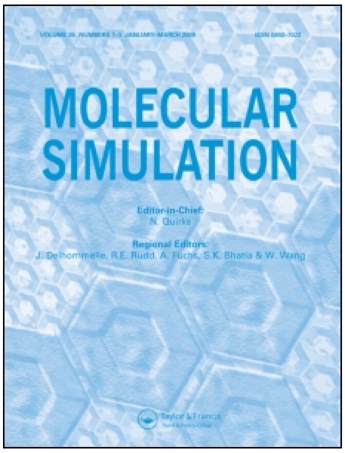

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

QSTR with extended topochemical atom (ETA) indices. 13. Modelling of hERG K^+ channel blocking activity of diverse functional drugs using different chemometric tools

Kunal Roy[a]; Gopinath Ghosh[a]
[a] Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Laboratory, 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. 13. Modelling of hERG K^+ channel blocking activity of diverse functional drugs using different chemometric tools', Molecular Simulation, 35: 15, 1256 — 1268

To link to this Article: DOI: 10.1080/08927020903015379

URL: http://dx.doi.org/10.1080/08927020903015379

Molecular Simulation
Vol. 35, No. 15, December 2009, 1256–1268

Taylor & Francis
Taylor & Francis Group

QSTR with extended topochemical atom (ETA) indices. 13. Modelling of hERG K^+ channel blocking activity of diverse functional drugs using different chemometric tools

Kunal Roy* and Gopinath Ghosh

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

(*Received 21 March 2009; final version received 22 April 2009*)

To accelerate the drug discovery process, early prediction of human ether a-go-go (hERG) K^+ channel affinity of drug candidates is becoming an important aspect. We have therefore developed quantitative structure–toxicity relationship models with extended topochemical atom (ETA) indices for hERG K^+ channel blocking activity of diverse functional drugs using different chemometric tools like factor analysis followed by multiple linear regression (FA-MLR), stepwise regression and partial least squares. The data set was divided into a training set of 50 compounds and a test set of 17 compounds based on *K*-means clustering technique. The ETA models were compared with those developed with a pool of other topological indices. Finally, an attempt was made to develop models from the combined pool of topological (ETA and non-ETA) descriptors. It was found that on using ETA parameters along with non-ETA ones, there was a considerable increase in the quality of the models. The best model came from stepwise regression using a combined set of descriptors ($Q^2 = 0.546$, $R^2_{\text{pred}} = 0.619$). The ETA model suggests that hERG channel blocking increases with the increase of molecular bulk and electron richness and decreases with the increase of functionalities of the carboxylic acid group and the aliphatic tertiary nitrogen fragment.

Keywords: QSAR; QSTR; hERG; ETA; TAU; PLS; stepwise regression; FA-MLR

1. Introduction

All drugs are foreign particles to the body and are capable of producing adverse effects along with their therapeutic effects. Thus, whenever a drug is consumed, a risk is taken [1]. An effect of a drug may be therapeutic in one context but a side effect in another context, e.g. depression of A-V conduction is the desired effect of digoxin in articular fibrillation, but the same may be undesirable when it is used for congestive heart failure [1]. Moreover, the post-marketing scenario of many pharmaceuticals used as therapeutic agents led to many unfortunate human experiences in the past, e.g. thalidomide with phocomelia, diethylstilbestrol with vaginal carcinoma and tetracyclines with discoloured and deformed teeth and also retardation of bone growth. These observations have raised many questions on the availability of sufficient and adequate information on the toxic effects of existing drugs. This is essential in order to protect human health and also for regulatory purposes [1,2–5]. Continuous research on the approved marketed drugs with specific therapeutic activity unveils their severe toxic effects on the hepatic, cardiovascular and nervous system. The high attrition rates of clinical candidates with undesirable absorption, distribution, metabolism, elimination and toxicity (ADMET) properties and/or also drug-induced liver injury, carcinogenicity, immuno-suppression and/or cardiac liability associated with the blockade of human ether a-go-go (hERG) K^+ channel, are the most frequent reasons for the withdrawal of some approved drugs from the market [5–7] and also a major cause of delay in the drug discovery process. The hERG-related gene potassium channels conduct the rapid component of the delayed rectifier potassium current, I_{Kr}, which is crucial for repolarisation of the cardiac action potentials. Moderate hERG blockade may produce a beneficial class III antiarrhythmic effect. However, a reduction in the hERG currents either due to genetic defects or adverse drug effects can lead to hereditary or acquired long QT syndromes characterised by action potential prolongation, lengthening of the QT interval on the surface ECG, and an increased risk for 'torsade de pointes' arrhythmias and sudden death. This undesirable side effect of the non-antiarrhythmic compounds has prompted the withdrawal of several blockbuster drugs from the market [8–11]. A research on the non-antiarrhythmic drugs with diversified structures used for different specific therapeutic activities shows cardiovascular toxicity associated with the undesirable blockade of hERG K^+ channel [11]. Thorough experimental studies on the mechanisms of hERG channel inhibition provide significant insights into the molecular factors that determine state-, voltage- and use-dependency of hERG K^+ current block [11]. A novel potassium channel gene

*Corresponding author. Email: kunalroy_in@yahoo.com

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020903015379
http://www.informaworld.com

associated with cardiac arrhythmia has been cloned and characterised. The gene encodes MinK-related peptide 1 (MiRP1), a small integral membrane subunit that assembles with hERG, a pore-forming protein, to alter its function. It has been proved that increased current and hereditary short QT syndrome with a high risk for life-threatening arrhythmias is produced due to mutations in K^+ channels with the hERG-related gene [10]. Mutants form channels with three missense mutations associated with long QT syndrome and ventricular fibrillations identified in the gene for MiRP1, which open slowly and close rapidly, thereby diminishing potassium currents. One variant, associated with clarithromycin-induced arrhythmia, increases channel blockade by an antibiotic [10]. The increased understanding of the hERG K^+ channel function and molecular mechanisms of the hERG K^+ current regulation could improve prevention and treatment of hERG-associated cardiac disorders and prevent human death due to drug induced cardiac abnormality [8]. Crucial properties of the high-affinity drug binding site in the hERG K^+ channel and its interaction with diverse drug molecules have been identified, providing the basis for more refined approaches in drug design, safety pharmacology and *in silico* modelling [8]. To accelerate the drug discovery process, early prediction of hERG K^+ channel affinity of drug candidates is becoming an important aspect. To gain insight into the molecular basis of drug–hERG K^+ channel interaction, both structure- and ligand-based approaches have been undertaken [11–14]. The medical needs of patients suffering from various disorders remain unmet and no current therapy constitutes a cure for these excluding chemical therapy. A great amount of concerted effort is being made worldwide to find a safe and an acceptable drug with considerably lower hERG K^+ channel blocking activity. The pharmaceutical industry has endeavoured to discover, develop and market efficacious drugs with reduced side effects so that a disease-specific treatment can be continued. This can be accelerated with the help of *in silico* tools like quantitative structure-activity relationships (QSARs) with the application of modern chemometric tools, as an alternative non-animal rational approach for the toxicity evaluation of the drugs [15–20]. An early identification of the adverse effects triggered by drugs and hypothetical drug candidates with *in silico* approach would be highly desirable in connection with safety and economy as well as a variety of ecological benefits such as sustainable resource management, reduction of animal models and possibly less risky clinical trials [21]. In *in silico* drug discovery, both existing and hypothetical drugs are now studied with the help of sophisticated software packages with fast and reproducible results, and typically based on human bioregulators [21]. The reliability and accuracy of toxicity predictions may be achieved by identifying toxicophores with the help of statistically robust QSAR models. These predictions can guide the design of chemical libraries for a 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. Several *in silico* approaches of considerable computational and statistical modelling efforts to define the features related with molecular structure have been attempted to unveil the hERG K^+ channel blockade property of approved marketed drug molecules or pre-clinical drug candidates or hypothetical drug candidates [11]. The primary aim of these approaches is to filter out potential hERG K^+ channel blockers in the context of virtual libraries; others involve an understanding of the structure–activity relationships governing hERG–drug interactions [12]. Different research groups have developed *in silico* models for hERG K^+ channel blockers. Different software modules such as DISCO in Sybyl [22,23], HypoGen in Catalyst [24,25] and Binary QSAR in MOE [26] have been successfully employed for the development of models for hERG K^+ channel blockers. Aptula and Cronin [27] have developed a QSAR model with a good statistical fit to predict the hERG K^+ channel blocking potency using structural knowledge. Du et al. [28] have introduced pharmacophore hypotheses of I(Kr) K^+ channel blocker as novel class III antiarrhythmic agents. Kramer et al. [29] have developed a composite model for hERG blockade without crystal structures of the hERG-encoded channel to save time and money and to differentiate between specific and non-specific drug binding. Binary classification models based on a combination of support vector machine method and using pharmacophore-based GRIND descriptor for large diverse compounds have been developed by Li et al. [30] in order to discriminate between the hERG blockers and non-blockers. A pharmacophore model in Catalyst, consisting of four hydrophobic features and four positive ionisable features has been developed by Ekins et al. using a training set of 15 mol [26]. A pharmacophore model using Catalyst 4.9 consisting of two aromatic rings, one hydrophobic group and one positive ionisable group along with the geometrical distances has been developed by Du et al. [24] from 34 mol. Cavilli et al. [31] have constructed a pharmacophore made up of three aromatic moieties connected through a nitrogen function (tertiary amine) and developed a QSAR model using 31 hERG channel blockers.

In connection with the rational approach of modelling of activity and toxicity of drugs, chemical graph theory and topological descriptors have been extensively applied. Topological descriptors are derived from hydrogen-suppressed 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) [32,33]. In consequence, the TIs can be defined as 2D descriptors that can be easily calculated from the molecular graphs, and do not depend on the way the graph is depicted or labelled and there is no need of energy minimisation of the chemical structure. They offer a simple way of measuring molecular branching, shape and size [34], which is used to develop the QSAR models to predict biological activity or toxicity of the existing and hypothetical drug molecules. In this background, we have introduced extended topochemical atom (ETA) indices [35–46] as an extension of the topochemically arrived unique (TAU) concept in the valence electron mobile (VEM) environment [47–50], and developed quantitative structure–toxicity relationship (QSTR) models with different toxicity data (phenol toxicity to *Tetrahymena pyriformis* [35], fish toxicity of substituted benzenes [36], nitrobenzene toxicity to *Tetrahymena pyriformis* [37], acute toxicity of phenylsulphonyl carboxylates to *V. fischeri* [38,39], acute toxicity of benzene derivatives to tadpoles (*Rana japonica*) [40], acute toxicity of benzene derivatives to *Saccharomyces cerevisiae* [41], inhibition of substituted phenols on the germination rate of *Cucumis sativus* [42], toxicity of 91 organic compounds to *Chlorella vulgaris* [43]) to establish the utility of ETA indices in modelling studies on chemical toxicity. We have also developed QSTR models with pharmaceuticals used as therapeutic agents such as human lethal concentration values of 26 organic compounds including some pharmaceuticals [44], acute NSAID cytotoxicity in rat hepatocytes [45] using the ETA indices and established that the ETA indices have sufficient power to encode the structural features of drugs to develop QSTR models using different chemometric tools.

In this present communication, we have developed QSTR models with ETA indices [35–46] for hERG K^+ channel blocking activity of diverse functional drugs [5] using different chemometric tools. The QSTR models developed with ETA descriptors have been compared with those developed from selected non-ETA (topological) and combined set of descriptors to show the modelling power and predictive nature of ETA indices. As the ETA indices are derived from a chemical graph theory approach, we have compared the ETA models with those derived from other topological indices that are not derived from an experiment or 3D approaches. Different statistical tools used in this communication are stepwise regression analysis, multiple linear regression with factor analysis as the pre-processing step for variable selection (FA-MLR) and partial least squares (PLS) regression.

2. Materials and methods

2.1 Data set

The hERG K^+ channel blocking activity data (pIC_{50}) of 67 diverse functional drugs (Table 1) [5] have been used as the model data set for the present work.

2.2 Descriptors

2.2.1 ETA descriptors

Definitions of some basic parameters used in the ETA [35,36] scheme to develop QSTR models with the current data set are given below.

The core count of a non-hydrogen vertex (α) is defined as [35,36]:

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

In Equation (1), PN stands for the period number while Z and Z^v represent the atomic number and valence electron number, respectively. As the hydrogen atom is being considered as the reference, the α value for hydrogen is taken to be zero. Again, another term ε, as a measure of electronegativity, has been defined [27,28] in the following manner:

$$\varepsilon = -\alpha + 0.3Z^V. \tag{2}$$

It is interesting to note that the α values of different atoms (which are commonly found in organic compounds) have high correlation ($r = 0.946$) [35] with (uncorrected) van der Waals volume while ε has a good correlation ($r = 0.937$) with Pauling's electronegativity scale (EN).

The VEM count β of the ETA scheme is defined as [35,36]:

$$\beta = \Sigma x\sigma + \Sigma y\pi + \delta. \tag{3}$$

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

Table 1. Observed and calculated values of the hERG K$^+$ channel blocking activity data (pIC$_{50}$).

Sl. no.	Molecules	pIC$_{50}$ [obs] [5]	pIC$_{50}$ [calc^a]	pIC$_{50}$ [calc^b]
Training set				
3	Astemizole	2.046	1.423	1.484
6	Chloroquine	−0.398	−0.311	−0.080
7	Chlorpheniramine	−0.204	−0.535	−0.626
8	Chlorpromazine	−0.167	−0.286	−0.359
9	Ciprofloxacin	−2.985	−2.326	−2.213
11	Clozapine	0.495	−0.960	−0.900
12	Clozapine-*N*-oxide	−2.125	−1.091	−1.323
13	Desipramine	−0.143	−0.490	−0.384
14	Diltiazem	−1.238	−0.163	−0.311
15	Diphenhydramine	−1.477	−0.368	−0.551
16	Dofetilide	1.921	1.524	1.875
17	Dolasetron	−0.775	0.205	0.108
18	Domperidone	0.791	0.891	1.209
19	Droperidol	1.495	0.862	0.926
20	E-4031	2.097	0.882	0.972
23	Granisetron	−0.572	−0.131	0.036
25	Halofantrine	0.706	−0.253	−0.277
27	Imipramine	−0.531	−0.408	−0.575
28	Lidocaine	−2.42	−1.559	−1.383
30	Lumefantrine	−0.91	0.374	0.290
31	Mesoridazine	0.26	0.032	−0.068
32	Mibefradil	−0.155	0.839	0.657
33	Mizolastine	0.456	0.459	0.372
34	Mefloquine	−0.422	−1.084	−0.916
35	Moxifloxacin	−2.111	−1.971	−1.815
36	*N*-desmethylclozapine	−0.652	−1.026	−0.690
37	Ofloxacin	−3.152	−2.412	−2.536
38	Olanzapine	−0.779	−0.963	−0.912
39	Ondansetron	0.091	0.102	0.061
40	Perhexiline	−0.892	−1.235	−0.990
43	Pimozide	1.745	1.229	1.228
44	Propafenone	0.356	0.362	0.455
46	Queitiapine	−0.761	−0.380	−0.733
47	Quinidine	0.397	−0.391	−0.499
48	Risperidone	0.777	0.601	0.523
49	Sertindole	1.854	0.714	0.800
50	Sparfloxacin	−1.255	−2.367	−2.224
51	Spironolactone	−1.362	−0.885	−0.953
52	Terfenadine	1.252	0.881	0.579
53	Thioridazine	1.447	−0.029	−0.148
54	Vardenafil	−1.107	−0.338	−0.266
55	Verapamil	0.845	0.034	−0.065
56	Ziprasidone	0.772	1.070	1.254
57	Acrivastine	−0.900	−1.439	−1.668
58	Amsacrine	0.690	0.622	0.949
60	Desmethylastemizole	0.000	1.396	1.469
62	Fentanyl	−0.260	0.401	0.176
64	Ketoconazole	−0.300	1.216	1.050
66	Meperidine	−1.870	−1.434	−1.555
67	Tadalafil	1.000	0.224	0.104
Test set				
1	Amiodarone	0.154	0.182	0.224
2	Amitryptyline	−1.000	−0.504	−0.654
4	Azimilide	0.252	1.057	0.920
5	Bepridil	0.260	0.244	0.021
10	Cisapride	2.155	0.784	0.950
21	Flecainide	−0.592	−0.150	0.261
22	Gatifloxacin	−2.114	−2.343	−2.203
24	Grepafloxacin	−1.699	−2.407	−2.283
26	Haloperidol	0.553	0.584	0.484
29	Loratadine	0.762	−0.059	−0.172

Table 1 – *continued*

Sl. no.	Molecules	pIC_{50} [obs] [5]	pIC_{50} [calc[a]]	pIC_{50} [calc[b]]
41	Phenobarbital	− 0.477	− 1.422	− 1.162
42	Phenytoin	− 2.380	− 0.597	− 0.328
45	Pyrilamine	− 0.041	− 0.158	− 0.309
59	Cocaine	− 0.640	− 0.582	− 0.703
61	Desmethylcarboxyloratadine	− 0.800	− 0.493	− 0.286
63	Fexofenadine	− 1.330	− 0.312	− 0.674
65	Lidoflazine	1.790	1.442	1.473

[a] Calculated from Equation (14).
[b] Calculated from Equation (17).

The VEM vertex count γ_i of the ith vertex in a molecular graph is defined as [35,36]:

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

In the above equation, α_i stands for the α-value of 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).

Finally, the composite index η is defined in the following manner [35,36]:

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

In Equation (5), both bonded and non-bonded interactions have been considered. The term r_{ij} stands for the topological distance between the ith atom and jth atom. Again, when all heteroatoms and multiple bonds in the molecular graph are replaced by carbon and single bond, respectively, the 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 have defined another term η'_F as η_F/N_V [35,36]. Again, one can determine contribution of a particular position, vertex or substructure to functionality in the following manner:

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

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 the dependence of this value on N_V, a related term $[\eta'_F]_i$ was the defined [35,36] as $[\eta_F]_i/N_V$.

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

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

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

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

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

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

In the present communication, utility of the ETA parameters has been demonstrated through a QSTR study taking the hERG K^+ channel blocking activity data (pIC_{50}) of 67 diverse functional drugs [5] as the model data set.

The calculations of η, η_R, η_F, η_B and the contributions of different vertices to η_F were done, using distance matrix and VEM vertex counts as inputs, by the GW-BASIC programs KRETA1 and KRETA2 developed by one of the authors [51]. The definitions of a few important ETA parameters are shown in Table 2.

2.2.2 *Non-ETA descriptors*

Performance of the ETA models was checked from the models developed from other topological (non-ETA) indices. The values for the non-ETA topological descriptors for the compounds have been calculated by QSAR+ and Descriptor+ modules of the Cerius 2

Table 2. Definitions of important ETA parameters used in exploring QSAR with the hERG K^+ channel blockers using chemometric tools.

Parameter	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 another vertex of the molecule
$[\eta'_F]$	Total functionality
$[\eta'_F]_{Cl}$	Functionality contribution for the chlorine atom
$[\eta'_F]_F$	Functionality contribution for the fluorine atom
$[\eta'_F]_{CH3}$	Functionality for the methyl group
$[\eta'_F]_{OCH3}$	Functionality for the methoxy group
$[\eta'_F]_{NH2}$	Functionality for the amino group
$[\eta'_F]_{OH}$	Functionality for the hydroxyl group
$[\eta'_F]_{COOH}$	Functionality for the carboxyl group
$[\eta'_F]_{-C{=}O}$	Functionality for the carbonyl group in ketones
$[\eta'_F]_{altN}$	Functionality for the aliphatic *tert* nitrogen atom
$[\eta'_F]_{alsN}$	Functionality for the aliphatic *sec* nitrogen atom
$[\eta'_F]_{al-N=}$	Functionality for the aliphatic nitrogen atom
$\Sigma\beta'_s$	$\Sigma\beta'_s$ is defined as $[\Sigma\beta_s]/N_v$ for non-hydrogen substituent(s); in case hydrogen is present in the substituent position, the value for that position is taken as zero
$\Sigma\beta'_{ns}$	$\Sigma\beta'_{ns}$ is defined as $[\Sigma\beta_{ns}]/N_v$ for non-hydrogen substituent(s); in case hydrogen is present in the substituent position, the value for that position is taken as zero
$[\eta'_B]_{unc}$	Branching without cyclicity correction

version 4.10 software [52]. The various topological indices calculated are Wiener W, Zagreb, connectivity indices (${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi_p$, ${}^3\chi_c$, ${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi_p^v$, ${}^3\chi_c^v$), kappa shape indices (${}^1\kappa$, ${}^2\kappa$, ${}^3\kappa$, ${}^1\kappa_\alpha$, ${}^2\kappa_\alpha$, ${}^3\kappa_\alpha$) and E-state parameters (S_sCH$_3$, S_ssCH$_2$, S_aaCH, S_dssC, S_aasC, S_aaaC, S_sNH$_2$, S_ddsN, S_sOH, S_dO, S_ssO, S_sCl, S_sBr). Finally, an attempt was made to develop models using combined set of descriptors (ETA and non-ETA). As the ETA class of descriptors represents invariants obtained from the graph-theoretic approach, we have not considered physicochemical or 3D descriptors for the comparison of model quality.

2.3 Model development

2.3.1 Statistical analyses performed

The statistical analyses were carried out using MLR and PLS as the statistical tools. Different methods of variable selection such as stepwise regression and factor analysis (FA) were used for the MLR analysis.

2.3.1.1 Factor analysis followed by multiple linear regression. In the case of FA-MLR, a classical approach of the multiple regression technique was used as the final statistical tool for developing QSTR relations and FA [53,54] was used as the data pre-processing step to identify the important predictor variables contributing to the response variable and to avoid collinearities among them. In a typical FA procedure, the data matrix is first standardised, and a correlation matrix and subsequently the reduced correlation matrix are constructed. An eigenvalue problem is then solved and the factor pattern can be obtained from the corresponding eigenvectors. The principal objectives of FA are to display multidimensional data in a space of lower dimensionality with minimum loss of information (explaining $>95\%$ of the variance of the data matrix) and to extract the basic features behind the data with the ultimate goal of interpretation and/or prediction.

FA was performed on the data set containing the activity data (pIC$_{50}$) of the training set of hERG K^+ channel blockers [5] (*vide infra*) and all descriptor variables, which were to be considered. The factors were extracted by principal component method and then rotated by VARIMAX rotation (a kind of rotation that is used in principal component analysis so that the axes are rotated to a position in which the sum of the variances of the loadings is the maximum possible) to obtain Thurston's simple structure. The simple structure is characterised by the

Table 3. Cluster membership for the compounds of the training and test sets.

Data set	Training set							Test set			
Cluster no.	1	2		3	4			1	2	3	4
No. of compounds in each cluster	10	18		5	17			3	6	2	6
Sl. no. of compounds	16	6	27	3	14	51		1	2	63	5
	18	7	28	32	19	53		4	41	65	21
	25	8	36	43	20	56		10	42		22
	30	9	38	52	31	57			45		24
	33	11	39	64	34	58			59		26
	48	12	40		35	62			61		29
	49	13	47		37	67					
	54	15	66		44						
	55	17			46						
	60	23			50						

property that as many variables as possible fall on the coordinate axes when presented in a common factor space, so that the largest possible number of factor loadings becomes zero. This is done to obtain a numerically comprehensive picture of the relatedness of the variables. Only variables with non-zero loadings in such factors, where the toxicity also has a non-zero loading were considered important in explaining variance of the toxicity. Further, variables with non-zero loadings in different factors were combined in a multivariate equation.

2.3.1.2 Stepwise regression. In stepwise regression [55], a multiple-term linear equation was built step-by-step. The basic procedures involve: (1) identifying an initial model; (2) iteratively 'stepping,' that is, 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 will then be included in the model, and the process starts again.

2.3.1.3 Partial least squares. PLS is a generalisation of regression, which can handle data with strongly correlated and/or noisy or numerous independent variables [56]. It gives a reduced solution that is statistically more robust than MLR. The linear PLS model finds 'new variables' (latent variables or independent scores) that 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 [57]. Application of PLS thus allows the construction of larger QSTR 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 QSTR equations are selected based on their ability to predict the data rather than to fit the data [58]. Based on the standardised regression coefficients, the variables with smaller coefficients were removed from the PLS regression, until there is no further improvement in the Q^2 value, irrespective of the components. The regression analyses and PLS analyses were carried out using SPSS software [59] and MINITAB 14 [60], respectively.

2.3.2 *Statistical parameters*

The statistical quality of the equations [61] was judged by parameters such as explained variance (R_a^2, i.e. adjusted R^2), correlation coefficient (r or R) and variance ratio (F) at specified degrees of freedom (*df*). PRESS (leave-one-out) statistics [56,62] were calculated and leave-one-out cross-validation R^2 (Q^2), predicted residual sum of squares (PRESS) were reported as internal validation parameters. 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 SE of estimate for a particular equation. The final selection of equation was done on the basis of external validation.

2.3.3 *External validation*

Any QSAR modelling should ultimately lead to statistically robust models capable of making accurate and reliable predictions of the biological activities of compounds. When QSAR models are developed, it is important to validate any fitted models to check whether it is plausible that its predictions will carry over to fresh data not used in the model fitting exercise. External validation has been considered more reliable to judge the prediction potential of QSAR models than internal validation techniques [43,63,64]. For maximum cases, appropriate external data set is not available for prediction purposes. That is why the original data set is divided into training and test sets. The selection of the training set is significantly important in QSAR analysis. In this paper, to check the predictability of ETA indices using external validation, the current data set was divided into a training set containing 50 hERG K^+ channel blockers and a test set containing 17 hERG K^+ channel blockers [5] (i.e. 75 and 25%, respectively, of the total number of compounds) based on clusters obtained from K-means clustering technique [65] applied on standardised descriptor matrix (Table 3). The (external) predictive capacity of a model was judged by its application for the prediction of activity of the test-set compounds and calculation of the predictive R^2 (R^2_{pred}) value for the test set as shown below:

$$R^2_{\text{pred}} = 1 - \frac{\Sigma\left(Y_{\text{pred(Test)}} - Y_{\text{obs(Test)}}\right)^2}{\Sigma\left(Y_{\text{obs(Test)}} - \bar{Y}_{\text{Training}}\right)^2}.$$

In the above equation, $Y_{\text{pred(Test)}}$ and $Y_{\text{obs(Test)}}$ indicate the predicted and observed activity values, respectively, of the test-set compounds and $\bar{Y}_{\text{Training}}$ indicates the mean activity value of the training set.

3. Results and discussion

The best QSTR models developed using hERG K^+channel blocking activity data (pIC_{50}) of the training set compounds [5] with different statistical analysis tools (i.e. FA-MLR, stepwise regression and PLS) and different types of topological descriptors [ETA, non-ETA (topological) and combined] and selected on the basis of R^2, R_a^2 and Q^2 values of the equations are described below for comparison. The final selection of the best equation was done on the basis of R^2_{pred} (predictive R^2) value for the test-set compounds. The significance of the equations is also described. The 95% confidence intervals of the regression coefficients for regression equations are mentioned within parentheses.

3.1 QSTR models with FA-MLR

The best QSTR models developed from FA-MLR using ETA, non-ETA (topological) and combined set of descriptors and selected on the basis of the statistical superiority of the equations are shown below.

3.1.1 QSTR with ETA descriptors

For the development of a QSTR model with FA-MLR technique, important ETA descriptors were selected using FA technique followed by which the MLR analysis was an performed. FA of the ETA descriptor matrix along with the activity (pIC_{50}) of the training set could resolve the data matrix into 16 factors explaining 95.67% of the variance. The best model derived using ETA descriptors is shown below:

$$\begin{aligned} pIC_{50} = {} & -15.150 + 2.046(\pm 0.981)\Sigma\alpha - 0.071(\pm 0.037) \\ & [\Sigma\alpha]^2 + 2.201(\pm 1.740)\Sigma\beta'_{\text{ns}} - 8.835(\pm 2.952) \\ & [\eta'_{\text{F}}]_{\text{COOH}} - 10.665(\pm 6.854)[\eta'_{\text{F}}]_{\text{al=N}-}, \\ & n_{\text{Training}} = 50, \quad Q^2 = 0.538, \\ & R^2 = 0.656, \quad R = 0.810, \quad R_a^2 = 0.617 \\ & F = 16.818(df\,5,\,44), \quad \text{PRESS} = 36.923, \\ & n_{\text{Test}} = 17, \quad R^2_{\text{Pred}} = 0.499. \end{aligned} \tag{9}$$

Equation (9) having five descriptors could explain and predict 61.7 and 53.8%, respectively, of the variance of activity (pIC_{50}) of the training set. The predictive R^2 value for the test set is 0.499. Equation (9) has a parabolic relation between variable $\Sigma\alpha$ and the activity (pIC_{50}). The positive coefficient of the variable $\Sigma\alpha$ indicates that the toxicity increases with the bulk of the compounds. However, the optimum $\Sigma\alpha$ value is found to be 14.4 beyond which the activity decreases. Again, the functionalities of COOH and aliphatic tertiary nitrogen atom show negative contributions to the hERG channel blocking activity (pIC_{50}), whereas the term $\Sigma\beta'_{\text{ns}}$ (electron richness) with a positive coefficient indicates its positive contribution.

3.1.2 QSTR with non-ETA descriptors

For the development of a QSTR model with FA-MLR technique, important non-ETA (topological) descriptors were selected using FA technique followed by which MLR analysis was performed. FA of the non-ETA descriptor matrix along with the activity (pIC_{50}) of the training set could resolve the data matrix into 17 factors explaining 95.68% of the variance. The best model developed with non-ETA descriptors and selected on the basis of the R^2_{pred} value (predictive R^2) for the test-set compounds is shown below:

$$\begin{aligned} pIC_{50} = {} & -1.607 + 0.337(\pm 0.170)^3\kappa - 5.388 \\ & (\pm 4.291)^3\chi_{\text{CH}} \quad - 0.115(\pm 0.108)\text{S\_sCH}_3 \\ & + 0.331(\pm 0.277)\text{S\_aaaC}, \quad n_{\text{Training}} = 50, \\ & Q^2 = 0.308, \quad R^2 = 0.442, \quad R = 0.664, \\ & R_a^2 = 0.392, \quad F = 8.895(df\,4,\,45), \\ & \text{PRESS} = 55.329, \quad n_{\text{Test}} = 17, \quad R^2_{\text{Pred}} = 0.364. \end{aligned} \tag{10}$$

Equation (10) having four non-ETA descriptors could explain and predict only 39.2 and 30.8%, respectively, of the variance of activity (pIC_{50}) of the training set of hERG K^+ channel blockers. On the basis of Equation (10), the predictive R^2 value for the test set is 0.364, which is drastically lower than that of Equation (9). Thus, using FA-MLR, the non-ETA topological descriptors give an inferior model than that with ETA descriptors.

3.1.3 QSTR with combined set of descriptors

When ETA descriptors were combined with the non-ETA descriptors for the development of QSTR model using FA-MLR technique, important ETA and non-ETA (topological) descriptors were selected using FA technique after which MLR analysis was performed. FA of the combined descriptor matrix along with the activity (pIC_{50}) of the training set could resolve the data matrix into 18 factors explaining 95.25% of the variance. The best model derived on the basis of good statistical quality is shown below:

$$\mathrm{pIC}_{50} = -16.418 + 0.165(\pm 0.162)^{3}\kappa + 2.037(\pm 0.976)\Sigma\alpha - 0.073(\pm 0.037)[\Sigma\alpha]^{2} + 1.803(\pm 1.560)\Sigma\beta' - 8.779(\pm 2.899)[\eta_{\mathrm{F}}^{'}]_{\mathrm{COOH}} - 9.905(\pm 6.947)[\eta_{\mathrm{F}}^{'}]_{\mathrm{al=N-}},\ n_{\mathrm{Training}} = 50,\ Q^{2} = 0.564,\ R^{2} = 0.677,\ R = 0.823,\ R_{\mathrm{a}}^{2} = 0.632,\ \mathrm{F} = 14.997(df\,6, 43),\ \mathrm{PRESS} = 34.897,\ n_{\mathrm{Test}} = 17,\ R_{\mathrm{Pred}}^{2} = 0.507. \quad (11)$$

Equation (11) with five ETA and one non-ETA descriptors could explain and predict 63.2 and 56.4%, respectively, of the variance of the activity (pIC_{50}) of the training-set compounds. The predictive R^2 value for the test set is 0.507. Like Equation (9), Equation (11) also has a parabolic relation between the variable $\Sigma\alpha$ and activity (pIC_{50}) of the hERG K^{+} channel blockers (optimum $\Sigma\alpha$ being 13.2).

On the basis of explained and predicted variance (internal and external) values, Equation (11) is superior to both the Equations (9) and (10). On using the ETA parameters along with the non-ETA parameters, there has been a considerable increase in the quality of the models.

3.2 QSTR models with stepwise regression

In search of robust QSTR models for the current data set, stepwise regression analysis was performed for the data matrices of the ETA, non-ETA (topological) and combined (ETA and non-ETA) set of descriptors along with the activity (pIC_{50}) as the response variable to select the important independent variables for regression analysis. The statistical quality of the best models has been judged on the basis of R^2, R_a^2 and Q^2 values of the equations. Finally, the best equation has been selected on the basis of R^2_{pred} (predictive R^2) value for the test-set compounds. The best models developed from stepwise regression analyses are shown below.

3.2.1 QSTR with ETA descriptors

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

$$\mathrm{pIC}_{50} = -15.164 + 2.223(\pm 1.028)\Sigma\alpha - 0.077(\pm 0.039)[\Sigma\alpha]^{2} - 9.394(\pm 3.090)[\eta_{\mathrm{F}}^{'}]_{\mathrm{COOH}} - 9.765(\pm 7.217)[\eta_{\mathrm{F}}^{'}]_{\mathrm{al=N-}},\ n_{\mathrm{Training}} = 50,\ Q^{2} = 0.517, \quad (12)\ R^{2} = 0.606,\ R = 778,\ R_{\mathrm{a}}^{2} = 0.571,\ \mathrm{F} = 17.286(df\,4,\ 45),\ \mathrm{PRESS} = 38.643,\ n_{\mathrm{Test}} = 17,\ R_{\mathrm{Pred}}^{2} = 0.508.$$

Equation (12) having 57.1% explained variance and 51.7% predictive variance contains four ETA-independent variables. The 95% confidence intervals of the regression coefficients are shown within parentheses. The predictive R^2 value of Equation (12) for the test set is 0.508. Like Equation (9), Equation (12) also contains the volume parameter $\Sigma\alpha$ with parabolic relationship (with optimum value of 14.4) and the functionalities of COOH and aliphatic tertiary nitrogen atom with negative contributions to the activity. On the basis of predictive power for the test-set compounds, Equation (12) is superior to Equation (9).

3.2.2 QSTR with non-ETA descriptors

With the non-ETA descriptors matrix and the same stepping criteria (F-to-enter 4 and F-to-remove 3.9), the following best equation on the basis of statistical quality was obtained from stepwise regression analysis.

$$\mathrm{pIC}_{50} = -3.595 + 0.246(\pm 0.144)^{1}\chi^{v} + 0.090(\pm 0.042)\mathrm{S\_aaCH} - 0.084(\pm 0.054)\mathrm{S\_sOH} + 0.023(\pm 0.021)\mathrm{S\_sF},\ n_{\mathrm{Training}} = 50,\ Q^{2} = 0.395,\ R^{2} = 0.515,\ R = 0.718,\ R_{\mathrm{a}}^{2} = 0.472,\ F = 11.967(df\,4,\ 45), \quad (13)\ \mathrm{PRESS} = 48.390,\ n_{\mathrm{Test}} = 17,\ R_{\mathrm{Pred}}^{2} = 0.415.$$

Equation (13) containing four non-ETA descriptors shows an equation with inferior statistics quality (explained variance 47.2% and predicted variance 39.5%) compared with Equation (12) developed with ETA descriptors. The predictive R^2 value of Equation (13) for the test-set compounds is 0.415, which is also inferior to that of Equation (12).

3.2.3 QSTR with combined set of descriptors

Combined descriptor matrix was used to develop a QSTR model using stepwise regression maintaining the same stepping criteria. The best equation developed and selected on the basis of superior statistical quality is shown below.

$$\begin{aligned} \mathrm{pIC}_{50} &= 0.343 - 1.460(\pm 0.713)JX + 0.178 \\ &(\pm 0.140)^{3}\kappa \ + 2.278(\pm 1.782)\Sigma\beta^{'}_{\mathrm{ns}} - 6.546 \\ &(\pm 2.955)[\eta^{'}_{\mathrm{F}}]_{\mathrm{COOH}} - 7.385(\pm 7.120)[\eta^{'}_{\mathrm{F}}]_{\mathrm{al{=}N-}}, \\ &n_{\mathrm{Training}} = 50,\ Q^{2} = 0.546,\ R^{2} = 0.640, \\ &R = 0.800,\ R^{2}_{\mathrm{a}} = 0.599,\ F = 15.663(df\,5,\,44), \\ &\mathrm{PRESS} = 36.342,\ n_{\mathrm{Test}} = 17,\ R^{2}_{\mathrm{Pred}} = 0.619. \end{aligned} \tag{14}$$

Equation (14) having 59.9% explained variance and 54.6% predicted variance contains three ETA descriptors and two non-ETA descriptors. The predictive R^2 value for the test-set compounds is 0.619. External validation shows that Equation (14) is superior to the other equations developed from FA-MLR and stepwise regression analysis.

3.3 QSTR models with PLS

In search of robust QSTR models for the current data set, we have further performed PLS analysis. The number of components (latent variables) was selected based on internal validation. The statistical quality of the best models has been justified on the basis of Q^2 values of the equations. Finally, the best equation was selected on the basis of R^2_{pred} (predictive R^2) for the test-set compounds. The best QSTR models developed for PLS analysis are shown below.

3.3.1 QSTR with ETA descriptors

In the 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 the case of PLS with ETA descriptors, the following best equation was obtained:

$$\begin{aligned} \mathrm{pIC}_{50} &= -12.177 + 1.453\Sigma\alpha - 0.049[\Sigma\alpha]^{2} \\ &\quad + 3.706\Sigma\beta^{'}_{\mathrm{ns}} - 6.927[\eta^{'}_{\mathrm{F}}]_{\mathrm{COOH}} - 11.309 \\ &[\eta^{'}_{\mathrm{F}}]_{\mathrm{al{=}N-}}, n_{\mathrm{Training}} = 50, Q^{2} = 0.506, \\ &R^{2} = 0.619, R = 0.787, R^{2}_{\mathrm{a}} = 0.585, \\ &F = 18.28(df\,4,\,45), \mathrm{PRESS} = 39.537, \\ &n_{\mathrm{Test}} = 17, R^{2}_{\mathrm{Pred}} = 0.465. \end{aligned} \tag{15}$$

Equation (15) is based on four PLS components and five ETA variables. Intercorrelation is not a disturbing factor in the case of PLS. The cross-validation statistic of Equation (15) is inferior to those of Equations (9) and (12). The predictive R^2 value of Equation (15) for the test-set compounds is 0.465. Like the other two equations (stepwise MLR and FA-MLR) developed with ETA descriptors, Equation (15) shows a parabolic relation between variable $\Sigma\alpha$ and the activity (optimum $\Sigma\alpha$ being 14.8).

3.3.2 QSTR with non-ETA descriptors

When PLS regression was performed with non-ETA topological descriptors for the present data set, the best model with eight non-ETA descriptors and one PLS component (selected by cross-validation) was obtained and the model is shown below.

$$\begin{aligned} \mathrm{pIC}_{50} &= -0.111 - 1.01JX + 0.183^{3}\kappa_{\alpha} - 3.775^{3} \\ &\chi_{\mathrm{CH}} + 0.052\mathrm{S\_aaCH} + 0.208\mathrm{S\_aaaC} + 0.052 \\ &\mathrm{S\_ssNH} - 0.148\mathrm{S\_ddssS} + 0.001\mathrm{S\_sF}, \\ &n_{\mathrm{Training}} = 50, Q^{2} = 0.476, R^{2} = 0.559, \\ &R = 0.748, R^{2}_{\mathrm{a}} = 0.550,\ F = 60.74(df\,1,\,48), \\ &\mathrm{PRESS} = 41.865,\ n_{\mathrm{Test}} = 17, R^{2}_{\mathrm{Pred}} = 0.344. \end{aligned} \tag{16}$$

On the basis of R^2, R^2_{a} and Q^2 values of the equations obtained from the training-set compounds and predictive R^2 for the test-set compounds, Equation (16) is inferior to Equation (15) and other equations developed with the same descriptor matrix using FA-MLR and stepwise regression analysis.

3.3.3 QSTR with combined set of descriptors

When a combined set of descriptors was used for PLS analysis, the following equation was obtained.

$$\begin{aligned} \mathrm{pIC}_{50} &= 0.141 - 1.449JX + 0.207^{3}\kappa_{\alpha} - 0.011 \\ &\mathrm{S\_aaCH}\ + 0.081\mathrm{S\_ssNH} + 2.701\Sigma\beta^{'}_{\mathrm{ns}} - 6.856 \\ &[\eta^{'}_{\mathrm{F}}]_{\mathrm{COOH}} - 9.238[\eta^{'}_{\mathrm{F}}]_{\mathrm{al{=}N-}},\ n_{\mathrm{Training}} = 50, \\ &Q^{2} = 0.540,\ R^{2} = 0.658,\ R = 0.811,\ R^{2}_{\mathrm{a}} = 0.619, \\ &F = 16.92(df\,5,\,44),\ \mathrm{PRESS} = 36.751,\ n_{\mathrm{Test}} = 17, \\ &R^{2}_{\mathrm{Pred}} = 0.617. \end{aligned} \tag{17}$$

Table 4. Comparison of statistical quality of different models.

	FA-MLR				Stepwise				PLS			
	Training			Test	Training			Test	Training			Test
Type of descriptors	Q^2	R^2	R^2_{a}	R^2_{pred}	Q^2	R^2	R^2_{a}	R^2_{pred}	Q^2	R^2	R^2_{a}	R^2_{pred}
ETA	0.538	0.656	0.617	0.499	0.517	0.606	0.571	0.508	0.506	0.619	0.585	0.465
Non-ETA	0.308	0.442	0.392	0.364	0.395	0.515	0.472	0.415	0.476	0.559	0.550	0.344
Combined	0.564	0.677	0.632	0.507	**0.546**	**0.640**	**0.599**	**0.619**	**0.540**	**0.658**	**0.619**	**0.617**

The best two models are shown in bold face.

Equation (17) containing three ETA and four non-ETA descriptors can predict 54.0% and explain 61.9% of the variance of the activity of hERG K^+ channel blockers. On the basis of the explained variance values, Equation (17) is superior to Equation (14), but on the basis of predictive variances (internal and external), Equation (17) is inferior to Equation (14).

4. Overview and conclusion

The models developed from different chemometric tools for the current data set using the ETA parameters show that volume parameter ($\Sigma\alpha$) and electron richness (unsaturation) have positive contributions to the hERG K^+ channel blocking activity but functionalities of COOH and aliphatic tertiary nitrogen fragments have negative contributions. From the models involving ETA parameters, it is also observed that there is a parabolic relation between the volume parameter ($\Sigma\alpha$) and the activity and the optimum value of $\Sigma\alpha$ is near 14 beyond which the activity decreases.

Comparisons of key statistical terms like R^2, R^2_{a}, Q^2 and R^2_{pred} of different models obtained by using different statistical tools and different descriptors have been shown in Table 4. In the case of ETA descriptors, the best model (Equation (12)) is obtained from stepwise regression analysis considering the predictive ability for the test-set compounds. Again, on the basis of the values of R^2, R^2_{a}, Q^2 parameters, the best model (Equation (16)) for non-ETA descriptors is obtained from PLS analysis. But when the value of R^2_{pred} for the test-set compounds is considered, the best model (Equation (13)) for non-ETA descriptors is obtained from stepwise regression analysis. These two models obtained from non-ETA descriptors are comparable in statistical quality. When we considered both ETA and non-ETA descriptors in combination, the statistical quality of the models improved. Considering predictive ability and statistical quality of the models form the combined descriptor matrix, the two best models (Equations (14) and (17)) are obtained from stepwise regression analysis and PLS analysis, respectively. The calculated activity values for the training- and test-set compounds on the basis of Equations (14) and (17) are shown in Table 1. As the best models were obtained for the current data set from the combined set of descriptors, it can be inferred that ETA descriptors increase the quality and predictive ability of the non-ETA models. As the ETA class of descriptors represents invariants obtained from the graph-theoretic approach, we have not considered physicochemical descriptors for a comparison of model quality.

Finally, we can conclude that the ETA descriptors are sufficiently rich in chemical information to encode the important structural features for hERG channel blockers and these may be used in combination with other topological descriptors for the development of predictive QSTR models.

References

[1] K.D. Tripathi, *Essential Medical Pharmacology*, 4th ed., Jaypee Brothers, New Delhi, 1999, pp. 69–70.

[2] J. Zurlo, D. Rudacille, and A.M. Goldberg, *Animal and Alternatives in Testing: History, Science, and Ethics*, Mary Ann Liebert, Inc., New Rochelle, NY, 1994.

[3] M.I. Greenberg and S.D. Phillips, *A Brief History of Occupational, Industrial and Environmental Toxicology*, in *Occupational, Industrial and Environmental Toxicology*, M.I. Greenberg, R.J. Hamilton, and S.D. Phillips, eds., Mosby, Philadelphia, 2003, pp. 2–5.

[4] G. Langley, *Acute Toxicity Testing without Animals "More Scientific and Less of a Gamble,"* European Coalition to End Animal Experiments (ECEAE), London, 2005.

[5] D. Garg, T. Gandhi, and C.G. Mohan, *Exploring QSTR and toxicophore of hERG K+ channel blockers using GFA and HypoGen techniques*, J. Mol. Graph. Mod. 26 (2008), pp. 966–976.

[6] 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.

[7] 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.

[8] D. Thomas, C.A. Karle, and J. Kiehn, *The cardiac hERG/IKr potassium channel as pharmacological target: structure, function, regulation, and clinical applications*, Curr. Pharm. Des. 12(18) (2006), pp. 2271–2283.

[9] X.L. Zhao, Z.P. Qi, C. Fang, M.H. Chen, Y.J. Lv, B.X. Li, and B.F. Yang, *HERG K+ channel blockade by the novel antiviral drug sophocarpine*, Biol. Pharm. Bull. 31(4) (2008), pp. 627–632.

[10] G.W. Abbott, F. Sesti, I. Splawski, M.E. Buck, M.H. Lehmann, K.W. Timothy, M.T. Keating, and S.A. Goldstein, *MiRP1 forms IKr potassium channels with HERG and is associated with cardiac arrhythmia*, Cell 97(2) (1999), pp. 175–187.

[11] K.M. Thai and G.F. Ecker, *Predictive models for HERG channel blockers: ligand-based and structure-based approaches*, Curr. Med. Chem. 14(28) (2007), pp. 3003–3026.

[12] A.M. Aronov, *Predictive* in silico *modeling for hERG channel blockers*, Drug Discov. Today 10(2) (2005), pp. 149–155.

[13] A.M. Aronov, *Ligand structural aspects of hERG channel blockade*, Curr. Top. Med. Chem. 8(13) (2008), pp. 1113–1127.

[14] M.C. Sanguinetti and M. Tristani-Firouzi, *hERG potassium channels and cardiac arrhythmia*, Nature 440(7083) (2006), pp. 463–469.

[15] H. González-Díaz, F. Prado-Prado, and F.M. Ubeira, *Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach*, Curr. Top. Med. Chem. 8 (2008), pp. 1676–1690.

[16] A. Duardo-Sánchez, G. Patlewicz, and A. López-Díaz, *Current topics on software use in medicinal chemistry: intellectual property, taxes, and regulatory issues*, Curr. Top. Med. Chem. 8 (2008), pp. 1666–1675.

[17] A.M. Helguera, R.D. Combes, M.P. González, and M.N. Cordeiro, *Applications of 2D descriptors in drug design: a DRAGON tale*, Curr. Top. Med. Chem. 8 (2008), pp. 1628–1655.

[18] M.P. González, C. Terán, L. Saíz-Urra, and M. Teijeira, *Variable selection methods in QSAR: an overview*, Curr. Top. Med. Chem. 8 (2008), pp. 1606–1627.

[19] J. Caballero and M. Fernández, *Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularized genetic neural networks (BRGNN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-1)*, Curr. Top. Med. Chem. 8 (2008), pp. 1580–1605.

[20] S. Vilar, G. Cozza, and S. Moro, *Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery*, Curr. Top. Med. Chem. 8 (2008), pp. 1555–1572.

[21] A. Vedani, M. Dobler, and M.A. Lill, *The challenge of predicting drug toxicity* in silico, Basic Clin. Pharmacol. Toxicol. 99(3) (2006), pp. 195–208.

[22] P. Matyus, A.P. Borosy, A. Varro, J.G. Papp, D. Barlocco, and G. Cignarella, *Development of pharmacophores for inhibitors of the rapid component of the cardiac delayed rectifier potassium current*, Int. J. Quant. Chem. 69 (1998), pp. 21–30.

[23] A.P. Borosy, K. Keseru, I. Penzes, and P. Matyus, *3D QSAR study of class I antiarrhythmics*, J. Mol. Struct. 503 (2000), pp. 113–129.

[24] L.P. Du, K.C. Tsai, M.Y. Li, Q.D. You, and L. Xia, *The pharmacophore hypotheses of I(Kr) potassium channel blockers: novel class III antiarrhythmic agents*, Bioorg. Med. Chem. Lett. 14 (2004), pp. 4771–4777.

[25] S. Ekins, W.J. Crumb, R.D. Sarazan, J.H. Wikel, and S.A. Wrighton, *Three dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channels*, J. Pharmacol. Exp. Ther. 301 (2002), pp. 427–434.

[26] A.M. Aronov, *Common pharmacophores for uncharged human ether-ago-go-related gene (hERG) blockers*, J. Med. Chem. 49 (2006), pp. 6917–6921.

[27] A.O. Aptula and M.T. Cronin, *Prediction of hERG K+ blocking potency: application of structural knowledge*, SAR QSAR Env. Res. 15(5–6) (2004), pp. 399–411.

[28] L.P. Du, K.C. Tsai, M.Y. Li, Q.D. You, and L. Xia, *The pharmacophore hypotheses of I(Kr) potassium channel blockers: novel class III antiarrhythmic agents*, Bioorg. Med. Chem. Lett. 14(18) (2004), pp. 4771–4777.

[29] C. Kramer, B. Beck, J.M. Kriegl, and T. Clark, *A composite model for hERG Blockade*, ChemMedChem. 3(2) (2008), pp. 254–265.

[30] Q. Li, F.S. Jørgensen, T. Oprea, S. Brunak, and O. Taboureau, *hERG classification model based on a combination of support vector machine method and GRIND descriptors*, Mol. Pharm. 5(1) (2008), pp. 117–127.

[31] A. Cavalli, E. Poluzzi, F. De Ponti, and M. Recanatini, *Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K(+) channel blockers*, J. Med. Chem. 45 (2002), pp. 3844–3853.

[32] H. González-Díaz, S. Vilar, L. Santana, and E. Uriarte, *Medicinal chemistry and bioinformatics-current trends in drugs discovery with networks topological indices*, Curr. Top. Med. Chem. 7 (2007), pp. 1015–1029.

[33] H. González-Díaz, Y. González-Díaz, L. Santana, F.M. Ubeira, and E. Uriarte, *Proteomics, networks and connectivity indices*, Proteomics 8 (2008), pp. 750–778.

[34] 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.

[35] K. Roy and G. Ghosh, *Introduction of extended topochemical atoms (ETA) indices in the valence electron mobile (VEM) environment as tool for QSAR/QSPR studies*, Internet Elect. J. Mol. Des. 2 (2003), pp. 599–620. Available at http://www.biochempress.com.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] K. Roy and G. Ghosh, *QSTR with extended topochemical atom indices. 10. Modeling of toxicity of organic chemicals to humans using different chemometric tools*, Chem. Biol. Drug Des. 72 (2008), pp. 383–394.

[45] K. Roy and G. Ghosh, *QSTR with extended topochemical atom (ETA) indices. 11. Comparative QSAR of acute NSAID cytotoxicity in rat hepatocytes using chemometric Tools*, Mol. Simul. 35 (2009), pp. 648–659.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] The GW-BASIC programs RRR98, KRETA1, KRETA2, KPRES1 and KPRES2 were developed by Kunal Roy and standardized using known data sets, 1998.

[52] Cerius 2 version 4.10 is a product of Accelrys Inc., San Diego, CA, USA. Available at http://www.accelrys.com/cerius2, 2005.

[53] F. Franke, *Theoretical Drug Design Methods*, Elsevier, Amsterdam, 1984, pp. 184–195.

[54] R. Franke and P.A. Gruska, *Principal component and factor analysis*, in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, ed., VCH, Weinheim, 1995, pp. 113–163.

[55] R.B. Darlington, *Regression and Linear Models*, McGraw-Hill, New York, NY, 1990.

[56] S. Wold and L. Eriksson, *Statistical validation of QSAR results*, in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, ed., VCH, Weinheim, 1995, pp. 312–317.

[57] G.M. Sperandio da Silva, C.M. SantLAnna, and E.J. 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.

[58] 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.

[59] SPSS is statistical software of SPSS, Inc., USA, 1998.

[60] MINITAB is a statistical software of MINITAB, Inc., USA.

[61] G.W. Snedecor and W.G. Cochran, *Statistical Methods*, Oxford & IBH Publishing Co. Pvt. Ltd, New Delhi, 1967, pp. 381–418.

[62] 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, NY, 2001, pp. 73–129.

[63] J.T. Leonard and K. Roy, *On selection of training and test sets for the development of predictive QSAR models*, QSAR Comb. Sci. 25 (2006), pp. 235–251.

[64] P. Roy and K. Roy, *On some aspects of variable selection for partial least squares regression models*, QSAR Comb. Sci. 27 (2008), pp. 302–313.

[65] B.S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, Edward Arnold, London, 2001.