Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Husowitz and Sanchez-Arias, J Biom Biostat 2017, 8:6
DOI: 10.4172/2155-6180.1000385

Research Article Open Access

A Machine Learning Approach to Designing Guidelines for Acute Aquatic Toxicity

Barry Husowitz[1]* and Reinaldo Sanchez-Arias[2]

[1]*Department of Applied Mathematics, Wentworth Institute of Technology, Boston, USA*
[2]*Department of Mathematics, St. Thomas University, USA*

Abstract

A support vector classification wrapper feature elimination approach was used to find the most relevant pairs of molecular features that adequately and accurately can predict acute aquatic toxicity. These pairs were then used to derive chemical thresholds or boundaries between chemical properties for toxic and nontoxic organic chemicals that can be used as a "rule of thumb" to design less toxic chemicals. The most relevant pairs were determined to be: Lowest Unoccupied Molecular Orbital (LUMO) and Aqueous Solubility (QPlogS), Difference between the LUMO and HOMO (dE) and Octonal-Water Partition Coefficient (QPlogo.w), and Difference between the LUMO and HOMO (dE) and Van der Waals surface area of polar nitrogen and oxygen atoms (PSA). Projected hyper planes were constructed for each pair and the following thresholds were found: for Lowest Unoccupied Molecular Orbital (LUMO) and Aqueous Solubility (QPlogS) they roughly correspond to QPlogS>-1 and LUMO>1, and for Octonal-Water Partition Coefficient (QPlogo.w) vs. difference between the LUMO and HOMO (dE) they roughly correspond to QPlogo.w<1 and dE>9. This study shows how a statistical approach such as support vector machines can be applied to the rational design of chemicals with reduced toxicity.

Keywords: Support vector machines; Toxicity; Computational models

Introduction

The number of chemicals synthesized every year is increasing exponentially [1,2]. Many of these compounds are toxic to humans or have a negative effect on the environment [2]. Only after the chemicals are introduced into the environment the toxic effects on humans and the environment are discovered. This occurs primarily because of poorly efficient risk assessment processes and limited information on hazard properties of chemicals. Experimental toxicological approaches can be used to assess the toxicity of new chemicals; however these approaches can be quiet expensive and very time consuming [3]. Consequently, there is a great need to develop computational and statistical models to predict the toxicity of chemicals post-production. The need to predict the toxicity of chemicals post-production has driven the development of new regulatory dispositions that were introduced in the European Community on June 1, 2007, with the chemical management system REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals) [4,5]. The objective of REACH is to characterize the toxicological properties of a large group of substances, manufactured or important quantities in excess of 1 ton per year. This regulation attempts to increase the production of useful data for better decision making in regards to human health and the environment. These regulations have been set forth requiring the chemicals to be carefully tested before entering the market, and forced the development of robust QSAR (Quantitative Structure Activity Relationships) models or computational toxicity screening approaches to aid in the assessment of potentially toxic chemicals. Regulatory agencies such as the European Union and the U.S. Environmental Protection Agency (US EPA) have used QSAR models to predict ecologic effects and the environmental fate of chemicals [6].

QSAR models for toxicity risk assessment offer a mathematical relationship between structural or descriptive features of a set of chemicals and the toxicity associated with them. In order to obtain the toxicity associated with a set of chemicals, for example a set of organic chemicals, one can investigate the acute toxicity 96 hour or 48 hour median lethal concentration (LC_{50}) or median effective concentration (EC_{50}) values for various aquatic species such as the fathead minnow, Japanese Medaka and *Daphnia magna* [7,8]. These values can then be incorporated into a QSAR model as the chemical activity or biological activity. Many traditional QSAR methods, which use these values, have been developed for acute aquatic toxicity over the years. Studies which use traditional linear QSAR methods such as ordinary regression, principle component regression, and partial least squares regression assume that structurally similar chemical class act through the same mode of action. However, Russom et al. illustrated that chemical of the same class may act through different modes of action [9]. Furthermore, classical linear QSAR methods have been shown to have problems associated with overfitting and are not able to handle nonlinear relationships [10-12]. QSAR methods have evolved over time to include Modes of Action [13,14]. For example, Verhaar et al. developed a QSAR model based on the mode of action (MOA) in which they classified a large number of organic pollutants into four classes: narcosis, polar-narcosis, reactive chemicals and specific mechanism of action [15,16]. The advantage of this type of model is that it allows one to make high quality predictions. Traditional QSAR models have recently been reconstructed to include more sophisticated theoretically based approaches and mathematical tools. More sophisticated QSAR approaches such as fragment-based two dimensional QSAR models and multiple field three-dimensional QSAR models have been recently developed [17,18]. These methods remarkably enhance the predictive

***Corresponding author:** Barry Husowitz, Assistant Professor of Applied Mathematics, Wentworth Institute of Technology, Boston, USA, Tel: 617-989-4341; E-mail: husowitzb@wit.edu

Received December 06, 2017; **Accepted** December 15, 2017; **Published** December 29, 2017

Citation: Husowitz B, Sanchez-Arias R (2017) A Machine Learning Approach to Designing Guidelines for Acute Aquatic Toxicity. J Biom Biostat 8: 385. doi: 10.4172/2155-6180.1000385

power of QSAR models and additionally provide one with more molecular structural information than conventional or traditional QSAR. However, since the MOAs cannot be correctly defined "a priori" most of these QSAR methods still tend to under or overestimate toxicity [19,20]. The overall goal of QSAR is to predict the biological activity or chemical activity of untested compounds for which the mode of action is uncertain. Nonlinear machine learning techniques such as support vector machines (SVM) and neural networks are modeling approaches that can be applied to toxicity that meets this requirement. For example, SVM methods have been used to predict the mechanism of toxic action for the fathead minnow [21,22].

Early QSAR models for acute aquatic toxicity were based mainly on the logarithm of the octonal-water coefficient [23-27]. In a more recent study done by Papa et al, they compared the predictability of chemicals classified according to their MOA for different molecular descriptors using multiple linear regression and a genetic algorithm-variable subset selection procedure for the acute aquatic toxicity of organic chemicals for the fathead minnow [28]. More advanced nonlinear and machine learning approaches such as neural networks and support vector machines (SVM) methods haven been used to predict the toxicity of organic compounds to the fathead minnow. In all these studies predictions were made for the acute aquatic toxicity of the fathead minnow, however these approaches do not attempt to define cut off/threshold values of chemical properties, which in turn can be used as design guidelines for designing safer or less toxic chemicals [29-31]. Only recent work done by Voutchkova et al. showed that organic chemicals with large differences in the Lowest Unoccupied Molecular Orbital (LUMO) and Highest Occupied Molecular Orbital energy and low octonal-water partition coefficients are likely to have low acute aquatic toxicity [32].

In this study we used a support vector classification Wrapper Feature Elimination approach to find the most relevant pairs of molecular features that adequately and accurately can predict acute aquatic toxicity. Furthermore, we used these pairs of features to derive chemical thresholds or boundaries between chemical properties for toxic and nontoxic organic chemicals that can be used as a "rule of thumb" to design less toxic chemicals. These thresholds or rules can also be used to screen toxic chemicals prior to their introduction into the environment. This study specifically shows how support vector classification can be used to derive threshold values for chemical properties, which in turn can tell someone whether a chemical will be toxic or non-toxic based on these values.

Only recently few studies have been concerned with the rational design of chemicals with reduced toxicity [32,33]. Very few QSAR or statistical approaches for that matter have been concerned with the rational design of chemicals with reduced toxicity. This paper shows how a statistical approach or QSAR approach such as support vector machines can be applied to the rational design of chemicals with reduced toxicity and how support vector machines using a Wrapper Feature Elimination approach can be used to reduce the dimensionality of a large data set (feature selection).

Methods

Data set and data preparation

The quality of the data to do any statistical analysis or toxicity prediction is as important as the methodology used. In this study three aquatic species where included in the analysis, the Fathead minnow (*Pimephales promelas*) *Pseudokirchneriella subcapitata* (green algae), *Oryzias latipes* (Japanese Medaka fish) and *Daphnia magna* (freshwater flea). The Fathead minnow data was taken from the U.S.-E.P.A. Duluth Fathead Minnow database which is considered to be a reliable source for quality toxicity measurments [34,35]. Medial lethal concentration values (LC50) 96 hour flow through acute toxicity assays were obtained from this database for 617 compounds. Only 570 compounds were used for analysis, since 37 did not have LC50 values while for 10 compounds we could not obtain their physicochemical properties. Furthermore, since these measurements are considered to be the "gold standard" for toxicity measurement, this data was used as the training set throughout this study, while the other data sets were used as test sets. Another reliable source of toxicity measurements is the Japanese Ministry of Environment which contain toxicity measurements for 948 compounds [36]. 285 compounds were for the Japanese Medaka (LC50 96 hour toxicity assays), while 363 compounds were for *Daphnia magna* (EC50 48 hour toxicity assays) and 300 were for *Pseudokirchneriella subcapitata* (EC50 72 hour toxicity assays). There were a total of 1218 acute toxicity data points, however only 570 compounds from the fathead minnow dataset where used for the training set while 231 unique chemicals for the Japanese Medaka, 288 unique chemicals for the *Daphnia magna* and 247 unique chemicals for *Pseudokirchneriella subcapitata* not include in the fathead minnow dataset were used as validation datasets. The multiple values that existed for some of the same compounds and species were converted to a single LC50 or EC50 value by a simple geometric mean.

Validation of model

The robustness and validity of a machine learning technique such as support vector machines is very important. Many methods are used to validate the accuracy of a classification model, such as sub-sampling test or n-fold cross-validation (n=5, 10, 15), jackknife test, bootstrapping and independent dataset test [37]. Among these choices the jackknife test is considered the most objective and rigorous cross-validation approach to test the accuracy of a classification model [38-40]. However, the jackknife test is computationally expensive and rather we adopted a 5 fold cross validation method. In this study 5 fold cross-validation was used in feature selection, the optimization of model parameters and to determine the hyper planes obtained from support vector classification. The data was first partitioned into 5 equally or nearly equal sets and for each iteration four subsets were chosen as the training set while the fifth set was the test set. Five iterations are run so that every subset is selected as a test set once. In machine learning techniques such as support vector machines the choice of parameters such as the regularization parameter C and the γ parameter for the various kernels are very important and need to be optimized properly since the wrong parameters can lead to over fitting. These parameters are explained later in the machine learning section. In this study 5-fold cross validation was used to optimize these parameters for the feature selection process and optimization of the hyper plane. External validation and a different machine learning technique were used to validate the model. External validation was done by considering the unique organic compounds obtained from the Japanese Medaka, *daphnia magna* and *Pseudokirchneriella subcapitata* as validation sets for our support vector classification model. Furthermore, a decision tree with 5-fold cross validation was run for all are pairs of selected features from the fathead minnow data set. The accuracies both for the positive (+1) and negative (-1) classes where considered with this decision tree model.

Molecular descriptor calculations

The physical and chemical properties used in this study were

previously calculated by Voutchkova et al. [32] Multistructure 2-dimensional SD (chemical table) files for the neutral organic compounds were converted to 3-dimensional structures by using the molecular coordinate convertor program babel [41]. The descriptors were then calculated by the Schrodinger program Qikprop version 3 a well-established and validated program used in drug discovery [42,43]. A total of 36 physical and chemical properties were calculated which include such descriptors as the number of hydrogen bond acceptors, hydrogen bond donors, number of amide, and amine groups as well as various partition coefficients (e.g. octanol/water, octanol/air octanol/gas etc.) and other properties such as globularity, molecular volume and solvent accessible surface area. The Qikprop program also provides descriptors relating to the molecules electronic structure based on the PM3 semi-empirical quantum chemistry approach. However, semi-empirical AM1 calculations for the frontier molecular orbitals i.e. highest occupied molecular orbital HOMO and lowest occupied molecular orbital LUMO using Gaussian 03 were carried out for each molecule by Voutchkova et al. [32]. In this previous study B3LYP density functional theory using the 6-31+G(d,p) basis set was used to calculate the HOMO and LUMO of a randomly selected set of 50 molecules. It was determined that the HOMO and LUMO values from DFT and AM1 compared well, with the exception of some compounds featuring acidic functional groups [31].

Classification scheme of toxicity

The EPA divides chemicals into four categories of concern for acute aquatic toxicity based on their LC50 and EC50 values [44]. The categorization of acute toxicity levels of concern are High, Moderate, Low and none. The LC50 value ranges for these different categories are show in Table 1 along with the number of chemicals that fall within each category for the fathead minnow.

Based on these categories the chemicals were separated into two classes of toxic and nontoxic chemicals. Chemicals were considered toxic if their LC50/EC50 values were less that 100 mg L^{-1} while nontoxic compounds had LC50/EC50 values either equal to this value or above this value. In this study we only considered a simple binary classification scheme rather than a more complicated multi-classification scheme.

Machine learning approach

Feature selection was accomplished by a pair-wise SVM Wrapper Elimination method. The procedure is explained in the next section. Support vector classification was used to generate the final hyper planes shown in the results section. In both cases the LIBSVM package implementation in R for support vector machines (SVM) developed by Chang and Lin was used [45]. There are many books explaining the theory and implementation involved in support vector machines, thus we will only briefly explain the C-classification support vector approach implemented in the LIBSVM package in R [45-49].

Category	EPA concern Level	Ranges of LC50/ EC50 (mg L^{-1}) values	Number of compounds
			Fathead minnow (96-h)
1	High	0-1	72
2	Moderate	1-100	333
3	Low	100-500	92
4	None	500>	73
		Total	570

Table 1: EPA categorization of acute toxicity showing levels of concern, ranges for each category and number of chemicals from the fathead minnow dataset that fall into each category [44].

In order to validate our model a decisions tree was used. The rpart package implementation in R for decision trees developed by Therneau and Atkinson was used. Since there are numerous books explaining the theory and implementation involved in CART (Classification and Regression Trees) and was only implement for the validation of our model, therefore we will not explain it here [50-52].

Given a training set of (x_i,y_i), with x_i being the input vectors or d-descriptors and y_i the class labels, the optimization problem in the dual form of C-classification can be written as:

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ &0 \le \alpha_i \le C, \quad i = 1,\ldots,l, \\ &\qquad y^T\alpha = 0, \end{aligned} \tag{1}$$

Where α are Lagrange multipliers, e is a vector of all ones, C is the regularization parameter, Q is an l by l positive semidefinite matrix, $Q_{ij}=y_iy_jk(x_i,x_j)$ and $k(x_i,x_j)$ is the kernel function, while C is the regularization parameter. Essentially C defines the trade-off between a large margin and misclassification error. The most widely used Kernel functions and the ones included in the LIBSVM program are the following:

$$\begin{aligned} &-\ \text{linear}: k\left(x_i, x_j\right) = x_i^T x_j \\ &-\ \text{polynomial}: \ k\left(x_i, x_j\right) = \left(x_i^T x_j + 1\right)^d \\ &-\ \text{radial basis}: \ k\left(x_i, x_j\right) = \exp\left(-\gamma\left|x_i - x_j\right|^2\right) \end{aligned} \tag{2}$$

The last two kernel functions preform a nonlinear mapping from the original data to some higher dimensional feature space. These two kernels are used to handle nonlinear or more complex relationships. The parameters C, d, and γ that appear in eqns. (1) and (2) are user-defined parameters, which need to be optimized prior to defining the optimal hyper plane for a data set. d and γ are kernel parameter and ultimately controls the curvature of the hyper plane. In other words, these parameters control the nonlinearity of the hyper plane. These parameters along with the C parameter control the complexity of the hyper plane and improper optimization of these parameters can lead to over fitting or under fitting. Support vector classification approaches rely on the maximum margin principle and try to determine a separating hyper plane with maximal distance between the two classes. In this approach most of the Lagrange multipliers are zero, the points or support vectors, which contribute to the hyper plane have nonzero values. For simple linear separable data, the support vectors are close to the hyper plane, which for two features is a line and are within a certain distance from that line. The hyper plane obtained from solving the optimization problem in eqn. (1) is the following:

$$f\left(x\right) = \sum_{i=1}^{n} \alpha_i k(x_i, x_j) + b \tag{3}$$

Where the sum runs over support vectors while (x_i,x_j) is a set of example data point and b is the bias. Consequently, for binary classification the class a point belongs to only depends on the sign of eqn. (3) while zero corresponds to a point on the hyper plane.

When applying support vector classification to any set of data the regularization constant C and kernel parameters (d and γ) have to be optimized prior to support vector classification. The easiest, however most computational expeNnive approach is to search a grid of parameters to find the optimal parameters that minimize the error of misclassification. In this work a grid search was used to find the

optimal parameters in all cases [45,47]. The radial basis function kernel was used in the feature selection process and in the determination of the hyper plane. In selecting the parameters for the feature selection process and hyper plane a set of discrete values, which are factors of 2 were chosen for the grid search. In this work an initial grid search was done to determine C the regularization constant for powers of 2 ranging from 2^{-10} to 2^{10}, while for the γ parameter in the radial basis function kernel we considered powers of 2 ranging from 2^{-10} to 2^{10}. These ranges of parameters provided the least amount of errors, best performance and resulted in a simple interpretable hyper plane.

Feature selection

Feature selection or minimizing the number of descriptors that define the physical properties associated with a toxic end point is very important in statistical analysis. Irrelevant descriptors are eliminated with the object to avoid over fitting, improve model performance and cost-effectiveness, and gain a deeper insight into the underlying process [53]. In this study a pair-wise SVM Wrapper Elimination method was used. The pair-wise SVM Wrapper Elimination method was used to find the best pair of descriptors from which we could determine the projected hyper plane and create a "rule of thumb" for acute aquatic toxicity. As previous show by Voutchkova et al. only two properties where used to design guidelines for acute aquatic toxicity [32]. In this paper we are interested in comparing their results with our results obtained by a more sophisticated machine learning approach. Furthermore, we wanted to show how a SVM Wrapper Feature Elimination approach can be applied to acute aquatic toxicity data and used in general for feature selection of data.

Initial preprocessing of the data was accomplished prior to the pair-wise SVM Wrapper Elimination methods. First any descriptors that had identical values for 80% of the features were removed. Also any descriptors that had a standard deviation less than 0.05 were eliminated. In both cases these were used to remove any descriptors that lack variability. The Pearson correlation coefficient was calculated for all the pairs of descriptors and one of any of the two descriptors was removed that had an absolute value of 0.95 or above. This was done to remove any redundant features that existed. A Pearson correlation coefficient, which is 0.95 or higher between two features, signifies that the two descriptors contain the same information. The descriptors number of atoms in 5- or 6-membered rings (in56), number of heavy atoms (nonhydrogen atoms, nonHatom), Highest Occupied Molecular Orbital (HOMO_AM1), electron affinity (EA. AM1) were eliminated based on their person correlation coefficient with the following descriptors number of atoms in rings (ringatoms), predicted polarizability (QPpolrz), ionization energy (IA.MA1), and Lowest Unoccupied Molecular Orbital (LUMO_AM1) respectively, which were not eliminated. Furthermore, the descriptors QPlogBB predicted brain/blood partition coefficient and QPlogKp predicted skin permeability, since these quantities are related to drug mechanisms in human's not aquatic life.

Support vector machine Wrapper Feature Elimination methods rank a subsets m features from a total of n features (m<n) based on their accuracy. This approach returns the best subset of features that gives the best overall accuracy and hence best prediction of the classifier. This method of evaluating the m features with an SVM is called an intensive "wrapper" feature selection approach which examines all combinations of r features or less and determines the combination that yields the best classification performance. Most Wrapper Feature Elimination methods try to explore all possible subsets of the feature, however we only considered pairs (r=2) since this was more tractable and showed to have good predictive power. Also we were interested in providing guidelines for toxic chemicals based on the projected hyper planes obtained by SVM, which can easily be viewed in 2D. The basic approach we took for a subset of m features is as follows:

Train a non-linear SVM (including Optimal Parameter grid search and 5-fold cross validation) for each individual feature with the classifier (-1 toxic and +1 non-toxic) and eliminate the features that under preform within a certain threshold (not highly correlated with classifier).

1. Train a non-linear SVM (including Optimal Parameter grid search and 5 fold cross validation) for all the possible subset of m features from n features ($_mC_n$ possible combinations).
2. Keep the subset of m features that gives the best accuracy for both classes.

The above procedure was run 25 times with 5-fold cross-validation for m=2 (best combination of two features). The best performance of the models was determined based on the Matthews correlation coefficient MCC as such:

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

Where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively. These two machine learning qualities of a binary classification take into account true and false positives, and true and false negatives creating a well balanced overall measurement even if the model only has a strong accuracy for one class or the classes are unbalanced. This method of evaluating the m features with an SVM is similar to a so called intensive "wrapper" feature selection approach which examines all combinations of r features or less and determines the combination that yields the best classification performance. In this case the intensive combinatorial approach involves a subset of r=2 features. Previous to examining the subsets of r=2 an initial elimination of features with r=1 was used (step 1 above), which eliminated any feature that had a Matthews correlation coefficient less that 0.15. This process helped initially eliminate any features that were not correlated with the classifier. Intensive "wrapper" feature selection approach has been shown to be a reliable method for feature selection [54]. For r=2 the most prevalent pairs of features were examined further and various hyper planes were examined, which can be used to describes the boundary between chemical properties for toxic and nontoxic chemicals and hence used as a rule to design less toxic chemicals. The explanation of how the hyper planes where determined is explained in the results and discussion.

Results and Discussion

The majority of QSAR models used to predict acute aquatic toxicity are based primarily on "a prior" classification of chemicals by their MOA. However, as mentioned in the introduction Machine learning approaches are independent of the mode of action [21,22]. Therefore, a mechanistic exploration of the data was not carried out and the molecules were not grouped by their mode of action. The only separation of the data that was accomplished was based on whether the chemicals were considered nontoxic or toxic by the criterion mentioned above. Furthermore, previous QSAR studies on acute aquatic toxicity were concerned with the prediction of toxicity and not concerned with the rational design of chemicals with reduced toxicity. The analysis that follows shows how support vector machines can be used to design chemicals with reduced acute aquatic toxicity.

The first part of the feature selection process was to test the best two features or best pairs of features. The following Table 2 shows the top 3 pairs of features that came out of this analysis, which had the overall best balance between their positive and negative rates.

The above descriptors are all related to a quantum mechanic property and solubility property. The quantum mechanical properties are: dE difference between the Lowest Unoccupied Molecular Orbital (LUMO) and Highest Occupied Molecular Orbital (HOMO) and LUMO_AM1 Lowest Unoccupied Molecular Orbital, while the solubility properties are: PSA Van der Waals surface area of polar nitrogen and oxygen atoms, QPlogS predicted aqueous solubility, and QPlogPo.w predicted octanol/water partition coefficient. In Figure 1 box plots of each of the normalized features separated based on whether they are toxic or nontoxic is displayed.

As one can see all of the features except the PSA feature partition between toxic and nontoxic along the number line. The interquartile range for the QPlogPo.w and QPlogS are perfectly partitioned between toxic and nontoxic. Furthermore, in Figure 2 we can see a similar behavior when we examine the frequency distributions for the toxic and nontoxic chemicals for each feature.

Above support vector machines were then used to determine the cut-offs or hyper plane that segregates the toxic from the nontoxic for these two properties. Prior to obtaining the optimal hyper planes for each run a grid search for the best parameter for the radial basis kernel function and the regularization constant C were performed by 5 fold cross validation. Five-fold cross validation was also used to obtain the final hyper plane with the optimized model parameters. In Figure 3 the optimal projected hyper planes for each of the pairs is shown.

As one can see nontoxic chemicals (class +1) have high aqueous solubility QPlogS or low octonal-water partition coefficient and high LUMO energies or large energy gaps between the LUMO and HOMO. The projected hyper plane for the PSA Van der Waals surface area of polar nitrogen and oxygen atoms pair does not have a clear separating projected hyper plane. One can see that the clustered nontoxic regions are in areas where dE is large and PSA is small. This pair of features does not provide us with a clear separating projected hyper plane to rationally design chemicals with reduced toxicity. From these graphs we can also obtain rough cutoff values for the features that give us nontoxic chemicals. For the lowest unoccupied molecular orbital (LUMO) vs. aqueous solubility (QPlogS) these cut-off values roughly correspond to QPlogS>-1 and LUMO>1, while for Octonal-Water Partition Coefficient (QPlogo.w) vs. difference between the LUMO and HOMO (dE) they roughly correspond to QPlogo.w<1 and dE>9 (Figure 4).

The prevalence of these properties can also be rationalized mechanistically. How these properties are relevant to acute toxicity has been described previously [32,55-61]. As mentioned in the introduction many of the initial classical QSAR models for acute aquatic toxicity were based on octonal-water coefficient. They can be easily rationalized since the value of a compounds octonal-water coefficient provides information on a chemicals ability to enter cells through lipid membranes. In turn the octonal-water coefficient of a compound can be related to the ability of a chemical to enter a fish through their gills [60]. A compounds octonal-water coefficient can also be related to how bioavailable the compounds will be to fish, since a low octonal-water coefficient means the compounds are more water-soluble and thus less bioavailable to fish compared to more lipid-soluble compounds with high octonal-water coefficient. This same rationalization also

Acc. Positive Samples	Acc. Negative Samples	MCC	Descriptor 1	Descriptor 2
0.97017	0.96988	0.99505	PSA	dE
0.70609	0.71687	0.95050	QPlogS	LUMO_AM1
0.60772	0.61446	0.94059	QPlogPo.w	dE

Table 2: Matthews Correlation Coefficient (MCC), the accuracies of the negative and positive samples for the three most relevant pairs of features from the fathead minnow data set.

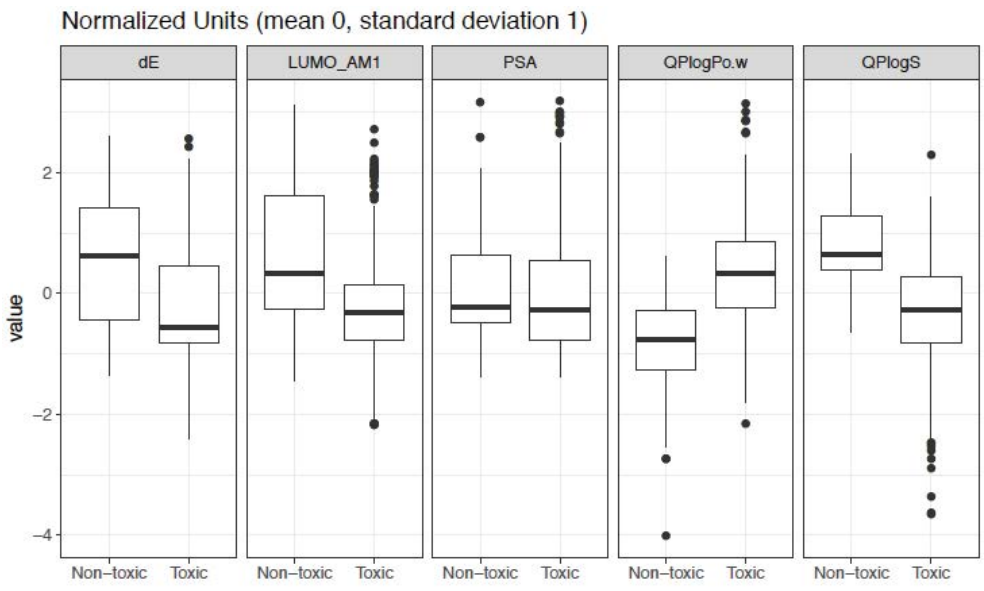

Figure 1: Box plots for the individual features obtained from the best pairs of features separated by whether they are toxic or nontoxic.

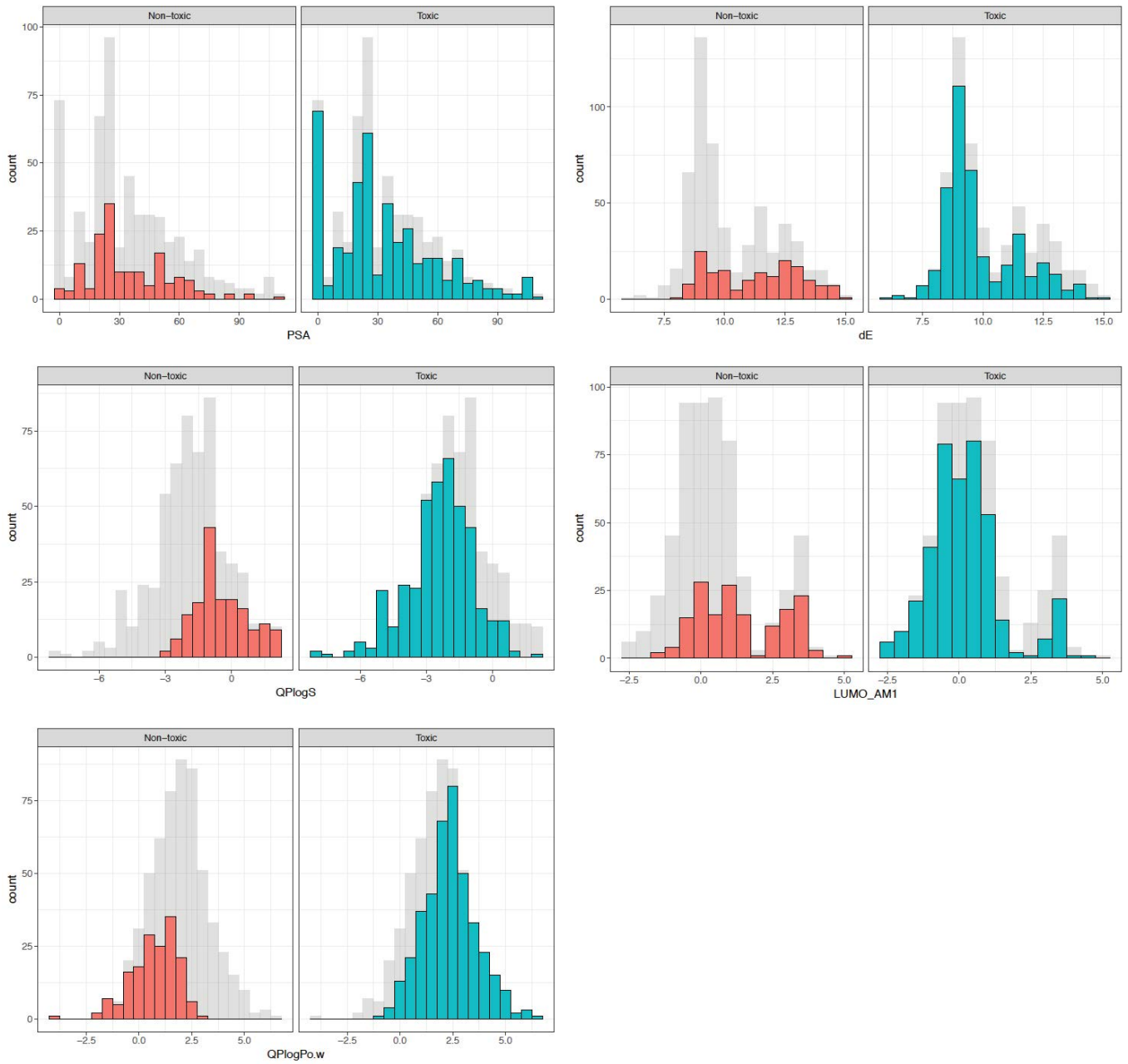

Figure 2: Frequency distributions for the individual features obtained from the best pairs of features separated by whether they are toxic or nontoxic.

applies to molecules having low aqueous solubilities. The surface area of polar nitrogen and polar oxygen (PSA) has been shown is a simple measure of hydrogen-bonding capacity and related to the energy involved in the membrane transport of a compound [62]. The energy involved in membrane transportation can be physically explained by the fact that polar groups are involved in desolation when they move from an aqueous environment to more lipophilic environment. Thus, a molecule with a small PSA will be more lipid-soluble and hence more bioavailable to fish. However, in the hyper plane above for the dE vs. PSA the nontoxic chemicals are clustered together for small PSA. Mechanistically this does not make sense, but as mentioned above there is not a clear separating plane for these two pairs of features and there is some clustering of the nontoxic chemicals for large PSA. For this pair of features the large differences between the LUMO and HOMO account for a chemicals non toxicity while the PSA can either be small or large. Furthermore, previous studies have shown that compounds with high octonal-water coefficient are more bioaccumulative [57,58]. The aqueous solubility and the polarity of molecule can be mechanistically related to how bioavailable the compounds will be to fish since molecules with high aqueous solubilities and high polarity will be more bioavailable to fish compared to compounds with smaller aqueous solubilities and small polarity. The dE on the other hand can be related mechanistically to acute aquatic toxicity by reactivity at site of action. A large HOMO-LUMO gap or difference in energy between the HOMO and LUMO implies high stability for a compound in the sense that its reactivity in chemical reactions is low. The reactivity at

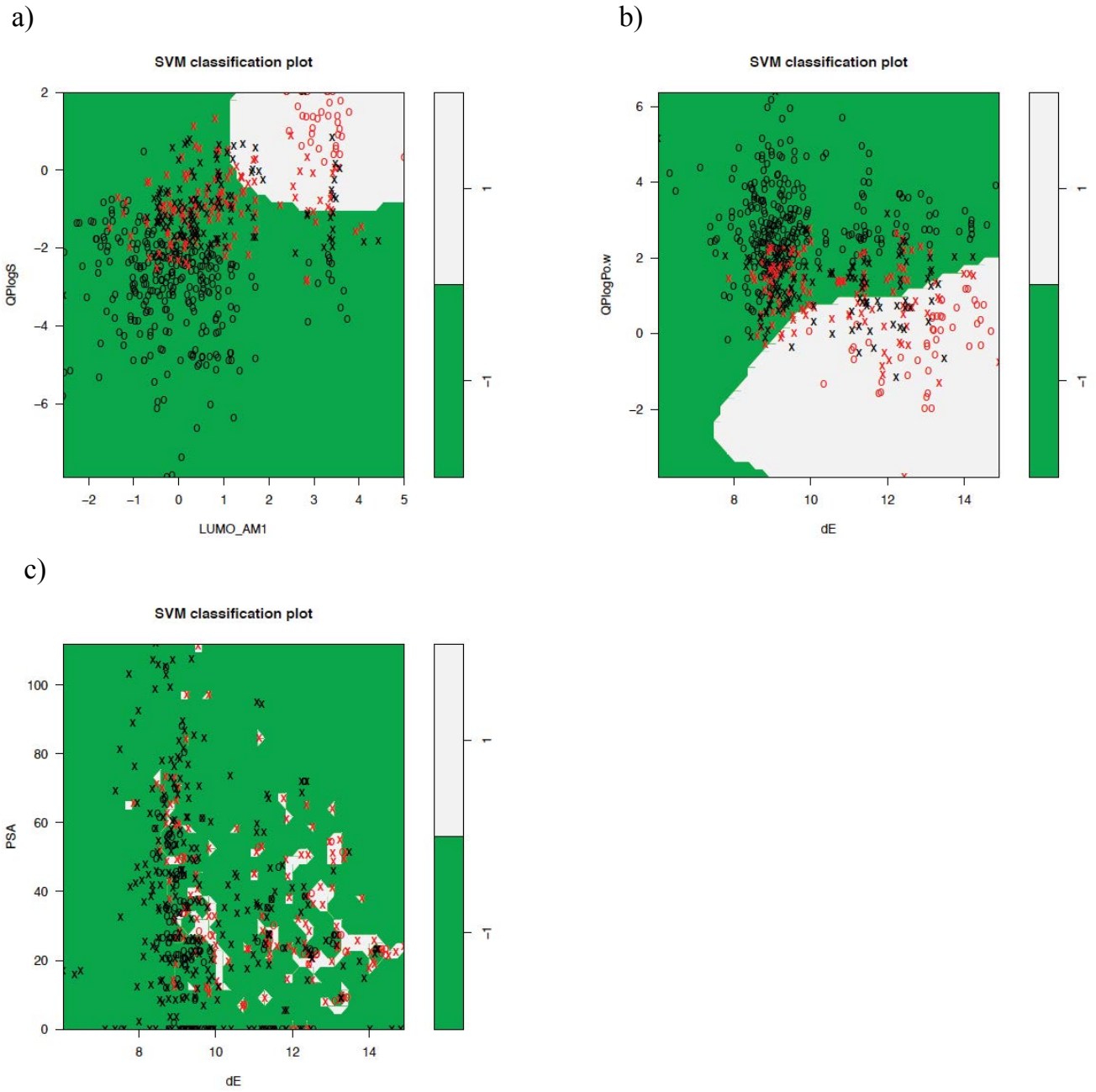

Figure 3: Hyper planes for the top three pairs of features for a) Lowest Unoccupied Molecular Orbital (LUMO) and Aqueous Solubility (QPlogS) b) Difference between the LUMO and HOMO (dE) and Octonal-Water Partition Coefficient (QPlogo.w) c) Difference between the LUMO and HOMO (dE) and Van der Waals surface area of polar nitrogen and oxygen atoms (PSA). The green regions correspond to toxic chemicals while white regions correspond to nontoxic chemicals. The original fathead minnow data points corresponding to toxic chemicals (black points) and non-toxic chemicals (red points) are also displayed in the graph above.

the site of action has been associated with frontier orbitals energies and in particular with the lowest unoccupied molecular orbital (LUMO) and the difference between the LUMO and HOMO previously [60]. Therefore, it is not surprising that once again the emerge as the dominant statistical descriptors for acute aquatic toxicity.

In order to further demonstrate that these are the best pairs of features to consider for acute aquatic toxicity we use a decision tree to see if we obtained comparable accuracies. The results of this analysis are show in Table 3.

As one can see the results of the design tree analysis are in some cases slightly less accurate than the support vector results. However, the results are comparable to the support vector results. The decision trees that were created for each pair is shown in Figure 2.

As one can see the decision tree for the pairs QPlogS/LUMO and QPlogPo.w/dE are much simpler as compared to the PSA/dE pair. This further shows that although the pair PSA/dE is the most accurate for the SVM model, it yields a fairly complicated decision tree and projected

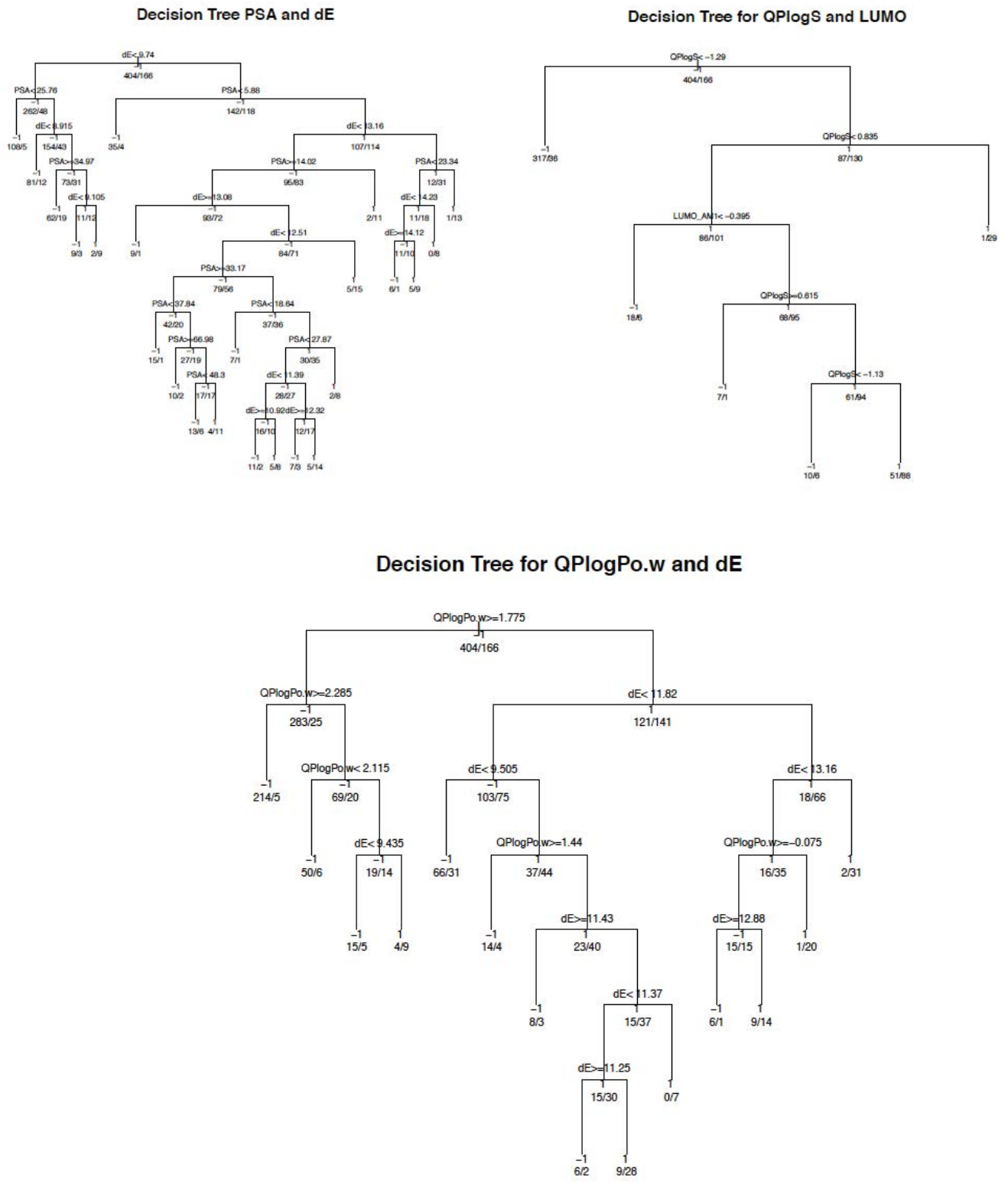

Figure 4: Constructed decision trees for the best pairs of features a) PSA and dE b) QPlogS and LUMO c) QPlogPo.w and dE.

hyper plane and cannot be used to rationally design chemicals with reduced toxicity.

Golbraikh and Tropsha have established that the only way to truly know that your model is reliable is through external validation [63]. In external validation, the data set used for validation should not be used in the training part of the model. Therefore, in this study organic compounds that were not included in the fathead minnow data set that only existed in the Japanese Medaka, *Daphnia magna* and

Acc. Positive Samples	Acc. Negative Samples	MCC	Descriptor 1	Descriptor 2
0.63855	0.92327	0.59736	PSA	dE
0.70482	0.87129	0.57310	QPlogS	LUMO_AM1
0.65663	0.93812	0.63721	QPlogPo.w	dE

Table 3: Matthews Correlation Coefficient (MCC), the accuracies of the negative and positive samples for the three most relevant pairs of features from the fathead minnow data set using a decision tree.

Species	MCC	Acc. Negative samples	Acc. Positive samples		
Pseudokirchneriella subcapitata	0.77502	0.991372	0.71429	PSA	dE
Daphnia magna	0.54006	0.95507	0.52976	PSA	dE
Oryzias latipes (Japanese Medaka)	0.35388	0.99953	0.35000	PSA	dE
Pseudokirchneriella subcapitata	0.80043	0.98792	0.78571	QPlogS	LUMO_AM1
Daphnia magna	0.68970	0.97853	0.63690	QPlogS	LUMO_AM1
Oryzias latipes (Japanese Medaka)	0.85771	0.99624	0.81667	QPlogS	LUMO_AM1
Pseudokirchneriella subcapitata	0.79491	0.99137	0.74490	QPlogPo.w	dE
Daphnia magna	0.68594	0.91808	0.76190	QPlogPo.w	dE
Oryzias latipes (Japanese Medaka)	0.58518	0.99857	0.566666	QPlogPo.w	dE

Table 4: Matthews Correlation Coefficient (MCC), the accuracies of the negative and positive samples for the three different aquatic species and three different pairs of features used to predict the hyper planes (Figure 1) obtained from the fathead minnow data set.

Pseudokirchneriella subcapitata datasets were used as validation sets. Another requirement for a good external validation set is determining whether the validation data set is in the applicability domain of the model. The applicability domain for a QSAR study refers to the scope of the model and whether it is appropriate to make predictions for a given query of chemicals. A modeling method cannot be expected to yield reliable results for chemical structures significantly different from the training compounds. Applicability domains can be based on the training set coverage in the models descriptor space, mechanisms and structural features just to mention a few. Interpolation rather than extrapolation within these domain or conditions should be used to make predictions [64]. Applicability domain based on training set coverage in the models descriptor space assessing whether the validation set or test set of chemicals fall in the relative space covered by the models training set feature space. Mathematically this range based approach estimates interpolation regions in a multivariate space. However, in this study we only consider two properties and therefore do not have to concern ourselves with a complex multivariate space. Furthermore, the validation data sets used in this paper are within the applicability domain, since they correspond to other organic compounds and the corresponding features of the compounds fall with the training set feature space. The overall performance of the different external validation data sets for the three different species and three different pairs of descriptors obtained from the hyper planes that appears in Figure 1 is shown in Table 4.

From the table above we can see that the hyper plane or curve predicted from the fathead minnow data sets suitably predictions the acute aquatic toxicity of the three species. However, the predictions obtained for the acute aquatic toxicity of the *Oryzias latipes* (Japanese Medaka) and *Daphnia magna* are fairly bad in regards to the MCC and accuracy of the positive examples if we considered all of the data. This is primarily due to the fact that the number of nontoxic chemicals that exist in this dataset was very small. Only 6 for the Japanese Medaka and 28 for the *Daphnia magna* compounds are considered nontoxic from a total of 270 and 315, respectively. Therefore, we considered the 6 nontoxic chemicals for the Japanese Medaka and randomly chose 71 other toxic chemicals, while for *Daphnia magna* we considered the 28 nontoxic chemicals and randomly chose 73 toxic chemicals for the analysis above. The randomly choosing of toxic chemicals was done 30 times and the average values are reported in the table above. The results for these created validation set show that the above analysis for the fathead minnow provides suitable predicts for the acute aquatic toxicity of the *Daphnia magna* and *Pseudokirchneriella subcapitata*. However, the predictions obtained for the acute aquatic toxicity of the *Oryzias latipes* (Japanese Medaka) are not as good as the results for the other species even with randomly choosing toxic chemicals to consider in the analysis. This is due to the fact that the number of nontoxic chemicals that exist in this dataset is only 6. Therefore, such measurements as the MCC and accuracy of the positive examples will not compare as well with the other species.

Conclusion

In this paper a support vector machine pairwise recursive feature extraction (RFE) method along with some initial preprocessing of the data was used to determine the two most relevant features for acute aquatic toxicity. The three most relevant pairs of features and optimal model parameters for this toxic end point were obtained based on 5-fold cross validation. A support vector classification model was then developed with these pairs of properties to determine the optimal dividing curves or projected hyper planes, which separates the toxic from the nontoxic. Validation of our model was done both by external validation and a decision tree model. The U.S.-E.P.A. Duluth Fathead Minnow database was used as the training set while a set of unique compounds not present in the fathead minnow dataset for the Japanese Medaka, *Daphnia magna* and *Pseudokirchneriella subcapitata* obtained from the Japanese ministry were used as a validation data set. This more sophisticated approach of using a machine learning technique to derive properties guidelines for designing safer chemicals yielded results similar to those obtained by Voutchkova et al. for acute aquatic toxicity. Chemicals of little concern (nontoxic chemicals) in regards to acute aquatic toxicity obtained from this model for the fathead minnow data set were determined to be for compounds with aqueous solubility QPlogS>-1 and lowest unoccupied molecular orbital LUMO>1, or Octonal-Water Partition Coefficient QPlogo.w<1 and difference between the LUMO and HOMO dE>9. These limits obtained from the fathead minnow data set are reasonably guidelines for other aquatic species, which is exemplified by the high performance accuracies that where obtained for the validation sets that, included three different aquatic species. This paper shows how effectively machine learning techniques, specifically support vector machines can be used to derive design guidelines for chemicals with reduced acute aquatic toxicity. In

future work we hope to apply this approach to other toxic endpoints and classes of chemicals along with a hyper plane for three features to obtain limits in 3D.

References

1. Binetti R (2008) Francesca Marina Costamagna, Ida Marcello. Exponential growth of new chemicals and evolution of information relevant to risk control. Ann Ist Super Sanità 44: 13-15.
2. U.S. EPA, http://www.epa.gov/tri/nationalanalysis/overview/2010TRINAOverview.pdf
3. Toussaint MW, Shedd TR, Van der Schalie WH, Leather GR (1995) A comparison of standard acute toxicity tests with rapid-screening toxicity tests. EnViron Toxicol Chem 14: 907-915.
4. European Union (2006) Corrigendum to Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC (OJ L 396, 30.12.2006). Off. J. Eur. Union 2007, L136, 50.
5. Registrations, Evaluation, Authorization and Restriction of Chemicals (REACH). http://ecb.jrc.it/reach/reach-legislation/ (accessed September 20, 2009).
6. Cronin MTD, Walker JD, Jaworska JS, Comber MHI, Watts CD, et al. (2003) Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances, Environ. Health Perspect 111: 1376-1390.
7. Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA (1997) Environmental Toxicology and Chemistry 16: 948-967.
8. Ministry of Environment, Japan: Ecotoxicity Test Results, http://www.safe.nite.go.jp/english/ sougou/doc/html/select_help_e.html#env_tox.
9. Russom CL, Bradbury SP, Broderius SJ, Drummond RA, Hammermeister DE (1997) Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas). Environ Toxicol Chem 16: 948-967.
10. Winkler DA (2004) Neural Networks as Robust Tools in Drug Lead Discovery and Development. Mol Biotechnol 27: 139-167.
11. Louis B, Agrawal VK, Khadikar PV (2010) Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. European Journal of Medicinal Chemistry.
12. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, et al. (2004) Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. J Chem Inf Comput Sci 44: 1257-1266.
13. Escher BI, Hermens JLM (2002) Modes of Action in Ecotoxicology: Their Role in Body Burdens, Species Sensitivity, QSARs, and Mixture Effects. EnViron Sci Technol 36: 4201-4217.
14. Bradbury SP (1994) Predicting modes of toxic action from chemical structure: An overview. SAR QSAR EnViron Res 2: 89-104.
15. Verhaar HJM, Van Leeuwen CJ, Hermens JLM (1992) Chemosphere 25: 471-479.
16. Verhaar HJM, Solbé J, Speksnijder J, Van Leeuwen CJ, Hermens JLM (2000) Chemosphere 40: 875-883.
17. Du QS, Huang RB, Wei YT, Pang ZW, Du LQ, et al. (2008) J Comput Chem 30: 295-304.
18. Du QS, Huang RB, Wei YT, Du LQ, Chou KC (2008) J Comput Chem 29: 211-219.
19. Escher BI, Hermens JLM (2002) Environ Sci Technol 36: 4201-4217.
20. Bradbury SP (1994) SAR QSAR Environ Res 2: 89-104.
21. Ivanciuc O (2002) Internet Electron J Mol Des 1: 157-172.
22. Ivanciuc O (2004) Internet Electron J Mol Des 3: 802-821
23. Schultz TW, Cronin MTD, Walker JD, Aptula AO (2003) Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective. J Mol Struct Theochem 3622: 1-22.
24. Posthumus R, Slooff W (2001) Implementation of QSARs in ecotoxicological risk assessments; RIVM Report 601516003.
25. Dearden JC (2002) Prediction of Environmental Toxicity and Fate Using Quantitative Structure-Activity Relationships (QSARs). J Braz Chem Soc 13: 754-762.
26. Schultz TW, Cronin MTD, Netzeva TI (2003) The present status of QSAR in toxicology. J Mol Struct Theochem 622: 23-38.
27. Cronin MTD, Dearden JC (1995) QSAR in Toxicology. 1. Prediction of Aquatic Toxicity. Quant Struct.-Act Rel 14: 1-7.
28. Papa E, Villa F, Gramatica P (2005) Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow). J Chem Inf Model 45: 1256-1266.
29. Mazzatorta P, Benfenati E, Neagu CD, Gini G (2003) J Chem Inf Comput Sci 43: 513-518.
30. Niculescu SP, Atkinson A, Hammond G, Lewis M (2004) SAR QSAR Environ Res 15: 293-309.
31. Tan NX, Li P, Rao HB, Li ZR, Li XY (2010) Prediction of the acute toxicity of chemical compounds to the fathead minnow by machine learning approaches. Chemo metrics and Intelligent Laboratory Systems 100 66-73.
32. Voutchkova AM, Kostal J, Steinfeld JB, Emerson JW, Brooks BW, et al. (2011) Towards rational molecular design: derivation of property guidelines for reduced acute aquatic toxicity. Green Chem 13: 2373.
33. Lai DY, Woo YT, Argus MF, Arcos JC (1996) Designing Safer Chemicals. In: DeVito S, Garrett R (eds.) Americal Chemical Society 62-73.
34. Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA (1997) Environmental Toxicology and Chemistry 16: 948-967.
35. http://www.epa.gov/med/Prods_Pubs/fathead_minnow.htm
36. Ministry of Environment, Ecotoxicity Test Results, Japan.
37. Chou KC, Zhang CT (1995) Crit Rev Biochem Mol 30: 275-349.
38. Cai YD (2001) Is it a paradox or misinterpretation? Proteins: Structure. Function and Genetics 43: 336-338.
39. Chou KC, Shen HB Nat Protoc 3: 153-162.
40. Chou KC, Shen HB (2007) Anal. Biochem. 370: 1-16.
41. Babel OS (2009) File Format Converter.
42. Jorgensen WL (2003) QikProp version 3.0, Schrodinger, LLC, New York.
43. Ioakimidis L, Thoukydidis L, Mirza A, Naeem S, Reynisson J (2008) QSAR Comb Sci 27: 445-456.
44. U.S. EPA, Ecotoxicity Categories for Terrestrial and Aquatic Organisms.
45. Chang CC, Lin CJ, LIBSVM (2001) A library for support vector machines.
46. Scholkopf B, Smola AJ (2002) Learning with Kernels. MIT Press B, Cambridge.
47. Cristianini N, Shawe-Taylor J (2003) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press 2003.
48. Vapnik V (1995) The Nature of Statistical Learning Theory, Springer, New York.
49. Vapnik VN (1998) Statistical Learning Theory, Wiley-Interscience, New York.
50. Therneau TM, Atkinson EJ (1997) An introduction to recursive partitioning using the rpart routines. Divsion of Biostatistics 61, Mayo Clinic.
51. Therneau TM (1983) A short introduction to recursive partitioning. Orion Technical Report 21, Stanford University, Department of Statistics.
52. Breiman L, Friedman JH, Olshen RA, Stone CJ (1983) Classification and Regression Trees. Wadsworth, Belmont.
53. Guyon I, Elisseeff A (2003) Journal of Machine Learning Research 3: 1157-1182.
54. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46: 389-422.

55. Arnot JA, Gobas FAPC (2006) Environ Rev 14: 257-297.

56. Arnot JA, Gobas FAPC (2003) QSAR Comb Sci 22: 337-345.

57. Zvinavashe E, Murk AJ, Vervoort J, Soffers AEME, Freidig A, et al. (2006) Environ Toxicol Chem 25: 2313-2321.

58. Bearden AP, Schultz TW (1997) Environ Toxicol Chem 16: 1311-1317.

59. Eriksson L, Verhaar HJM, Hermens JLM (1994) EnvironToxicol Chem 13: 683-691.

60. Zvinavashe E, Du TT, Griff T, van den Berg HHJ, Soffers AEMF, et al. (2009) Chemosphere 75: 1531-1538.

61. Mckim JM, Schmieder PK, Erickson RJ (1986) Aquat Toxicol 9: 59-80.

62. Waterbeemd HV (2007) Comprehensive Medical Chemistry II 5: 669-697.

63. Tropsha A, Gramatica P, Gombar VK (2003) QSAR Comb Sci 22: 69-77.

64. Nikolova-Jeliazkova N, Jaworska J (2005) Altern Lab Anim 33: 461.