SAR and QSAR in Environmental Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gsar20

QSTR modelling of the acute toxicity of pharmaceuticals to fish

G. Tugcu [a], M. Türker Saçan [a], M. Vracko [b], M. Novic [b] & N. Minovski [b]

[a] Bogazici University, Institute of Environmental Sciences, Istanbul, Turkey

[b] Kemmijski Institute/National Institute of Chemistry, Ljubljana, Slovenia

Available online: 02 Mar 2012

To cite this article: G. Tugcu, M. Türker Saçan, M. Vracko, M. Novic & N. Minovski (2012): QSTR modelling of the acute toxicity of pharmaceuticals to fish, SAR and QSAR in Environmental Research, DOI:10.1080/1062936X.2012.657678

To link to this article: http://dx.doi.org/10.1080/1062936X.2012.657678

SAR and QSAR in Environmental Research
2012, 1–14, iFirst

QSTR modelling of the acute toxicity of pharmaceuticals to fish[$£]

G. Tugcu[a], M. Türker Saçan[a]*, M. Vracko[b], M. Novic[b] and N. Minovski[b]

[a]*Bogazici University, Institute of Environmental Sciences, Istanbul, Turkey;*
[b]*Kemmijski Institute/National Institute of Chemistry, Ljubljana, Slovenia*

(*Received 1 September 2011; in final form 4 November 2011*)

Extensive use of pharmaceuticals as human and veterinary medication raises concerns for their adverse effects on non-target organisms. The purpose of this study was to employ multiple linear regression (MLR) to predict the toxicities of a diverse set of pharmaceuticals to fish. The descriptor pool consisted of about 1500 descriptors calculated using Dragon 5.4, Spartan 06 and Codessa 2.2 software. Descriptor selection was made by the heuristic method available in Codessa 2.2. The data set was divided into training and test sets using Kohonen networks. The training set contained approximately 65% of the compounds of the full data set (99 compounds). The training set model contained eight descriptors from all dimensions, all of which were obtained from Dragon 5.4. The statistical parameters of the model for the training set are $R^2 = 0.664$, $F = 13.588$, and R^2_{cv} (LOO) = 0.542 while it achieves $R^2 = 0.605$ for the test set. The training, test and external sets have no response outliers considering the standardized residual greater than three. The external validation of the model was made with a set of pharmaceuticals obtained from several databases. The R^2_{pred} is 0.777, reflecting a relatively good predictive power for the external set.

Keywords: pharmaceuticals; QSAR; QSTR; toxicity; fish; MLR

1. Introduction

Concerns have been raised about pharmaceuticals and their metabolites because of their extensive and increasing usage [1] and ubiquitous presence in the aquatic environment [2,3]. Although the current concentrations in both surface waters and effluents are low, their possible adverse effects on aquatic life and ultimately on human health are not well understood [4].

Determination of the toxic effects of the pharmaceuticals on aquatic life can be elicited by either ecotoxicity testing or by applicable models. The experimental studies on pharmaceuticals summarized by Lange and Dietrich [5] are in general very limited relative to the studies on industrial chemicals [6]. Qualitative structure–activity relationships (QSARs) have been used successfully for many years for modelling purposes [7,8]. In terms of pharmaceutical toxicity, the models in the literature mainly focus on ranking and categorization of these compounds [9,10].

*Corresponding author. Email: msacan@boun.edu.tr
[$]Dedicated to the memory of Professor Corwin H. Hansch (1918–2011).
[£]Presented at CMTPI 2011: Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (Maribor, Slovenia, 3–7 September 2011).

ISSN 1062–936X print/ISSN 1029–046X online

http://dx.doi.org/10.1080/1062936X.2012.657678
http://www.tandfonline.com

In the literature, there are studies on the general risk assessment, mode of action (MOA) estimation, and importance of chronic toxicity estimations of pharmaceuticals. Crane et al. [11] reviewed the chronic aquatic toxicity of pharmaceuticals. They stated that there are QSAR models for prioritization purposes using acute toxicity data, but there are not sufficient studies for chronic toxicity of pharmaceuticals. The authors suggested that the solution is in testing the acute and chronic toxicity of a representative range of model substances on a representative range of aquatic organisms. In an extensive survey by Khetan and Collins [12], the background of problems related to the release of pharmaceuticals to the environment and uncertainties were discussed. Calculation of risk quotient, which is used in environmental risk assessments, was depicted as a ratio of predicted environmental concentration to predicted no-effect concentration. They summarized the benefits of the risk assessments approaches. Sanderson and Thomsen [2] examined modes of action of active pharmaceutical ingredients (APIs). They found that majority of the acute MOA of the studied APIs were non-specific narcosis. They additionally stated that the size and the conformation of the toxicants are crucial for the expression of the compound's potential excess toxicity. However, Fent et al. [1] reported that particular pharmaceuticals may have additional MOAs. Escher et al. [13] focused on hospital wastewater as a primary pharmaceutical source. They evaluated the risk potential of the mixtures of 100 pharmaceuticals with the toxicity data. A QSAR model to predict baseline toxicity was generated and the risk analysis was performed by calculating the risk quotient.

In a recent quantitative structure–toxicity relationship (QSTR) study by Kar and Roy [6], two interspecies correlation models were developed for diverse pharmaceuticals. The authors presented linear models for toxicity prediction of fish and Daphnia, which can be used for ecotoxicological hazard assessment for pharmaceuticals where data gaps exist for both species.

There are ecological structure–activity relationship (ECOSAR) based models in the literature for prioritization purposes. Madden et al. [9] studied the toxicity classification of pharmaceuticals with the use of ECOSAR. They concluded that toxicity estimation of these chemicals with ECOSAR should be used with caution in terms of applicability domain. In another QSAR study of pharmaceuticals [10], Sanderson and Thomsen explored whether ECOSAR could predict the toxicity class of pharmaceuticals accurately. They emphasized that the majority of the pharmaceuticals have the narcosis MOA and that $\log K_{ow}$ is an important parameter in the expression of acute toxicity of these chemicals. However, Fent et al. [1] could not find any correlation between the $\log K_{ow}$ of pharmaceuticals and the acute toxicity of a certain species.

The present studies in the literature on QSAR models target pharmaceutical categorization, hazard assessment, and MOA estimation mainly depending on hydrophobicity. Although structurally diverse pharmaceuticals are assumed to cover a broad range of toxic mechanisms (e.g. non-specific narcosis, electrophilic and specific mechanisms), there is no distinct classification of these chemicals in terms of their mode of actions, therapeutic uses and chemical classes. These compounds are nonetheless important from environmental point of view like other industrial chemicals. They are designed to be biologically active compounds. Therefore, their potential effects on non-target aquatic species should be considered. Developing a QSAR model for diverse pharmaceuticals will be beneficial given the lack of knowledge of the potential harmful effects of potent, continually and increasing amounts of pharmaceuticals released into the aquatic environment. This study aims to develop a QSAR model to estimate acute pharmaceutical toxicity using multiple descriptors to make point estimations for fish.

2. Materials and methods

2.1 *Data set*

Sanderson and Thomsen [10] compiled a set of 147 pharmaceuticals with toxicity (LC_{50}) values for various freshwater fish species (rainbow trout, fathead minnow, guppy, bluegill, etc.). Their pre-processing of data included choosing the lowest measured acute effect concentration obtained from different test systems (static, flow-through, etc.) by screening seven publicly available databases. Different test systems and fish species as well as other factors (hardness, age and gender of the fish, physical and chemical parameters of the environment, etc.) may affect the acute toxicity of chemicals to fish. However, Mayer and Ellersieck [14] stated that there is no difference between static and flow-through test pairs for the acute toxicities of 37% of chemicals they studied. Additionally, they stated that the acute fish toxicities of organic chemicals for many fish species are highly correlated. On the other hand, the pharmaceutical data for acute fish toxicity is very limited [1]. Therefore, we used the data set compiled by Sanderson and Thomsen [10], although it has some drawbacks.

For the model development, LC_{50} values were converted to molar basis (mM) and then the negative logarithm of the concentrations (pT) was used as observed toxicity values. In this study, some of the compounds were excluded from the data set because of their atypical nature (metal-containing compounds, salts, disconnected compounds, water-containing compounds and ionized compounds) and some were excluded due to the inconsistency between their names and CAS numbers. Therefore, a diverse set of 99 pharmaceuticals spanning a wide range of pharmacological classes (analgesics, anti-inflammatories, depressants, anti-depressants, diuretics, hormones, hormone antagonists, anti-neoplastics, β-blockers, antibiotics, lipid regulating agents, gastrointestinal agents, cardiovascular agents, respiratory system agents, anti-coagulants and fatty acid synthesis inhibitors) participated in the modelling exercise. The data set used is presented in Appendix 1 in the supplementary material which is available via the multimedia link on the online article webpage.

The external set of 14 compounds was compiled from the TerraTox database [15], Roche database [16], US Environmental Protection Agency (USEPA) ECOTOX database [17] and Nassef et al. [18].

2.2 *Molecular descriptors and subset selection*

A large number of molecular descriptors were calculated for each chemical using three software packages, namely, Dragon v.5.4 [19], Spartan 06 [20] and Codessa 2.2 [21]. Before the calculation step, the structures of the compounds were sketched using Spartan 06 software package and geometrically optimized employing the semi-empirical PM3 method. The molecular geometries corresponding to the lowest energy conformer were selected for calculations of the molecular descriptors. The total pool of 1393 Dragon, 161 Codessa and 10 Spartan descriptors were computed. The Dragon descriptors belong to the following classes: charge and geometrical descriptors, connectivity indices, 3D-MoRSE, GETAWAY, RDF and WHIM descriptors, etc. The Codessa descriptor set including constitutional, topological, geometrical and electrostatic descriptors. The Spartan descriptors set is constructed of dipole moment (μ), the energy of the lowest unoccupied molecular orbital (E_{LUMO}), the energy of the highest occupied molecular orbital (E_{HOMO}) and the gas phase energy (E). The rest of the calculated descriptors

such as $E_{LUMO} - E_{HOMO}$ gap, hardness, electronegativity, softness and electrophilicity (ω) were calculated from the energies obtained from Spartan 06 and using the formula reported for each by LoPachin et al. [22].

The significant descriptors for MLR models were selected by the heuristic method (HM) running in Codessa 2.2. The HM algorithm selects the descriptors according to the following criteria [23]. The program calculates all correlations between individual descriptors and property (toxicity) and eliminates descriptors considering the following criteria: *F*-test's value is less than one; correlation coefficient is less than the set value (0.1); and the *t*-value is less than the set value (0.1). This method also takes into account correlations between molecular descriptors. The criterion is set at 0.99, i.e. if the descriptors are highly correlated, then the descriptor with the lower squared correlation coefficient in the one-parameter equations is removed from the descriptor list. Additionally, descriptors with variance inflation factor (VIF) values greater than five [24] were tested after heuristic analysis. The best descriptor groups having high correlations to toxicity are chosen for the model development.

2.3 *Model development and validation*

According to Organisation for Economic Co-operation and Development (OECD) principles, a QSAR model should have appropriate measures of goodness-of-fit, robustness and predictivity. While the internal performance of a model is determined using a training set, the predictivity is determined by using an appropriate test set [25]. Composition of the training and the test sets should guarantee that these sets are scattered over similar descriptor spaces and that the training set is a representative set of the whole data set. Therefore, the data set was divided into training and test sets using Kohonen Neural Network alias Self-Organizing Maps (SOM). SOM are able to select a representative training set and a test set similar to it [26,27].

Kohonen networks project multi-dimensional space onto a two-dimensional (2D) array of neurons. The projection, which is called network learning, runs in two steps. In the first step, an object (represented by a vector) is presented to all neurons and the algorithm selects the most similar neuron, called the 'winning neuron'. In the second step, the weights of the winning neuron are modified to the vector values and at the same time the neighbouring neurons are modified to become similar to it [28]. Since this division method is based on similarity analysis, the test set of the compounds is structurally similar to the training set of the chemicals in order to maintain the same chemical domain. However, the developed models could be predictive for chemicals only in the test set [29]. In order to eliminate this biased situation, we preferred to have an additional external set of compounds, which was not used in the model development [8], to confirm the predictivity of our model. The division of the data set is performed by the program developed by Zupan et al. [30].

An additional external set of 14 pharmaceuticals (antibiotics, anticonvulsant, hormone and analgesics), which were not utilized during the model development, were used for external validation of the model [15–18]. This set of compounds was representative of the chemical space of the training set, i.e. considering the applicability domain of the model.

The MLR model for the training set was obtained using the SPSS 17.0 statistical software package [31]. Models with varying numbers of descriptors were examined. The model was checked for overfitting due to high number of descriptors (Topliss ratio) [25]

and variable multicollinearity. Large VIF values (over five) were not allowed to avoid multicollinearity [24]. For robustness of the model, the number of compounds (n), squared correlation coefficient (R^2), adjusted (for degrees of freedom) squared correlation coefficient (R^2_{adj}) and Fischer statistics (F) were reported. For training and test sets of the model, root mean square error ($RMSE$) and average absolute error (AAE) were calculated. Internal validation of the model was tested with the leave-one-out (LOO) procedure and cross-validation correlation coefficient (R^2_{cv}) was calculated using MDM 2011.2.6.0 software [32].

The reliability and robustness of the MLR model were also tested using a response randomization (Y-scrambling) procedure. The significantly low correlation coefficients of the new models indicate that there is no correlation by chance [33]. For model randomization, the dependent variables of the training set were shuffled and new correlation coefficients were calculated. The process was repeated several times using MDM software [32].

The predictive power of the regression model developed on the training set was estimated on the predicted values of the external set chemicals by the predictive R^2_{pred} or (Q^2_{ext}) as stated by Gramatica et al. [34]. Additionally, we also applied the conditions described by Golbraikh and Tropsha [35] to the external set results for further external validation, which controls the fit of the regression line to $y = x$:

$$\begin{aligned} \text{i.e.:}\quad & \mathbf{I} \quad R_0^2 \text{ or } R_0'^2 \text{ close to } R^2 \\ & \quad\ \ (\mathbf{a}) \quad (R^2 - R_0^2)/R^2 < 0.1 \text{ and } 0.85 \le k \le 1.15 \text{ or} \\ & \quad\ \ (\mathbf{b}) \quad (R^2 - R_0'^2)/R^2 < 0.1 \text{ and } 0.85 \le k' \le 1.15 \\ & \mathbf{II} \quad \left|R_0^2 - R_0'^2\right| < 0.3, \end{aligned}$$

where R^2 is predicted vs. observed, R'^2 is observed vs. predicted, k and k' are slopes, R_0^2 and $R_0'^2$ are squared correlation coefficients (without intercept).

The applicability domain (AD) of the model was verified by using the ranges of descriptors and toxicity values, and then applying the leverage approach. We identified a compound as a response outlier if its standardized residual was higher than three. Chemicals structurally very influential in determining model parameters (i.e. creating leverage effect) were determined using a Williams plot. In general, critical hat value is set at $3p'/n$, where p' is the number of descriptors plus one and n is the number of compounds in the model [33]. If a compound's leverage value is greater than the critical hat value, than this compound is structurally distant from the compounds whose leverage values are smaller than the critical hat value.

3. Results and discussion

The data set of 99 compounds was divided into training and test sets for validation of the model. Kohonen networks were used for data set splitting. We used different networks for the developed model and approximately 65% of the data set was allocated to the training set. Selection of a 10×10 network and 400 epochs resulted in 64/35 division for the training/test sets. The combination of descriptors, which are highly correlated with the fish toxicity, was selected by the heuristic method. The descriptors from other software did not appear to be representative for this data set.

The model obtained for the prediction of acute toxicity of pharmaceuticals to fish, using the training set of 64 compounds, is the following linear model (Equation (1)) with the reported statistical parameters:

$$\begin{aligned} \mathrm{pT} &= 3.621(\pm 1.279) + 0.986(\pm 0.218)\mathrm{GATS3p} + 0.481(\pm 0.136)\mathrm{EEig05d} \\ &\quad - 1.104(\pm 0.362)\mathrm{BEHe3} + 1.259(\pm 0.488)\mathrm{Mor32u} - 1.794(\pm 0.663)\mathrm{HATS2u} \\ &\quad - 0.684(\pm 0.122)\mathrm{C\text{-}040} + 0.334(\pm 0.115)\mathrm{O\text{-}060} + 0.105(\pm 0.018)\mathrm{MLOGP2} \\ n_{\mathrm{training}} &= 64, \quad R^2 = 0.664, \quad R^2_{\mathrm{adj}} = 0.615, \quad R^2_0 = 0.583 \\ R^2_{\mathrm{cv}}, &= 0.542, \quad F_{8,55} = 13.588, \quad \mathrm{RMSE} = 0.602, \quad AAE = 0.491 \end{aligned} \tag{1}$$

Generally, QSAR models are functions of a molecule's structure, electronic properties and hydrophobicity [36]. In this model, Mor32u, HATS2u, C-040, and O-060 stand for structure; GATS3p, EEig05d, and BEHe3 stand for electronic properties; and MLOGP2 stands for hydrophobicity. The descriptions of the independent variables in the model equation were given in Table 1, together with their VIF and *t*-values.

GATS3p, EEig05d, O-060, Mor32u and MLOGP2 showed direct correlation with the acute fish toxicity. MLOGP2 and C-040 are descriptors that made the most contribution to the model considering the *t*-values (Table 1).

3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction) descriptors describe the distribution of the atoms in three-dimension geometrical molecules. When atomic properties act as weighting factor; these descriptors encode the distribution of the atomic properties in molecules.

Table 1. Definitions, VIF and *t*-values, and classification of the descriptors in the model.

Descriptor	*VIF*	*t-value*	*Description*	*Category*
GATS3p	1.264	4.5	2D descriptor; Geary autocorrelation – lag 3 (weighted by atomic polarizabilities)	2D autocorrelations
EEig05d	4.152	3.5	2D descriptor; Eigenvalue 05 from edge adjacency matrix (weighted by dipole moments)	Adjacency indices
BEHe3	3.869	−3.0	2D descriptor; highest eigenvalue number 3 of Burden matrix (weighted by atomic Sanderson electronegativities)	Burden eigenvalues
Mor32u	1.568	2.6	3D descriptor; 3D-MoRSE signal 32 (unweighted)	3D-MoRSE descriptors
HATS2u	2.524	−2.7	3D descriptor; leverage-weighted autocorrelation of lag 2 (unweighted)	GETAWAY descriptors
C-040	1.580	−5.6	1D descriptor; R–C(=X)–X/R–C#X/X=C=X	Atom-centred fragments
O-060	1.388	2.9	1D descriptor; Al–O–Ar/Ar–O–Ar/R··O··R/R–O–C=X	Atom-centred fragments
MLOGP2	1.534	5.7	Other descriptors; squared Moriguchi octanol–water partition coefficient	Molecular properties

Computation of GATS3p, EEig05d, and BEHe3 involves the structure of the molecule, but weighting components have been embedded in these descriptors in terms of polarizability, dipole moment and electronegativity, respectively. These descriptors address the topology of the structure in association with electronic property. BEHe3 can include connectivity information and atomic properties (e.g. atomic charge, polarizability, hydrogen-bonding) that are relevant to intermolecular interactions. BEHe3, in particular, demonstrates for the electronegativity of atoms that are separated by three bonds.

Autocorrelation descriptors calculated for 3D-spatial molecular geometry are based on interatomic distances collected in the geometry matrix G. In GATS3p, for instance, '3' indicates the autocorrelation vector of lag 3 corresponding to the number of edges in the fragment unit considered in the computation and the character 'p' refers to the atomic polarizabilities. HATS indices of GETAWAY (GEometry, Topology and Atom-Weights AssemblY) group are based on the diagonal elements of the molecular influence matrix (MIM). HATS2u is a 3D GETAWAY descriptor and it accounts for the effective position of substituents and fragments in the 3D molecular space [37].

Atom-centred fragments (ACF) are simple molecular descriptors defined as the number of specific atom types in a molecule. The fragment, C-040, a descriptor from ACF class, represents the number of carbon atoms attached to the heteroatom by single or multiple bonding and one valence is satisfied by an alkyl group. Another ACF class descriptor O-060, for example, represents a presence of the Al–O–Ar, Ar–O–Ar, R··O··R or R–O–C = X fragment in a molecular structure (Al: aliphatic group, Ar: aromatic group, X: any heteroatom, and R: any group linked through carbon). In our case, a electronegative O atom is located in different positions in the molecules of the diverse set of pharmaceuticals. The groups connecting to O atom with –O– fragment and heteroatom-attached carbon atoms influence the toxicity because these points are primary reaction centres in the molecules.

The descriptors used in the present model have been utilized in previous QSAR models in the literature. In a study by Bozorgi et al. [38], GATS3p was used for IC_{50} estimation of telomerase inhibition for cancer cells in a linear model. The blood–brain–barrier (BBB) penetration coefficient (logBB) was modelled with EEig05d by Soto and co-workers [39] in a MLR model. Gonzales et al. [40] assessed the acute toxicity of 69 benzene derivatives using the BEHe3 descriptor. A QSTR model for predicting the guppy toxicity was developed by Duchowicz et al. [41] with Mor32u for a set of benzene derivatives. A QSAR model on mouse oral LD_{50} data of 58 chemicals developed by Bhhatarai and Gramatica [42] employed HATS2u. In a study by Duchowicz et al. [43], C-040 and O-060 were used to model *Tetrahymena pyriformis* growth inhibition by phenol derivatives. Shao et al. [44] utilized MLOGP2 in their models to explain *Tetrahymena pyriformis* toxicity.

MLOGP2 represents hydrophobicity, whereas the descriptors appearing in the developed model represent topological information and electronic properties. The participating descriptors of the model therein suggest that atomic properties in 3D molecular space (Mor32u), effective position of substituents and fragments (HATS2u), atom-centred fragments (C-040 and O-060) and molecule's weighting components in terms of electronegativity (BEHe3), polarizability (GATS3p) and dipole moment (EEig05d) revealed an influence on the toxicity of diverse pharmaceuticals to fish.

Figure 1 shows the calculated/predicted and observed values for the training, test and external set compounds highlighted by different markers. Test set compounds achieve a squared correlation coefficient of 0.605, R_0^2 of 0.594, *RMSE* of 0.671, and *AAE* of 0.512

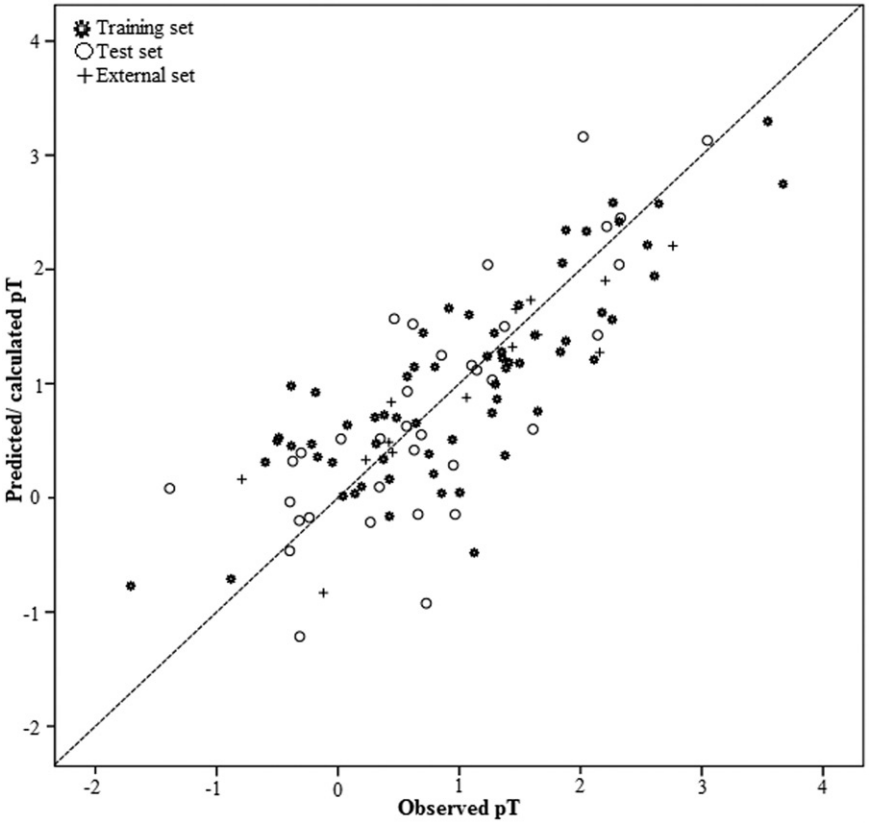

Figure 1. Calculated/predicted pT vs. experimental pT for the training, test and external set compounds.

while external set scores squared correlation coefficient of 0.788, AAE of 0.348, and R^2_{pred} of 0.777, which implies a predictive model. The presented model was subjected to the test for the criteria of external validation as recommended by Golbraikh and Tropsha [35]. The calculations resulted in $|R_0^2 - R_0'^2| = 0.031$, $(R^2 - R_0^2)/R^2 = 0$ and $k = 1.1082$ revealing a high predictive power of the model.

The reliability of the model was checked using a response randomization test. Random shuffling of response was repeated 25 times for the equation. R^2 values ranged between 0.012 and 0.151 with a mean value of 0.055. The results reveal that the proposed model is well founded and not just the result of a correlation by chance.

We define the applicability domain of the model as the descriptor space of the 64 chemicals of the training set (Table 2). The applicability domain of the model was also analysed using a Williams plot (Figure 2), where the vertical reference line is the critical leverage value (h^*), and the horizontal reference lines are $\pm 3\sigma$, the cut-off value for response outliers. As seen in Figure 2, there is no response outlier of training, test and external sets with a three standard deviation unit. Acroleine, amidosulfonic acid and epichlorohydrine are the high leverage compounds of the training set and influential for model development. In fact, these compounds influence the regression line and their residuals are small. Ethyl bromide is the high leverage compound of the test set.

Table 2. Descriptor and toxicity space of the model.

Variable	*Minimum value*	*Maximum value*
Toxicity	−1.708	3.671
GATS3p	0.000	2.056
EEig05d	−1.446	3.628
BEHe3	1.093	3.749
Mor32u	−0.735	0.185
HATS2u	0.173	1.576
C-040	0.000	4.000
O-060	0.000	5.000
MLOGP2	0.002	22.665

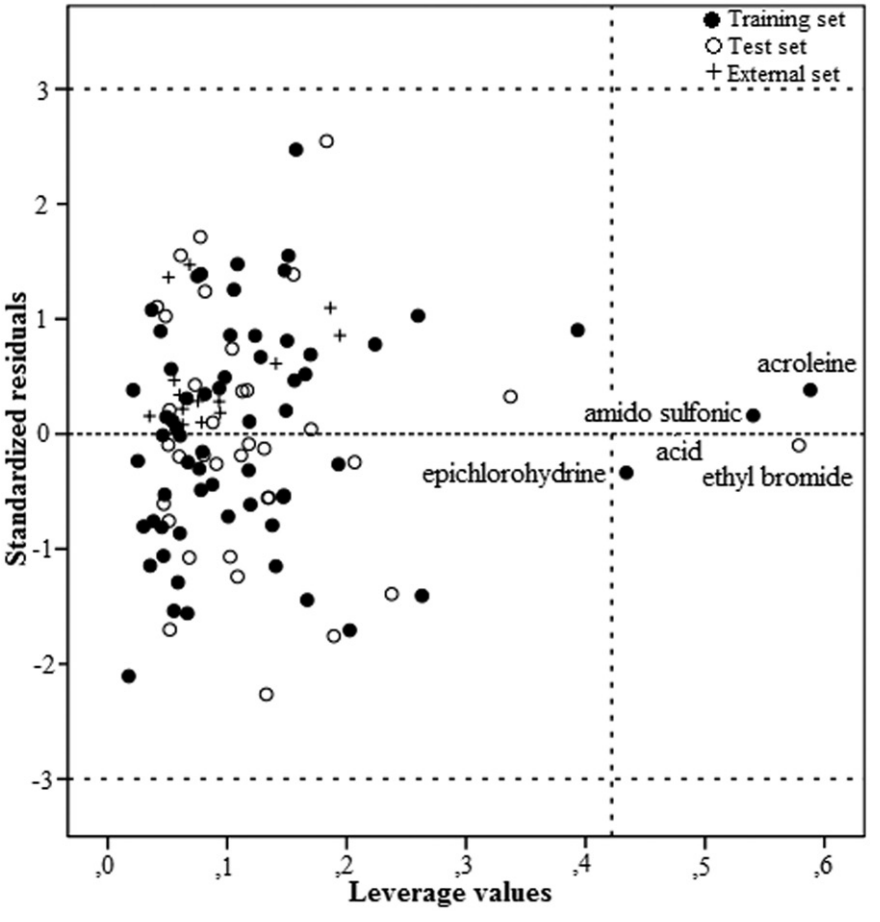

Figure 2. Projection of the standardized residuals vs. leverage values of the training, test and external set compounds. Critical hat value (h^*) is set at 0.422. Response outlier limits are set at $\pm 3\sigma$.

This compound has a unique structure as the only bromoalkane in the data set. Ethyl bromide shows an extreme trend, e.g. it has the minimum values for GATS3p, BEHe3, C-040 and O-060, and a maximum value for HATS2u among the test set compounds. Visual inspection of the four high leverage pharmaceuticals reveals that these compounds have the simplest structures with relatively low number of atoms within the data set. It is also important to note that the external set has no leverage value higher than the critical hat value.

Sanderson and Thomsen [10] compiled a very large database that contains a diverse set of pharmaceuticals. This database is a result of selective study of collecting seven publicly available databases. However, each database elicits toxicity information from a different and vast number of laboratories. For each of the chemicals in a single database, the LC_{50} value for a specific organism may yield thousands of entries and variability of results can exceed several orders of magnitude. Potential causes of the data variability include influence of biological and physical factors [45]. The performance of our model should be considered together with the lack of consistency in the literature data in experimental conditions, duration, endpoints measured, and species used as stated by Madden et al. [9]. Despite these disadvantages, our model performs very well in terms of its predictive power.

Kar and Roy [6] developed a linear model with the same pharmaceutical and fish data taken from Sanderson and Thomsen [10]. Their model included *Daphnia magna* toxicity and two additional descriptors as independent variables. In the present work, we developed a model with only theoretical descriptors. We have 71 out of 77 pharmaceuticals in common. Six compounds were not included in our data set either because of inconsistency between the names and their CAS numbers (e.g. carbacystine, chlorolactam, verapamil) or the presence of replicates (e.g. aspirin and acetyl salycilic acid). The average absolute error of the predicted 71 compounds for our and their models is 0.462 and 0.526, respectively.

It is of our interest to compare the predictions of the external set compounds with those of ECOSAR [46] models in which log *P* based toxicity models were developed. ECOSAR classes and corresponding predictions for each chemical in the external set are given in Table 3. The average absolute error of their predictions obtained by our model and ECOSAR is 0.348 and 0.872, respectively. Although we used the closest ECOSAR predictions to the reported literature values among the available predictions to calculate *AAE*, our model had better predictions than that of ECOSAR for the external set compounds.

Although ECOSAR is a publicly available program that can be easily applied for the prediction of ecotoxicity of chemicals, there are wide variations in the predicted values of a pharmaceutical from log *P* depending on the ECOSAR classes in which the chemical belongs to. Variations in performances of the ECOSAR classes are discussed by Reuschenbach et al. [47]. Additional drawbacks of the ECOSAR program are discussed by Fent et al. [1]. Therefore, ECOSAR could fill a gap where a better QSAR model is not present. The level of accuracy of the developed model is good enough considering the many sources of error (e.g. variations in test conditions) that may impact the model. Our model seems to be more complex compared to ECOSAR, but the descriptors used in this MLR are attractive because they can be calculated easily and rapidly.

4. Conclusions

This QSTR study involved 99 pharmaceuticals with an additional external set of 14 pharmaceuticals modelled for their toxicity to fish based on MLR with descriptors

Table 3. ECOSAR predictions for the external set.

Pharmaceuticals	*ECOSAR class(es)*	*ECOSAR prediction(s) from log P (mM)*	*Reported literature pT (mM)*	*References*
Amobarbital	Baseline toxicity (2010)*	0.089	0.420	[14]
	Carbonyl ureas (11)	3.404		
Ampicillin	Baseline toxicity (2010)	−0.405	−0.120	[14]
	Aliphatic amines (acid) (90)	−0.642		
	Amides (acid) (28)	−0.866		
Carbamazepine	Baseline toxicity (2010)	0.300	1.590	[14]
	Substituted ureas (21)	0.757		
Chloramphenicol	Baseline toxicity (2010)	−0.888	0.450	[14]
	Benzyl alcohols (11)	−0.437		
	Haloacetamides (10)	0.920		
Cotinine	Baseline toxicity (2010)	−1.406	0.230	[14]
	Amides (28)	−0.663		
Diclofenac	Neutral organics (acid) (296)	0.896	1.467	[17]
Estradiol	Baseline toxicity (2010)	1.830	2.205	[16]
	Phenols (203)	2.237		
Flutamide	Baseline toxicity (2010)	1.439	2.158	[16]
	Amides (28)	1.603		
Gentisic acid	Baseline toxicity (2010)	−0.126	0.440	[14]
	Poly−acid phenols (24)	0.131		
	Hydroquinones (acid) (6)	1.986		
Isotretinoin	Neutral organics (acid) (296)	4.332	2.762	[15]
Metronidazole	Baseline toxicity (2010)	−1.713	−0.792	[16]
	Imidazoles (12)	2.237		
Norfloxacin	Baseline toxicity (2010)	−1.986	1.060	[14]
	Aliphatic amines (90)	−1.799		
	Vinyl/allyl ketones (acid) (7)	−2.765		
Pyrimethamine	Baseline toxicity (2010)	0.453	1.648	[15]
	Anilines (unhindered) (49)	1.131		
	Anilines (hindered) (13)	0.649		
	Anilines (amino-meta) (2)	0.902		
Sulfamethoxazole	Baseline toxicity (2010)	−1.276	1.440	[14]
	Anilines (unhindered) (49)	−0.210		
	Amides (28)	−0.559		

*Numbers in parantheses display the number of compounds used in ECOSAR modelling.

calculated by Dragon software and selected by a heuristic method. The proposed model was validated internally and externally with proper statistical tools. The results indicate that the model we built is robust and satisfactory, and that the selected descriptors are able to explain the toxicity of a diverse set of pharmaceuticals to fish. Given the model's descriptors and given the *t*-values, we can conclude that the toxicity of these compounds mainly depends on their hydrophobicity and heteroatom-bonded carbon atom. Descriptors weighted by polarizability, dipole moment and electronegativity are also involved in pharmaceutical toxicity modelling. It is likely that different factors affect the toxicity of heterogeneous pharmaceuticals. Therefore, more complex descriptors (i.e. 3D-MoRSE and GETAWAY) seem to be useful in their modelling of fish toxicity. The QSTR model developed in this study can provide a useful tool for point toxicity

estimations of pharmaceuticals within the applicability domain relying on the high predictive squared correlation coefficient.

Acknowledgments

The support of this study by TUBITAK-ARRS (Project No: 108Y119) is gratefully acknowledged. The financial support of Boğaziçi University Research Fund (Project Number: 6052) and the Slovenian Ministry of Higher Education, Science and Technology (grant P1-0017 and BI-TR/09-11-003) is also appreciated.

References

[1] K. Fent, A.A. Weston, and D. Caminada, *Ecotoxicology of human pharmaceuticals*, Aquat. Toxicol. 76 (2006), pp. 122–159.
[2] H. Sanderson and M. Thomsen, *Ecotoxicological quantitative structure–activity relationships for pharmaceuticals*, Bull. Environ. Contam. Toxicol. 79 (2007), pp. 331–335.
[3] O.A.H. Jones, N. Voulvoulis, and J.N. Lester, *Aquatic environmental assessment of the top 25 English prescription pharmaceuticals*, Water Res. 36 (2002), pp. 5013–5022.
[4] V.L. Cunningham, M. Buzby, T. Hutchinson, F. Mastrocco, N. Parke, and N. Roden, *Effects of human pharmaceuticals on aquatic life: Next steps*, Environ. Sci. Technol. 40 (2006), pp. 3457–3462.
[5] R. Lange and D. Dietrich, *Environmental risk assessment of pharmaceutical drug substances—conceptual considerations*, Toxicol. Lett. 131 (2002), pp. 97–104.
[6] S. Kar and K. Roy, *First report on interspecies quantitative correlation of ecotoxicity of pharmaceuticals*, Chemosphere. 81 (2010), pp. 738–747.
[7] M.T.D. Cronin, in *Predicting chemical toxicity and fate in humans and the environment—an introduction, in Predicting Chemical Toxicity and Fate*, M.T.D. Cronin and D. Livingstone, eds., CRC Press, Boca Raton, FL, 2004, pp. 18–28.
[8] A. Tropsha, *Best practices for QSAR model development, validation, and exploitation*, Mol. Inf. 29 (2010), pp. 476–488.
[9] J.C. Madden, S.J. Enoch, M. Hewitt, and M.T.D. Cronin, *Pharmaceuticals in the environment: Good practice in predicting acute ecotoxicological effects*, Toxicol. Lett. 185 (2009), pp. 85–101.
[10] H. Sanderson and M. Thomsen, *Comparative analysis of pharmaceuticals versus industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q)SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action*, Toxicol. Lett. 187 (2009), pp. 84–93.
[11] M. Crane, C. Watts, and T. Boucard, *Chronic aquatic environmental risks from exposure to human pharmaceuticals*, Sci. Total Environ. 367 (2006), pp. 23–41.
[12] S.K. Khetan and T.J. Collins, *Human pharmaceuticals in the aquatic environment: A challenge to green chemistry*, Chem. Rev. 107 (2007), pp. 2319–2364.
[13] B.I. Escher, R. Baumgartner, M. Koller, K. Treyer, J. Lienert, and C.S. McArdell, *Environmental toxicology and risk assessment of pharmaceuticals from hospital wastewater*, Water Res. 45 (2011), pp. 75–92.
[14] F.L. Mayer and M.R. Ellersieck, *Manual of Acute Toxicity: Interpretation and Data Base for 410 Chemicals and 66 Species of Freshwater Animals*, Resource Publication 160, United States Department of the Interior, U.S. Fish and Wildlife Service, 1986.
[15] http://www.terrabase-inc.com/ (last accessed July 2011).
[16] http://www.roche.com (last accessed July 2011).
[17] http://cfpub.epa.gov/ecotox/ (last accessed July 2011).

[18] M. Nassef, S. Matsumoto, M. Seki, I.J. Kang, J. Moroishi, Y. Shimasaki, and Y. Oshima, *Pharmaceuticals and personal care products toxicity to Japanese medaka fish (Oryzias latipes)*, J. Fac. Agric., Kyushu Univ 54 (2009), pp. 407–411.
[19] *DRAGON for Windows 5.4*, Talete srl, 2006; software available at http://www.talete.mi.it/
[20] *SPARTAN 06*, Wavefunction Inc., Irvine, USA, 2006.
[21] *CODESSA 2.20*, Semichem Inc., Shawnee Mission, KS, 1994–1996.
[22] R.M. LoPachin, T. Gavin, B.C. Geohagen, and S. Das, *Neurotoxic mechanisms of electrophilic type-2 alkenes: Soft-soft interactions described by quantum mechanical parameters*, Toxicol. Sci. 98 (2007), pp. 561–570.
[23] A.R. Katritzky, V.S. Lobanov, and M. Karelson, *CODESSA: Training Manual*, University of Florida, Gainesville, FL, 1995.
[24] L. Xi, H. Sun, J. Li, H. Liu, X. Yao, and P. Gramatica, *Prediction of infinite-dilution activity coefficients of organic solutes in ionic liquids using temperature-dependent quantitative structure–property relationship method*, Chem. Eng. J. 163 (2010), pp. 195–201.
[25] OECD, *Guidance document on the validation of (quantitative) structure–activity relationship [(Q)SAR] models*, ENV/JM/MONO (2007)2, OECD Environment Health and Safety Publications Series of Testing and Assessment No. 69, Organisation for Economic Co-operation and Development, Paris, 2007.
[26] J. Devillers, *Neural Networks in QSAR and Drug Design*, Academic Press, San Diego, CA, 1996.
[27] J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Wiley-VCH, Weinheim, 1999.
[28] M. Vracko, *Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies*, Curr. Comp.-Aid. Drug Des. 1 (2005), pp. 73–78.
[29] H. Liu and P. Gramatica, *QSAR study of selective ligands for the thyroid hormone receptor β*, Bioorg. Med. Chem. 15 (2007), pp. 5251–5261.
[30] J. Zupan, M. Novic, and I. Ruisánchez, *Kohonen and counter propagation artificial neural networks in analytical chemistry*, Chemom. Intell. Lab. Syst. 38 (1997), pp. 1–23.
[31] *SPSS Statistics 17.0 for Windows* (Statistical Package for Social Scientists), SPSS Inc., 2008.
[32] *MDM 2011.2.6.0* (Molegro Data Modeller) Molegro ApS., 2011.
[33] E. Papa, J.C. Dearden, and P. Gramatica, *Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors*, Chemosphere. 6 (2007), pp. 351–358.
[34] P. Gramatica, E. Giani, and E. Papa, *Statistical external validation and consensus modeling: A QSPR case study for Koc prediction*, J. Mol. Graph. Model 25 (2007), pp. 755–766.
[35] A. Golbraikh and A. Tropsha, *Beware of q2!*, J, Mol. Graph. Model. 20 (2002), pp. 269–276.
[36] H. Sanderson, D.J. Johnson, T. Reitsma, R.A. Brain, C.J. Wilson, and K.R. Solomon, *Ranking and prioritization of environmental risks of pharmaceuticals in surface waters*, Reg. Tox. and Pharm. 39 (2004), pp. 158–183.
[37] V. Consonni, R. Todeschini, and M. Pavan, *Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors*, J. Chem. Inf. Comput. Sci. 42 (2002), pp. 682–692.
[38] A. H. Bozorgi, H. T. Ghomi, and A. Jouyban, *QSAR and pharmacophore studies of telomerase inhibitors*, Med. Chem. Res. (2011), DOI 10.1007/s00044-011-9594-4.
[39] A.J. Soto, R.L. Cecchini, G.E. Vazquez, and I. Ponzoni, *Multi-objective feature selection in QSAR using a machine learning approach*, QSAR Comb. Sci. 28 (2009), pp. 1509–1523.
[40] M.P. Gonzalez, A.M. Helguera, and M.A. Cabrera, *Quantitative structure–activity relationship to predict toxicological properties of benzene derivative compounds*, Bioorg. Med. Chem. 13 (2005), pp. 1775–1781.
[41] P.R. Duchowicz, J.J. Marrugo, E.V. Ortiz, E.A. Castro, and R. Vivas-Reyes, *QSAR Study for the fish toxicity of benzene derivatives*, J. Argent. Chem. Soc. 97 (2009), pp. 116–127.
[42] B. Bhhatarai and P. Gramatica, *Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse*, Mol. Divers. 15 (2011), pp. 467–476.

[43] P.R. Duchowicz, A.G. Mercader, F.M. Fernández, and E.A. Castro, *Prediction of aqueous toxicity for heterogeneous phenol derivatives by QSAR*, Chem. Int. Lab. Sys. 90 (2008), pp. 97–107.
[44] L. Shao, L. Wu, X. Fan, and Y. Cheng, *Consensus ranking approach to understanding the underlying mechanism with QSAR*, J. Chem. Inf. Model. 50 (2010), pp. 1941–1948.
[45] M. Hrovat, H. Segner, and S. Jeram, *Variability of in vivo fish acute toxicity data*, Reg. Tox. Pharm. 54 (2009), pp. 294–300.
[46] ECOSAR v. 1.1; software available at http://www.epa.gov/oppt/newchems/tools/21ecosar.htm
[47] P. Reuschenbach, M. Silvani, M. Dammann, D. Warnecke, and T. Knacker, *ECOSAR model performance with a large test set of industrial chemicals*, Chemosphere. 71 (2008), pp. 1986–1995.