mathrm{S\_sssCH} \\ &\quad + 0.117(\pm 0.033)\sum\beta \quad n = 15, \\ Q^2 &= 0.792, \\ R^2 &= 0.867, R = 0.931, R_{\mathrm{a}}^2 = 0.845, \\ F &= 39.198(\mathrm{df2}, 12), \mathrm{PRESS} = 1.445. \end{aligned} \tag{20}$$

Equation (20) is a two-descriptor QSTR model with 79.2% predicted variance and 84.5% explained variance. In Equation (20), both ETA and non-ETA descriptors are included and predicted variance of Equation (20) is better than the FA-MLR models developed separately with ETA and non-ETA descriptors. Thus, on using ETA descriptors along with the non-ETA ones, predicted variance of the FA-MLR model developed with non-ETA indices increased. The positive coefficients of the variables S_sssCH and $\Sigma\beta$ indicate their positive contributions to the hepatotoxicity of NSAID compounds.

Next, another model-building attempt was made from the combined (ETA and non-ETA) descriptors using the stepwise regression method using the same stepping criteria as used in previous stepwise equations. The best model obtained with combined set of descriptors using stepwise regression analysis is identical to Equation (15) developed with non-ETA descriptors matrix for same data set using stepwise regression. Note that no ETA descriptors were selected in the stepwise process.

For PCRA, FA was performed using combined ETA and non-ETA descriptor matrix (omitting the toxicity values) and factor scores thus obtained were used as the predictor parameters in a multiple regression equation and the following best equation was obtained:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -3.057 + 0.434(\pm 0.227) \\ &\quad f1 + 0.334(\pm 0.227)f2 + 0.283(\pm 0.227)f4 \\ n &= 15, Q^2 = 0.632, R^2 = 0.765, R = 0.874, \\ R_{\mathrm{a}}^2 &= 0.700, \quad F = 11.903(\mathrm{df3}, 11), \mathrm{PRESS} = 2.562. \end{aligned} \tag{21}$$

The statistical quality of Equation (21) is inferior to that of Equation (20) developed from FA-MLR analysis with the same data matrix. However, on using ETA indices along with non-ETA ones, there has been an enhancement of statistical quality of the PCRA model (Equation (21) compared with Equation (16)).

In the case of PLS regression with the current data set, the best model is obtained with five variables and two PLS components and the same is shown below:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -6.308 + 0.024\mathrm{SC\_3\_P} + 0.118\,{}^3\chi_{\mathrm{p}} \\ &\quad + 0.548\,{}^3\chi_{\mathrm{C}} + 1.138\mathrm{S\_sssCH} - 1.776\sum\beta'_{\mathrm{ns}} \\ n &= 15, Q^2 = 0.834, R^2 = 0.896, R = 0.947, \\ R_{\mathrm{a}}^2 &= 0.879, \quad F = 51.61(\mathrm{df2}, 12), \mathrm{PRESS} = 1.159. \end{aligned} \tag{22}$$

Equation (22) consists of one subgraph count, two connectivity parameter, one E-state index and one ETA parameter. Except $\Sigma\beta'_{\mathrm{ns}}$, other indices with positive coefficients indicate their positive contributions to the hepatotoxicity of NSAID compounds. On the basis of prediction ability, Equation (22) ($Q^2 = 0.834$) is superior to Equation (17) ($Q^2 = 0.796$). This indicates that the presence of ETA descriptors has increased predicted variance of the non-ETA models

We have also performed GFA with the combined descriptor matrix, and in this case, the best model developed is same as Equation (18). In this case, no ETA parameters were selected.

G/PLS was applied to the combined descriptor matrix to model the hepatotoxicity data at 1000 iterations using scaled variables and options such as no fixed length of the equations and initial equation length set at 4. The best model is given below:

$$\begin{aligned} -\log \mathrm{LD}_{50} &= -6.741 + 2.811\sum\beta'_{\mathrm{ns}} \\ &\quad + 1.339\mathrm{S\_sssCH} + 11.197\sum[\alpha]_{\mathrm{X}}/\sum\alpha + 0.307\,{}^3\chi_{\mathrm{p}} \\ n &= 15, Q^2 = 0.854, R^2 = 0.919, R_{\mathrm{a}}^2 = 0.906, \\ F &= 67.92(\mathrm{df2}, 12), \mathrm{PRESS} = 1.018. \end{aligned} \tag{23}$$

The predicted variance of Equation (23) is superior to both Equations (14) and (19), i.e. the G/PLS models developed from ETA and non-ETA descriptors, respectively.

Table 3. Comparative statistical quality of different models.

Type of descriptors	ETA	Non-ETA	Combined
FA-MLR			
Q^2	0.571	0.652	0.792
R^2	0.691	0.871	0.867
R_a^2	0.639	0.836	0.845
Stepwise regression			
Q^2	0.540	0.652[b]	
R^2	0.695	0.871	
R_a^2	0.644	0.836	
PCRA			
Q^2	0.482	0.629	0.632
R^2	0.623	0.771	0.765
R_a^2	0.560	0.709	0.700
PLS			
Q^2	0.640	0.796	*0.834*[a]
R^2	0.725	0.873	0.896
R_a^2	0.704	0.863	0.879
GFA			
Q^2	0.643	*0.871*[a,b]	
R^2	0.843	0.924	
R_a^2	0.800	0.903	
G/PLS			
Q^2	0.606	0.809	*0.854*[a]
R^2	0.739	0.905	0.919
R_a^2	0.719	0.889	0.906

[a] Based on Q^2 values, the best three models are shown in italics.
[b] Same models are obtained for non-ETA and combined set of descriptors.

Thus, it is suggested that ETA descriptors supplement the information content of the non-ETA descriptors to generate better quality models. The statistical quality of Equation (23) is slightly inferior to the best GFA model Equation (18).

4. Overview of the models

A comparative study table (Table 3) has been given on the basis of the statistical qualities of the best models developed from different descriptor matrices (ETA, non-ETA and combined descriptors) using FA-MLR, stepwise regression, PCRA, PLS, GFA and G/PLS as the statistical tools. In the case of ETA descriptors, the best model (Equation (13)) ($Q^2 = 0.643$, $R^2 = 0.843$, $R_a^2 = 0.800$ and PRESS = 2.484) was obtained from GFA tool. For non-ETA descriptors, the best model (Equation (18); $Q^2 = 0.871$, $R^2 = 0.924$, $R_a^2 = 0.903$ and PRESS = 0.896) was also obtained from GFA technique. But in case of combined descriptors, two models of comparable statistical quality [GFA model, which is same as the one obtained with non-ETA descriptors, i.e. Equation (18) and G/PLS model (Equation (23); $Q^2 = 0.854$, $R^2 = 0.919$, $R_a^2 = 0.906$ and PRESS = 1.018] were obtained. Statistical quality of the non-ETA models using different statistical methods (except stepwise regression and GFA tools) increase on addition of ETA descriptors. Furthermore, statistical qualities of the best three models reported in the paper [Equation (18): $Q^2 = 0.871$, $R^2 = 0.924$; Equation (23): $Q^2 = 0.854$, $R^2 = 0.919$; Equation (22); $Q^2 = 0.834$, $R^2 = 0.896$] are comparable with that of the previously reported equations using physicochemical parameters [10].

Randomisation test was applied for model development process of the genetic models (results are shown in Table 4). The results show that the GFA models are superior to the G/PLS models as the mean values of the correlation coefficients of the random models for G/PLS are higher in comparison with the corresponding values

Table 4. Results of randomisation test of the model development process for the genetic models.

Eq. no	13	18	14	19	23
Model type	GFA	GFA	G/PLS	G/PLS	G/PLS
Descriptor type	ETA	Non-ETA	ETA	Non-ETA	Combined
r from non-random model	0.918	0.961	0.896	0.960	0.959
Confidence level	90%	90%	90%	90%	90%
Mean value of r from random trials ± SD	0.553* ± 0.162	0.536* ± 0.181	0.679 ± 0.078	0.772 ± 0.069	0.775 ± 0.105

[*] In case of the GFA models, mean values of R of random models are significantly lower than those of the corresponding non-random models

Table 5. Results of randomisation test applied to selected genetic models.

Eq. no	18	22	23
Model type	GFA	PLS	G/PLS
Descriptor type	Non-ETA	Combined	Combined
r from non-random model	0.961	0.947	0.959
No. of random r's less than non-random r	99	99	99
No. of random r's more than non-random r	0	0	0
Confidence level	99%	99%	99%
Mean value of r from random trials ± SD	0.424 ± 0.168	0.177 ± 0.265	0.109 ± 0.243

Table 6. Intercorrelation (r) matrix for the best model Equation (18).

	S_sssCH	$^3\chi_c^v$	SC_3_P
S_sssCH	1.000	0.231	−0.122
$^3\chi_c^v$	0.231	1.000	0.544
SC_3_P	−0.122	0.544	1.000

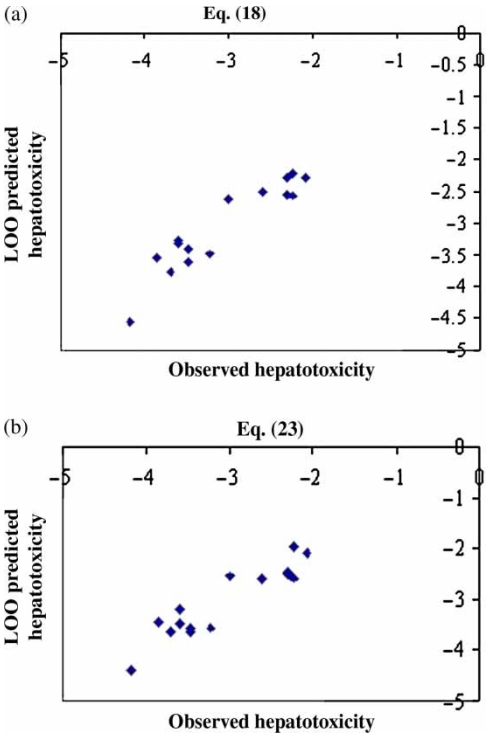

Figure 2. Scatter plots of observed versus LOO predicted hepatotoxicity of NSAIDs according to (a) Equation (18) and (b) Equation (23).

of GFA models, and hence the probability of obtaining the G/PLS models by chance is higher. Randomisation test was also applied to the best three models at 99% level and the results are shown in Table 5. Intercorrelation matrix for the best model Equation (18) is shown in Table 6, which indicates the absence of high intercorrelation among the predictor variables of Equation (18). Scatter plots of observed versus LOO predicted hepatotoxicity of NSAIDs according to Equations (18) and (23) are shown in Figure 2.

Use of ETA parameters suggests that the hepatotoxicity of NSAIDs increases with bulk and degree of branching. Moreover, the toxicity increases with heteroatom count and decreases with the degree of unsaturation.

5. Conclusion

Statistically significant QSAR relations were developed using topological parameters for hepatotoxicity of NSAIDs and statistical quality of the best model was comparable with that of the previously reported equations using physicochemical parameters [10]. The models developed with ETA descriptors alone were not statistically superior to the models developed from comparatively larger pool of non-ETA descriptors. However, the use of ETA descriptors along with the non-ETA ones increased the statistical quality of the non-ETA models in different techniques, except stepwise regression and GFA. This suggests that the ETA indices are sufficiently rich in chemical information to decode the structural features for QSAR/QSPR/QSTR modelling.

References

[1] D.L. DeWitt, *Cox-2-selective inhibitors: the new super aspirins*, Mol. Pharmacol. 55 (1999), pp. 625–631.
[2] P. Brooks, *Use and benefits of nonsteroidal anti-inflammatory drugs*, Am. J. Med. 104 (1998), pp. 9S–13S.
[3] M.W. James and C.J. Hawkey, *Assessment of non-steroidal anti-inflammatory drug (NSAID) damage in the human gastrointestinal tract*, Br. J. Clin. Pharmacol. 56 (2003), pp. 146–155.
[4] U.A. Boelsterli, H.J. Zimmerman, and A. Kretz-Rommel, *Idiosyncratic liver toxicity of nonsteroidal antiinflammatory drugs: molecular mechanisms and pathology*, Crit. Rev. Toxicol. 25 (1995), pp. 207–235.
[5] Y. Pirson and C. van Ypersele de Strihou, *Renal side effects of nonsteroidal antiinflammatory drugs: clinical relevance*, Am. J. Kidney Dis. 8 (1986), pp. 338–344.
[6] S.K. Kulkarni and N.K. Jain, *Coxibs: the new super aspirins or unsafe pain killer*, Indian J. Pharmacol. 37 (2005), pp. 86–89.
[7] K.D. Tripathi, *Essentials of Medical Pharmacology*, 5th ed., Jaypee Brothers Medical Publishers, New Delhi, 2003, pp. 167–184.
[8] J.L. Goldstein, P. Correa, W.W. Zhao, A.M. Burr, R.C. Hubbard, K.M. Verburg, and G.S. Geis, *Reduced incidence of gastrodudenal ulcers with celecoxib, a novel cycloxygenase-2 inhibitor, compared to naproxen in patients with arthritis*, Am. J. Gastroenterol. 96 (2001), pp. 1019–1027.
[9] G.H. Hakimelahi and G.A. Khodarahmi, *The identification of toxicophores for the prediction of mutagenicity, hepatotoxicity and cardiotoxicity*, J. Iranian Chem. Soc. 2 (2005), pp. 244–267.
[10] A.G. Siraki, T. Chevaldina, and P.J. O'Brien, *Application of quantitative structure-toxicity relationships for acute NSAID cytotoxicity in rat hepatocytes*, Chem.-Biol. Interact. 151 (2005), pp. 177–191.
[11] D.F. Lewis, C. Ioannides, and D.V. Parke, *A retrospective study of the molecular toxicology of benoxaprofen*, Toxicology 65 (1990), pp. 33–47.
[12] G.W. Milne, *Mathematics as a basis for chemistry*, J. Chem. Inf. Comput. Sci. 37 (1997), pp. 639–644.
[13] O. Ivanciuc and A.T. Balaban, *The graph description of chemical structures*, in *Topological Indices and Related Descriptors in QSAR and QSPR*, J. Devillers and A.T. Balaban, eds., Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 59–167.
[14] 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. 2 (2003), pp. 599–620, http://www.biochempress.com.
[15] K. Roy and G. Ghosh, *QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 559–567.
[16] K. Roy and G. Ghosh, *QSTR with extended topochemical atom indices. 3. Toxicity of nitrobenzenes to* Tetrahymena pyriformis, QSAR Comb. Sci. 23 (2004), pp. 99–108.
[17] K. Roy and G. Ghosh, *QSTR with extended topochemical atom indices. 4. Modeling of the acute toxicity of phenylsulfonyl carboxylates to* Vibrio fischeri *using principal component factor*

analysis and principal component regression analysis, QSAR Comb. Sci. 23 (2004), pp. 526–535.
[18] K. Roy and G. Ghosh, *QSTR with extended topochemical atom indices. Part 5. Modeling of the acute toxicity of phenylsulfonyl carboxylates to* Vibrio fischeri *using genetic function approximation*, Bioorg. Med. Chem. 13 (2005), pp. 1185–1194.
[19] K. Roy and G. Ghosh, *QSTR with extended topochemical atom* (*ETA*) *indices. VI. Acute toxicity of benzene derivatives to tadpoles* (Rana japonica), J. Mol. Model. 12 (2006), pp. 306–316.
[20] K. Roy and I. Sanyal, *QSTR with extended topochemical atom indices. 7. QSAR of substituted benzenes to* Saccharomyces cerevisiae, QSAR Comb. Sci. 25 (2006), pp. 359–371.
[21] K. Roy and G. Ghosh, *QSTR with extended topochemical atom (ETA) indices. 8. QSAR for the inhibition of substituted phenols on germination rate of* Cucumis sativus *using chemometric tools*, QSAR Comb. Sci. 25 (2006), pp. 846–859.
[22] K. Roy and G. Ghosh, *QSTR with extended topochemical atom (ETA) indices. 9. Comparative QSAR for the toxicity of diverse functional organic compounds to* Chlorella vulgaris *using chemometric tools*, Chemosphere 70 (2007), pp. 1–12.
[23] K. Roy, I. Sanyal, and P.P. Roy, *QSPR of the bioconcentration factors of nonionic organic compounds in fish using extended topochemical atom (ETA) indices*, SAR QSAR Environ. Res. 17 (2006), pp. 563–582.
[24] K. Roy, I. Sanyal, and G. Ghosh, *QSPR of* n-*octanol/water partition coefficient of nonionic organic compounds using extended topochemical atom (ETA) indices*, QSAR Comb. Sci. 25 (2006), pp. 629–646.
[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. 27B (1988), pp. 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. 28B (1989), pp. 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. 29B (1990), pp. 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. 31B (1992), pp. 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. 38B (1999), pp. 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. 40B (2001), pp. 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. 9 (2003), pp. 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. 2 (2003), pp. 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. 2 (2003), pp. 475–491, http://www.biochempress.com.
[34] K. Roy, S. Chakroborty, C.C. Ghosh, and A. Saha, *QSPR with TAU indices: molar thermochemical properties of diverse functional acyclic compounds*, J. Indian Chem. Soc. 81 (2004), pp. 115–125.
[35] K. Roy and A. Saha, *QSPR with TAU indices: boiling points of sulfides and thiols*, Indian J. Chem. 43A (2004), pp. 1369–1376.
[36] K. Roy and A. Saha, *QSPR with TAU indices: molar refractivity of diverse functional acyclic compounds*, Indian J. Chem. 44B (2005), pp. 1693–1707.
[37] K. Roy and A. Saha, *QSPR with TAU indices: part 5. Liquid heat capacity of diverse functional organic compounds*, J. Indian Chem. Soc. 83 (2006), pp. 351–355.
[38] The GW-BASIC programs RRR98, KRETA1, KRETA2, KRPRES1 and KRPRES2 were developed by Kunal Roy and standardized using known data sets.
[39] Cerius 2 version 4.10 is a product of Accelrys Inc., San Diego, CA, USA, http://www.accelrys.com/cerius2
[40] P.J. Lewi, *Multivariate data analysis in structure activity relationships*, in *Drug Design*, E.J. Ariens, ed.,Vol. 10, Academic Press, New York, 1980, pp. 307–342.
[41] R. Franke and A. Gruska, *Principal component and factor analysis*, in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, ed.,Vol. 2, VCH, Weinheim, 1995, pp. 113–163.
[42] R.B. Darlington, *Regression and Linear Models*, McGraw-Hill, New York, 1990.
[43] S. Wold, *PLS for multivariate linear modeling*, in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, ed., VCH, Weinheim, 1995, pp. 195–218.
[44] G.M. Sperandio da Silva, C.M. Sant'Anna, and E. Barreiro, *A novel 3D-QSAR comparative molecular field analysis (CoMFA) model of imidazole and quinazolinone functionalized p38 MAP kinase inhibitors*, Bioorg. Med. Chem. 12 (2004), pp. 3159–3166.
[45] S.S. Kulkarni and V.M. Kulkarni, *Three-dimensional quantitative structure-activity relationship of interleukin 1-beta converting enzyme inhibitors: a comparative molecular field analysis study*, J. Med. Chem. 42 (1999), pp. 373–380.
[46] Y. Fan, L.M. Shi, K.W. Kohn, Y. Pommier, and J.N. Weinstein, *Quantitative structure-antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies*, J. Med. Chem. 44 (2001), pp. 3254–3263.
[47] D. Rogers and A.J. Hopfinger, *Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure-property relationships*, J. Chem. Inf. Comput. Sci. 34 (1994), pp. 854–866.
[48] W.J. Dunn III and D. Rogers, *Genetic partial least squares in QSAR*, in *Genetic Algorithms in Molecular Modeling*, J. Devillers, ed., Academic Press, London, 1996, pp. 109–130.
[49] K. Hasegawa, Y. Miyashita, and K. Funatsu, *GA strategy for variable selection in QSAR studies: GA based PLS analysis of calcium channel antagonists*, J. Chem. Inf. Comput. Sci. 37 (1997), pp. 306–310.
[50] SPSS is statistical software of SPSS Inc., USA.
[51] MINITAB is a statistical software of MINITAB Inc., USA.
[52] G.W. Snedecor and W.G. Cochran, *Statistical Methods*, Oxford and IBH Publishing Co. Pvt. Ltd, New Delhi, 1967, pp. 381–418.
[53] A.K. Debnath, *Quantitative structure-activity relationship (QSAR): a versatile tool in drug design*, in *Combinatorial Library Design and Evaluation*, A.K. Ghose and V.N. Viswanadhan, eds., Marcel Dekker, Inc., New York, 2001, pp. 73–129.