Available online at www.sciencedirect.com
ScienceDirect
Chemosphere 71 (2008) 1225–1232
CHEMOSPHERE
www.elsevier.com/locate/chemosphere

An evaluation of global QSAR models for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*

S.J. Enoch [a], M.T.D. Cronin [a], T.W. Schultz [b], J.C. Madden [a,*]

[a] *School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, England, United Kingdom*
[b] *The University of Tennessee, College of Veterinary Medicine, 2407 River Drive, Knoxville, TN 379961-4543, USA*

Received 26 June 2007; received in revised form 16 November 2007; accepted 7 December 2007
Available online 7 February 2008

Abstract

This study presents an analysis of the ability of a two-parameter response surface, a multiple linear regression and a neural network model to produce global quantitative structure–activity relationships (QSARs) to predict the toxic potency of phenols to *Tetrahymena pyriformis*. The phenolic toxicity data set analysed is characterised by multiple mechanisms of toxic action. The study aimed to evaluate the confidence that can be applied to the modelling of the differing mechanisms of action. Assessment of confidence was decided in terms of whether the statistics for the global models reflect the ability of the QSARs to model the individual mechanisms of toxic action present in the data set. The results showed that the global statistics only reflected the ability of models to predict the two non-covalent mechanisms (polar narcosis and respiratory uncoupling), with the metabolically transformed and electrophilic mechanism (pre-electrophiles and soft electrophiles) being modelled poorly by all three model building methods. The results confirm the difficulty in modelling electrophilic mechanisms of toxic action. The results also highlight the fact that this poor predictivity is often 'hidden' in good statistical fit of some global models. In particular these results emphasise that for practical predictive purposes the mechanistic applicability domain is required to give confidence to estimated toxicity values.

Keywords: Toxicity; QSAR; Evaluation; *Tetrahymena pyriformis*; Phenol; Mechanism of action

1. Introduction

It is envisaged that *in silico* methods, and (quantitative) structure–activity relationships ((Q)SARs) in particular, will play an important role in the reduction of animal testing required for the risk assessment of chemical substances, especially under new legislation such as REACH (Worth et al., 2007). These models will, by necessity, be developed utilising information from the numerous toxicological databases that are currently available (Cronin, 2005). Two approaches to modelling such databases have been suggested; the first of these is the development of "global" models which (in this study) are defined as QSAR models that cover a number of different mechanisms of action for a given toxicological endpoint. The use of the term global model in this study is distinct from that used to define QSAR models based on chemicals with similar modes of action allowing interspecies correlations (Dimitrov et al., 2003). The second is the development of a number of "local" models, each covering a single mechanism of action present in the database. Several publications have investigated the ability of QSAR approaches to fulfil regulatory requirements (Worth et al., 2007; Yuan et al., 2007), in addition the OECD principles for the validation of QSARs aim to offer guidance on this issue (OECD, 2004).

A fundamental requirement for QSAR model building is high quality data (Cronin, 2005). A good example of a toxicological database with high quality potency values is that for the ciliated protozoan *Tetrahymena pyriformis* for 166 phenols, first published in 1996 (Cronin and Schultz,

* Corresponding author. Tel.: +44 151 231 2032; fax: +44 151 231 2170.
E-mail address: j.madden@ljmu.ac.uk (J.C. Madden).

0045-6535/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chemosphere.2007.12.011

1996), and then later extended to 250 compounds (Cronin et al., 2002). Whilst limited to compounds containing a benzene ring with a hydroxy substituent, these data reflect the distributions of substituents observed in commercially available industrial organic substances. Mechanisms of action were assigned to each of the phenols in the database by Schultz et al. (1997) utilising a series of simple structural rules. Mechanisms were: polar narcosis, weak acid respiratory uncouplers, soft electrophiles, precursor to soft electrophiles (pre-electrophiles), and precursor to redox cyclers (pro-redox cyclers).

Analysis of the mechanistic makeup of the phenols database published by Cronin et al. in 2002 using the rules derived by Schultz et al. (1997) revealed that, as expected approximately 70% (173 chemicals out of 250) of the chemicals act mainly by narcosis (Bradbury and Lipnick, 1990). In addition, four other reactive electrophilic mechanisms were also identified, these being: pre-electrophiles (27 chemicals), pro-redox (4 chemicals), respiratory uncouplers (19 chemicals) and the generic category soft electrophiles (27 chemicals). These reactive chemicals exhibit (acute) toxicity in excess of that predicted by narcosis, primarily acting via non-reversible covalent interactions between the toxic chemical and biological macromolecules (Veith and Mekenyan, 1993; Cronin, 2003; Roberts et al., 2006). The make up of the phenol database in terms of the relative proportion of each type of mechanism of action is considered to be typical of both the chemical universe as a whole, and of that available for QSAR modelling.

A number of QSAR studies have utilised the high quality, single source toxic potency values of phenols to *T. pyriformis* (Cronin et al., 2002) to develop both local and global models. Two studies have shown the ability of the mechanistically interpretable 'response surface' analysis to model successfully a limited set of phenols acting by non-electrophilic mechanisms (Cronin and Schultz, 1996; Cronin et al., 2002). The response surface model was based on log P to model hydrophobicity and LUMO (energy of the lowest unoccupied molecular orbital) to model electrophilicity. These local models produced statistics of $r^2 = 0.90$ for a training set of 120 phenols and $r^2 = 0.81$ for a larger set of 160 phenols. In addition, studies utilising the extended 250 phenol data set in which five mechanisms of action have been identified (Schultz, 1997) have produced linear and non-linear QSAR models with varying degrees of quality (Cronin and Schultz, 1996; Devillers, 2004). These models utilised varying numbers and types of descriptors, with only limited improvement in the resulting statistical relationships. Example training statistics for the best linear and non-linear models from these studies produced models with $r^2 = 0.67$, and $r^2 = 0.82$, respectively, where the number of training chemicals was 200 in both cases (both models used the same set of descriptors). The latter results suggest that, as expected the non-linear methods may be better at formulating global models, however no diagnostic statistics were provided for the individual mechanisms within the global model.

Previous studies have identified that chemicals belonging to the pre-electrophile mechanism of action are frequently found to be statistical outliers to QSARs for acute eco-toxicity (Cronin and Schultz, 1996; Devillers, 2004). However, no study has investigated whether the global model statistics reported for these models applies equally to each of the mechanisms of action within the data set. Knowledge of whether individual mechanisms of action present in a global data set are being modelled as well as is suggested by the global statistics is important in being able to assign a confidence to a given prediction made by a QSAR model. In other words, when making a prediction of toxicity using a global model should one consider whether a molecule fits into a global or local applicability domain?

The aim of this study, therefore, was to investigate what level of confidence should be applied to predictions from each of the previously identified mechanisms of action for the database of 250 phenol toxicity values for *T. pyriformis*. The level of confidence was assessed for three common QSAR model building methods, namely a previously reported two-parameter response surface, a stepwise multiple linear regression and neural network approaches. The assessment of confidence was performed for individual mechanisms to determine the real success of global approaches for modelling.

2. Methods

2.1. Data set and toxicity values

A database of toxicity values for 250 phenolic compounds was acquired from the literature (Cronin et al., 2002). The database was split into training and validation sets (200 and 50 compounds, respectively) as described previously (Cronin et al., 2002). All models used in the study utilised the same 200 training and 50 validation chemicals. This database contained five mechanisms of action classified from a previous study (Schultz et al., 1997). The number of training and validation compounds, respectively, for each mechanism is shown in parentheses: weak acid respiratory uncouplers (15:4), soft electrophiles (22:5), pre-electrophiles (22:5), pro-redox cyclers (3:1) and polar narcotics (138:35). Global models were constructed using all 200 compounds and validated using a validation set of 50 compounds. The performance of each model was also assessed in terms of its ability to fit the training data, and predict the validation data for four of the five mechanisms of action identified. The exception to this was the pro-redox cycling mechanism for which insufficient compounds exist in the data set for such an analysis to be statistically relevant.

The toxicity values were obtained from a population growth impairment test utilising the freshwater ciliate *T. pyriformis* (strain GL-C), performed following the protocol previously described by Schultz (Schultz, 1997). The endpoint, population density, of this static 40-h assay was measured spectrophotometrically at 540 nm. Test conditions

allow for 8–9 cell control cultures. Compounds were tested in a range finder prior to testing in duplicate for three definitive replicates. Two controls, one without chemical but inoculated with *T. pyriformis*, and the other, a blank with neither test chemical or ciliates, were used to provide a measure of the test acceptability and as a basis for interpretation of the treatment data. Each definitive test replicate consisted of six to eight different concentrations with duplicate flasks of each concentration. Only replicates with control-absorbency values between 0.60 and 0.75 were used in the analysis.

The 50% growth inhibition concentration, IGC_{50} was determined for each compound using the probit analysis routine in the statistical analysis system (SAS) software (SAS Institute, Inc). All statistical analyses were performed on nominal concentrations; chemical analyses of concentrations were not performed.

2.2. Molecular descriptors

A total of 168 descriptors were calculated for each chemical, representing the physico-chemical, structural and topological properties that may be related to the toxicity of phenols to *T. pyriformis* (Table 1).

Logarithms of the 1-octanol/water partition coefficient (log *P*) and pK_a values were calculated using the ACD/Labs software (1995, Advanced Chemistry Development Inc, Toronto Canada). The distribution coefficient (log D) at pH 7.35 was calculated according to (1).

$$\log D = \log P - \log(1 + 10^{\mathrm{pH}-\mathrm{p}K_a}) \quad (1)$$

The remaining descriptors were calculated with Chem-X, version 2000.1 (Oxford Molecular Limited, Oxford, England), TSAR, version 3.3 (Oxford Molecular Limited, Oxford, England) and QSARis, version 1.1 (SciVision, Academic Press, San Diego, CA).

2.3. Linear regression models

The two-parameter response surface model based on log *P* and LUMO was used as described previously (Cronin et al., 2002). The stepwise multiple linear regression model was developed using the Minitab (version 14) statistical software. Forward stepwise regression was performed with descriptors requiring a Fisher statistic (*F*) value of 10 or greater to be considered for inclusion. Model quality was assessed for fit based on the coefficient of determination (r^2), the coefficient of determination adjusted for the number of degrees of freedom (r^2_{adj}), the leave-one-out cross-validated coefficient of determination (r^2_{cv}) and the root mean square error (RMSE). Predictivity of the model was assessed based on the external coefficient of determination (q^2_{ext}) and the RMSE for the 50 compound validation set.

2.4. Neural network model

Neural network analysis was carried out using Statistica V6.1 (StatSoft, Inc. (2004). Forward stepwise feature selection was performed using 'Feature Selection' algorithms available in the 'Neural Network' analysis tools. The forward stepwise feature selection efficiently trains a number of probabilistic and generalised regression neural networks enabling the most relevant descriptors to be selected for further neural network analysis. The number of selected descriptors was controlled by a sampling parameter, with a larger value resulting in fewer descriptors being included. This value is akin to the *F* statistic used to control the number of selected features in the stepwise multiple linear regression analysis. In order to reduce the descriptor pool sufficiently this values was set to 0.005. Upon completion the number of descriptors had been reduced from 168 to 20.

Neural network analysis was then carried out with the intelligent problem solver algorithms within Statistica

Table 1
Physico-chemical descriptors calculated for the chemicals in this study

Software	Descriptors
ACD labs	Logarithm of the octanol–water partition coefficient (log *P*), negative logarithm of the ionisation constant (pK_a), molar refractivity (MR), molar parachor (PAR), molar polarisability (POL), surface tension (ST)
MOPAC (Chem-X)	Energy of the highest occupied molecular orbital (HOMO), energy of the lowest unoccupied molecular orbital (LUMO), total nucleophilic (S^N) and electrophilic (S^E) superdelocalisability, maximum nucleophilic (S^N_O) and electrophilic superdelocalisability on the oxygen atom of the hydroxy group, maximum partial charge (Q_O) on the oxygen atom of the hydroxy group
Chem-X	Volume, enclosed by isopotential surface with electrostatic potential (EP): EP = −20 kcal/mol (EP_{M20}); EP = −10 kcal/mol (EP_{M10}); EP = 0 kcal/mol (EP_{ZERO}); EP = 10 kcal/mol (EP_{P10}); EP = 20 kcal/mol (EP_{20}). Coded by EP molecular VdW surface: EP > 10 kcal/mol (C_{POS}); EP < −10 kcal/mol (C_{NEG}); −10 kcal/mol < EP < 10 kcal/mol (C_{MID}). Coded by EP molecular VdW surface, in percents: P_{POS}, P_{NEG}, P_{MID}
TSAR	Molecular volume (MVol); molecular surface area (MSA); lipole, Kier simple and valence-corrected molecular connectivity indices: zero order, 2nd to 6th order path, 3rd order cluster, 4th order path-cluster; shape flexibility index; Kappa shape indices; Weiner, Randic and Balaban topological indices; number of H-bond donor (N_{HDON}) and acceptor (H_{ACC}) centres; atom counts (oxygen, nitrogen, fluorine, chlorine, bromine, iodine); group counts (methyl, amino, hydroxyl, nitro); molecular weight (MW); heat of formation (HF); ionisation potential (IP); total energy (E_{TOT}); dipole moment (μ)
QSARis	The sum of absolute values of the charges of a molecule, in electrons (ABSQ); sum of absolute charges on nitrogen and oxygen atoms in a molecule (ABSQon); ovality of a molecule; specific polarisability; magnitude of dipole moment (*P*), magnitude of principle quadrapole moment (*Q*); the largest positive charge on a hydrogen atom (MaxHp); the largest negative charge in a molecule (MaxNeg); the largest positive charge in a molecule (MaxQp); *E*-state indices; HE-state indices; dipolar descriptors

using log $(ICG_{50})^{-1}$ as the dependent variable and the 20 descriptors selected by the feature selection procedure as independent variables. The intelligent problem solver algorithm attempts to build the optimum neural network by training and validating a number of linear, radial basis function and three-layer perceptron neural networks. In this study the algorithm was allowed to run for 500 cycles, with a training set of 200 compounds (with 20 being utilised as a subset for the estimation of the training error). In addition the same 50 validation chemicals used for the linear regression models were used to assess external predictivity. Following completion of the 500 cycles 50 neural networks were retained for further analysis. These networks were selected to be diverse in terms of the type and architecture of the networks tested during the Intelligent Problem Solver routine. It was from these 50 networks that the final network was selected, this selection was based on the balance between training and validation error in conjunction with perceived model complexity, with simpler architectures being preferred.

The model quality of the final neural network was assessed for fit using the coefficient of determination (r^2) and RMSE calculated for the training data. Predictivity was assessed using the coefficient of determination (q^2_{ext}) and RMSE for the external validation set. Toxicity values, SMILES and descriptors used in the models are available from the corresponding author upon request.

3. Results

Toxicity data to *T. pyriformis* for 250 phenols were collected from a single literature data source known to be of high quality (Cronin et al., 2002). The data set consisted of a range of phenolic compounds varying in structure, reactivity, and hydrophobicity. Three global QSAR models were constructed using a variety of statistical approaches and modelling philosophies: a two-parameter response surface, a forward stepwise multiple linear regression and a neural network model.

3.1. Linear models

Eq. (2) shows the two-parameter response surface model (taken from Cronin et al., 2002), whilst Eq. (3) shows the multiple linear regression model developed from a stepwise selection of the 168 calculated descriptors.

$$\log(IGC_{50})^{-1} = -0.24(0.075) + 0.42(0.029)\log D - 0.70(0.065)\mathrm{LUMO}$$
$$r^2 = 0.54, \quad r^2_{adj} = 0.54, \quad r^2_{cv} = 0.53, \quad \mathrm{RMSE} = 0.56 \quad (2)$$
$$\log(IGC_{50})^{-1} = 4.92(1.04) + 0.49(0.029)\log P - 1.25(0.23)\,\mathrm{AHard} + 0.22(0.058)NH_{Don} + 1.15(0.19)\mathrm{SdssC}$$
$$r^2 = 0.66, \quad r^2_{adj} = 0.66, \quad r^2_{cv} = 0.64, \quad \mathrm{RMSE} = 0.48 \quad (3)$$

where log D is the distribution coefficient calculated according to (1), LUMO is the energy of the lowest unoccupied molecular orbital, log P is the logarithm of the octanol–water partition, AHard is a measure of molecular hardness, NH_{Don} is the number of hydrogen bond donors and SdssC is the E-state index for a carbon atom with two single and a double bond.

3.2. Neural network model

Stepwise feature selection resulted in the reduction of the 168 descriptors to 20 descriptors. These 20 descriptors were used in the Statistica 'IPS' routine as discussed in the methods and from analysis of the 50 retained models the network shown in Eq. (4) was selected as the best performing network. The selected network had an architecture corresponding to six input variables connected via four hidden nodes to a single one output variable (6:6-4-1:1).

Three-layer perceptron 6:6-4-1:1

$$\log P,\ \log D,\ \mathrm{Vol},\ {}^{O}\chi^{v}_{A},\ {}^{4}\chi_{PC},\ {}^{3}\chi_{P}$$
$$r_{train} = 0.84, \quad r_{test} = 0.85, \quad E_{train} = 0.11, \quad E_{test} = 0.10,$$
$$r^2 = 0.71, \quad \mathrm{RMSE} = 0.45 \quad (4)$$

where log P and log D are as for the linear models, Vol is the molecular volume, ${}^{o}\chi^{v}_{A}$, ${}^{4}\chi_{PC}$ and ${}^{3}\chi_{P}$ are Kier shape and size indices.

Log P and log D were correlated; however removal of either log P or log D diminished the quality of the networks developed (data not shown).

4. Discussion

The characterisation and evaluation of models for *in silico* toxicology will become an important aspect in the application of alternatives to animal testing. Global *in silico* QSARs offer the possibility of producing a single model for a given toxicological endpoint that may cover a wide range of chemical space and a number of different mechanisms of toxic action. These global models are usually described and evaluated in terms of the complete chemical space of the model. However for regulatory usage it is clearly important to understand how the statistics for these global models relate to the predictions of the individual mechanisms of toxic action within the applicability domain. Such an understanding enables confidence to be better assigned to the resulting predictions when the models are used in a real life scenario. This study has therefore investigated the levels of confidence for the different mechanisms of toxic action present in global models developed from a high quality toxicity data set.

In order to understand the performance of global models several different approaches were attempted in this study. The results of these analyses (Table 2) indicate that the neural network model produced the best overall statistics for all compounds in the data set. This is demonstrated by the highest r^2 and q^2_{ext}, coupled with lowest RMSE

Table 2
Global statistics for the QSAR models developed in this study

Number of compounds		Response surface	Multiple linear regression	Neural network
200	r^2	0.54	0.66	0.71
	RMSE (training)	0.56	0.48	0.45
50	q^2_{ext}	0.48	0.72	0.73
	RMSE (validation)	0.62	0.47	0.44

Table 3
Correlation coefficients (r^2) and RMSE statistics for individual mechanisms for the training data (number in parentheses indicates number of chemicals in each mechanism)

		Response surface	Multiple linear regression	Neural network
Polar narcotics (138)	r^2	0.78	0.81	0.84
	RMSE	0.44	0.38	0.35
Pre-electrophiles (22)	r^2	0.14	0.34	0.39
	RMSE	0.99	0.78	0.77
Soft electrophiles (22)	r^2	0.18	0.17	0.25
	RMSE	0.48	0.52	0.55
Pro-redox cyclers (3)	r^2	0.14	0.08	0.06
	RMSE	1.12	1.46	0.92
Respiratory uncouplers (15)	r^2	0.6	0.67	0.75
	RMSE	0.73	0.51	0.44

(training) and RMSE (validation) statistics. The quality of the stepwise multiple linear regression and neural network models in this study are also in keeping with the results of previous studies (Cronin et al., 2002; Devillers, 2004).

Inspection of the statistical fit for the individual mechanisms of action (Table 3) reveals that only the polar narcosis and respiratory uncoupling mechanisms of toxic action were well modelled. Both of these mechanisms involve non-covalent interactions. The electrophilic and metabolically activated mechanisms are poorly modelled, regardless of the model building method, thus only low confidence can be assigned to a prediction for these mechanisms. Figs. 1–3 illustrate the poor modelling of these two mechanisms, regardless of modelling method (response surface, multiple linear regression and neural network, respectively). It can be clearly seen that significant numbers of the pre-electrophiles and soft electrophile chemicals (○ and ◇, respectively) are the worst modelled chemicals.

Analysis of the data for the test or validation set gives an indication of the predictivity each of the models possesses on a per mechanism basis (with the omission of the pro-redox cycling mechanism as there is only a single chemical of this type in the validation data). The results for the validation data (Table 4) confirm that, in keeping with the training data, all three modelling techniques were able to predict the polar narcotic validation chemicals accurately. In addition, the respiratory uncouplers were reasonably predicted by both the multiple linear regression and neural network models. In contrast the response surface model performed significantly less well for this mechanism. All of these results are in keeping with the fit observed for the training data, suggesting that the two non-covalent mechanisms can be equally well modelled by either the multiple linear regression or neural network models.

The pre-electrophile and soft electrophile mechanisms both have poor validation statistics however, for differing reasons. In the case of the pre-electrophiles none of the

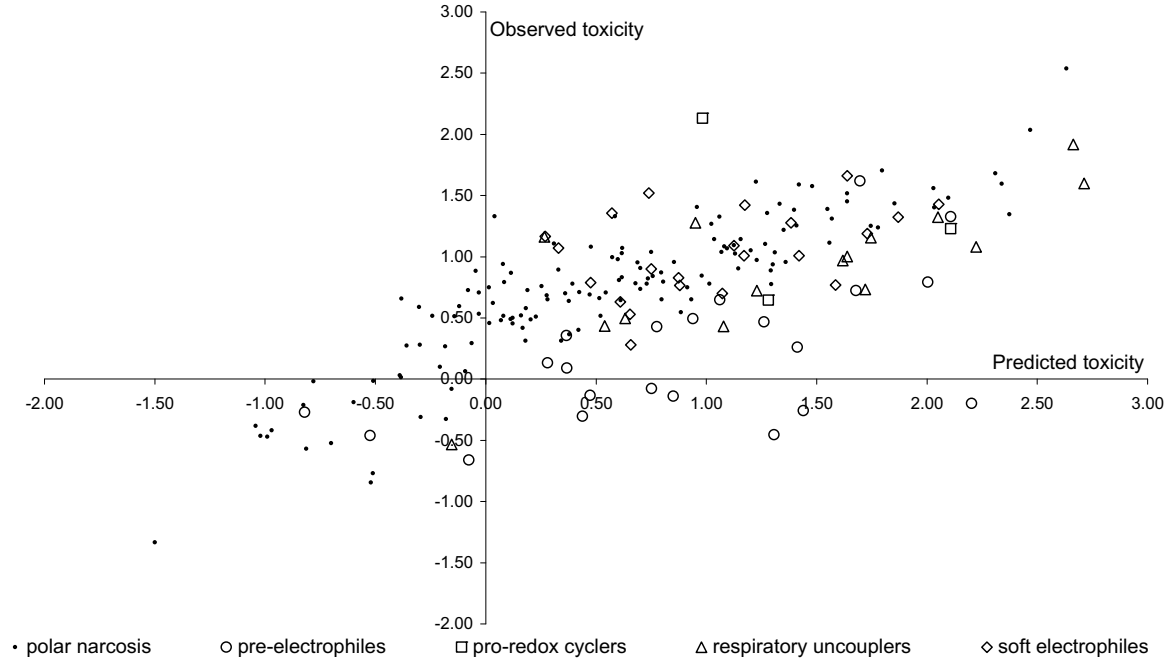

Fig. 1. Observed versus predicted toxicity for the training data used to build the two-parameter response surface model.

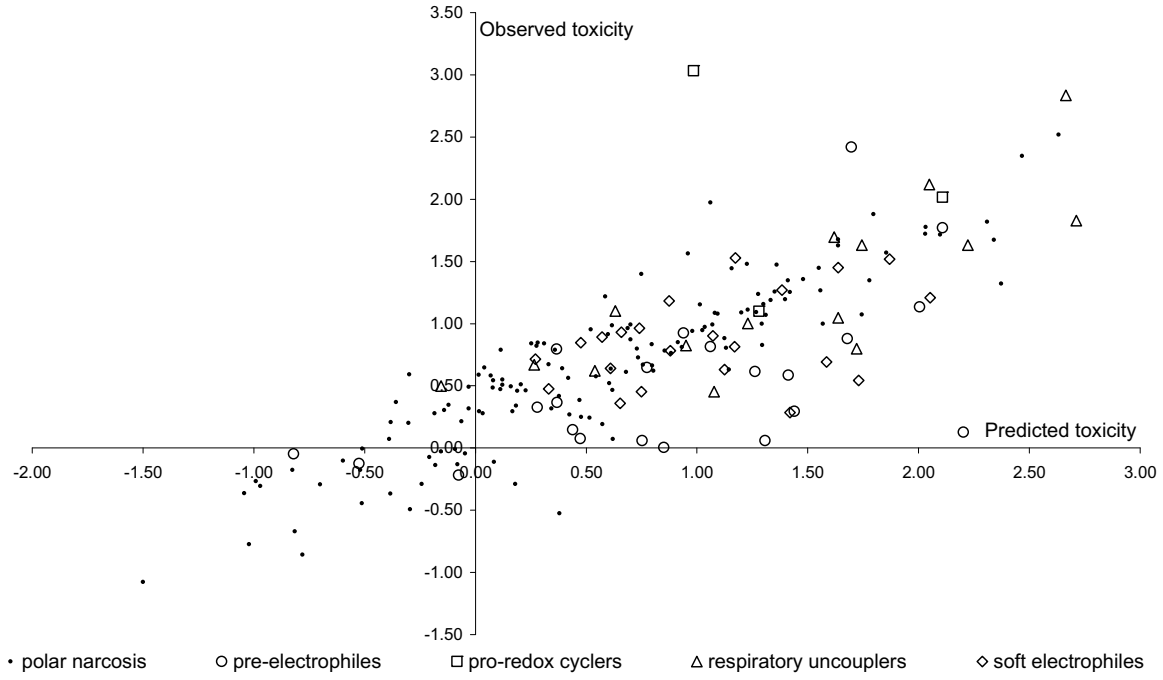

Fig. 2. Observed versus predicted toxicity for the training data used to build the multiple linear regression model.

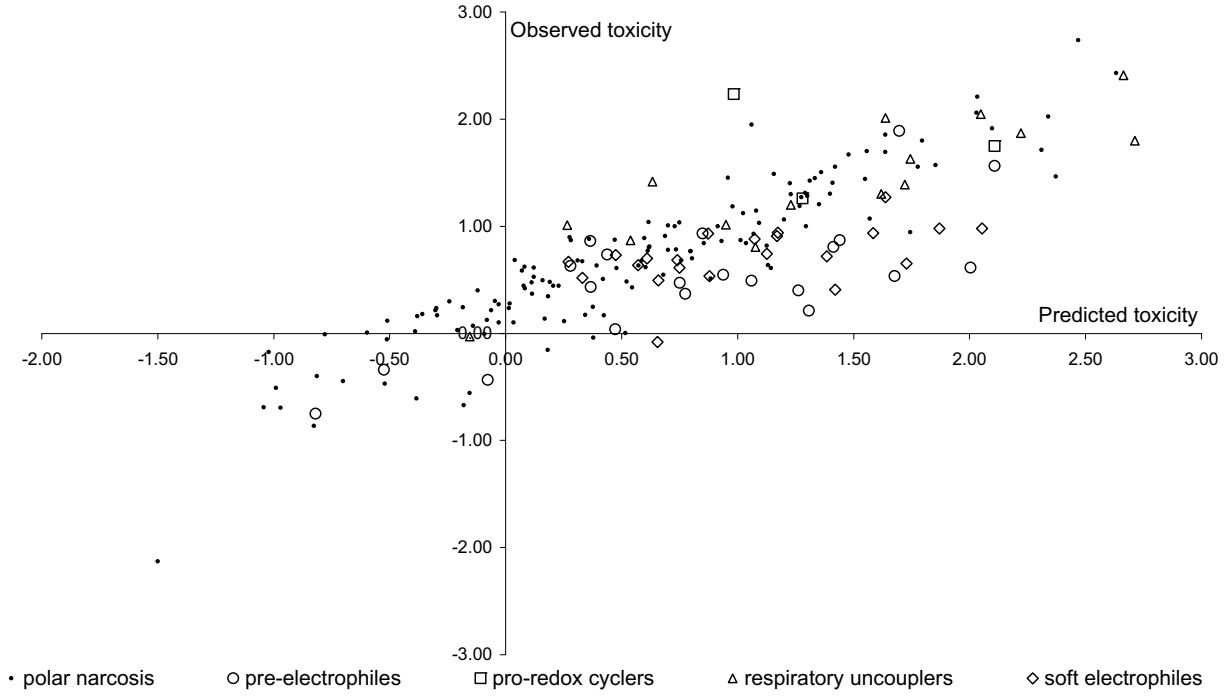

Fig. 3. Observed versus predicted toxicity for the training data used to build the neural network model.

modelling methods gave a low RMSE value in conjunction with high q^2_{ext}; these results are in keeping with those seen in the training data. Interestingly the validation results for the multiple linear regression and neural network models for the soft electrophiles show both low RMSE and high q^2_{ext}. This is in contrast to the extremely poor fit as indicated by the r^2 and RMSE observed in the training data for this mechanism. These contrasting results for this mechanism suggest that the area of model space describing the soft electrophiles is poorly modelled. The variability between training and validation results lead to low confidence in the ability to model these mechanisms. Such observations are in common with other QSAR studies for electrophiles (Cronin et al., 2002; Schultz and Yarbrough, 2004) and

Table 4
Correlation coefficients (q^2_{ext}) and RMSE statistics for the individual mechanisms for the validation data (number in parentheses indicates number of chemicals in each mechanism)

		Response surface	Multiple linear regression	Neural network
Polar narcotics (35)	q^2_{ext}	0.81	0.82	0.82
	RMSE	0.46	0.42	0.39
Pre-electrophiles (5)	q^2_{ext}	0.69	0.57	0.67
	RMSE	1.42	0.96	0.90
Soft electrophiles (5)	q^2_{ext}	0.011	0.73	0.86
	RMSE	0.64	0.36	0.23
Respiratory uncouplers (4)	q^2_{ext}	0.45	0.83	0.92
	RMSE	0.75	0.38	0.56

the poor modelling of these compounds is 'hidden' if only the global model statistics are considered.

Considering the training and validation results together for each of the mechanisms suggests that only the non-covalent mechanisms (polar narcosis and respiratory uncoupling) can be modelled with any confidence (Tables 3 and 4). Mechanisms involving either metabolism (pre-electrophiles and pro-redox cyclers) or electrophilic reactivity (soft electrophiles) are poorly modelled no matter what the model building method. The results also highlight that for the data set studied the multiple linear regression model and neural network models perform almost equally as well as one-another. In addition, the simplest response surface model performs nearly as well for the polar narcotic chemicals as either of the remaining modelling methods.

These results suggest that, for this endpoint, local, individual mechanism of action-based models may be of more value as compared to their global counterparts. In a real-life scenario, a prediction of toxicity may be required for a compound of unknown mechanism of action. If either the global multiple linear regression or neural network models developed in this study was applied, statistics would suggest that the resulting prediction should be of reasonable quality. However, closer inspection of the fit and predictivity (which would not usually be available when developing a global model) shows that this would only be the case if the compound of interest was a polar narcotic or respiratory uncoupler. For example, predictions for a compound where the mechanism of action was known (or at least suspected) to be polar narcosis could be made using any of the models, with some confidence. For regulatory purposes, it is likely that, in keeping with the OECD principles of algorithm transparency and interpretability (Walker et al., 2003), the user will prefer the simplest, most transparent predictive model available. In this study, the QSAR fulfilling these criteria is the two-parameter response surface model rather than either of the other two equally predictive models.

One of reasons for the inability of any of the statistical methods utilised to develop a single predictive global QSAR model across mechanisms of toxic action, is the dominance in the data set of compounds acting by polar narcosis. This dominance leads to a set of descriptors being chosen that best describe this mechanism. As noted in the discussion of the interpretability of the model, for narcosis this leads to descriptors for the ability of a chemical to partition through cell membranes, typically log P or log D. In contrast, for the remaining electrophilic toxicity mechanisms in *T. pyriformis*, partitioning through cell membranes is significantly less important. Instead, electrophilic reactivity is believed to be crucial to these mechanisms (Aptula et al., 2005,2006; Aptula and Roberts, 2006; Roberts et al., 2006,2007). The pre-electrophilic chemicals require abiotic transformation before they exert their toxicity, it is likely that modelling these chemicals in their 'inactive' forms is responsible for the poor results. The soft electrophiles (and the transformed pre-electrophiles) exert their toxicity via irreversible covalent bond formation. This has been suggested to involve nucleophilic attack of these chemicals by nitrogen or sulphur containing amino acid protein residues (Fig. 4). Unfortunately there are currently no descriptors which are easy to calculate that accurately describe reactive electrophilicity and hence this phenomenon is poorly characterised. In addition, accurate modelling of the soft electrophiles is complicated further by the fact that this mechanism was originally defined as halo-nitro-containing phenols that were not uncouplers (Schultz, 1997). This group was poorly distributed across the log D and LUMO descriptor space suggesting that it may contain phenols acting by more than one electrophilic mechanism of action. Further studies are required to elucidate the differing electrophilic mechanisms that can occur for this set of phenols.

The dominance of polar narcotics in the phenol data set investigated in this study provides an important illustration into issues relating to data set bias. In particular, most chemical inventories, and their associated databases, will be biased to some mechanisms of action, and will not be expected to have equal proportions of all mechanisms. In this study, it is important to be aware of the bias in the data set towards polar narcotic compounds; this bias is perhaps one reason why the trained neural networks are less predictive for the mechanisms represented by fewer compounds in the data set. It is possible that if a data set in which sufficiently large numbers of compounds of each mechanism existed (for example more than 100 of each) then neural network analysis might be able to produce significantly improved results. In terms of the development of models for regulatory purposes the mechanistic make up of the

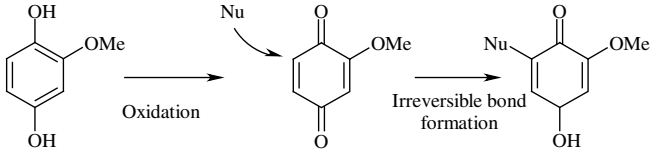

Fig. 4. Example chemistry of poorly modelled pre-electrophilic compound methoxyhydroquinone (where Nu = nitrogen or sulphur group of a protein amino acid).

data set (in terms) of the relative proportions of chemicals in each mechanism of toxic action will influence the final model. This information is not captured in current attempts to describe applicability domains which define coverage of chemical space whilst ignoring mechanistic relevance. The results of this study suggest that mechanistic relevance and the associated model confidence for a particular mechanism are crucial for making predictions in a regulatory environment.

5. Summary

This study has investigated the ability of three global models to model a multi-mechanism data set for the toxicity of phenols to *T. pyriformis*. It has demonstrated that global model statistics can be improved by the application of more powerful (non-linear) modelling methods. However, the study has also highlighted the fact that none of the modelling methods were able to predict, with confidence, the toxicity of compounds acting by mechanisms of action that involve covalent binding (pre-electrophiles and soft electrophiles). The poor predictions for these compounds were hidden, or disguised, by the relatively encouraging statistics for the global models. These results suggest the need for definition of the domain of applicability for a particular mechanism. In addition, the study has also demonstrated that when modelling compounds within domains based on a common mechanism of toxic action, linear regression methods using mechanistically interpretable descriptors perform as well as stepwise multiple linear regression and neural network analyses.

Acknowledgement

The funding of the European Union 6th Framework CAESAR Specific Targeted Project (SSPI-022674-CAESAR) is gratefully acknowledged.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chemosphere.2007.12.011.

References

Aptula, A.O., Patlewicz, G., Roberts, D.W., 2005. Skin sensitization: reaction mechanistic applicability domains for structure-activity relationships. Chem. Res. Toxicol. 18, 1420–1426.

Aptula, A.O., Patlewicz, G., Roberts, D.W., Schultz, T.W., 2006. Non-enzymatic glutathione reactivity and in vitro toxicity: a non-animal approach to skin sensitization. Toxicol. in vitro 20, 239–247.

Aptula, A.O., Roberts, D.W., 2006. Mechanistic applicability domains for non animal-based prediction of toxicological end points: general principles and application to reactive toxicity. Chem. Res. Toxicol. 19, 1097–1105.

Bradbury, S.P., Lipnick, R.L., 1990. Structural-properties for determining mechanisms of toxic action – introduction. Environ. Health Perspect. 87, 181–182.

Cronin, M.T.D., 2003. Quantitative structure–activity relationships for acute aquatic toxicity: the role of mechanism of toxic action in successful modelling. In: Benigni, R. (Ed.), Quantitative Structure–Activity Relationship (QSAR) Models of Mutagens and Carcinogens. CRC Press, Boca Raton FL, USA, pp. 235–258.

Cronin, M.T.D., 2005. Toxicological information for use in predictive modelling: quality, sources and databases. In: Helma, C. (Ed.), Predictive Toxicology. CRC Press, Boca Raton FL, USA, pp. 93–133.

Cronin, M.T.D., Aptula, A.O., Duffy, J.C., Netzeva, T.I., Rowe, P.H., Valkova, I.V., Schultz, T.W., 2002. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. Chemosphere 49, 1201–1221.

Cronin, M.T.D., Schultz, T.W., 1996. Structure–toxicity relationships for phenols to *Tetrahymena pyriformis*. Chemosphere 32, 1453–1468.

Devillers, J., 2004. Linear versus nonlinear QSAR modeling of the toxicity of phenol derivatives to *Tetrahymena pyriformis*. SAR QSAR Environ. Res. 15, 237–249.

Dimitrov, S.D., Mekenyan, O.G., Sinks, G.D., Schultz, T.W., 2003. Global modeling of narcotic chemicals: cilate and fish toxicity. J. Mol. Struct. (Theochem). 622, 63–70.

OECD, 2004. The Report from the Expert Group on (Quantitative) Structure–Activity Relationship ([Q]SARs) on the Principles for the Validation of (Q)SARs. Organisation for Economic Cooperation and Development, Paris.

Roberts, D.W., Aptula, A.O., Patlewicz, G., 2006. Mechanistic applicability domains for non-animal based prediction of toxicological endpoints. QSAR analysis of the Schiff base applicability domain for skin sensitization. Chem. Res. Toxicol. 19, 1228–1233.

Roberts, D.W., Aptula, A.O., Patlewicz, G., 2007. Electrophilic chemistry related to skin sensitization. Reaction mechanistic applicability domain classification for a published data set of 106 chemicals tested in the mouse local lymph node assay. Chem. Res. Toxicol. 20, 44–60.

Schultz, T.W., 1997. Tetratox: *Tetrahymena pyriformis* population growth impairment endpoint – A surrogate for fish lethality. Toxicol. Meth. 7, 289–309.

Schultz, T.W., Sinks, G.D., Cronin, M.T.D., 1997. Identification of mechanisms of toxic action of phenols to *Tetrahymena pyriformis* from molecular descriptors. In: Chen, F., Schuurmann, G. (Eds.), Quantitative Structure–Activity Relationships in Environmental Sciences VII. SETAC Press, Pensacola FL, USA, pp. 329–342.

Schultz, T.W., Yarbrough, J.W., 2004. Trends in structure–toxicity relationships for carbonyl-containing alpha, beta-unsaturated compounds. SAR QSAR Environ. Res. 15, 139–146.

Veith, G.D., Mekenyan, O.G., 1993. A QSAR approach for estimating the aquatic toxicity of soft electrophiles [QSAR for soft electrophiles]. Quant. Struct.–Act. Relat. 12, 349–356.

Walker, J.D., Jaworska, J., Comber, M.H.I., Schultz, T.W., Dearden, J.C., 2003. Guidelines for developing and using quantitative structure–activity relationships. Environ. Toxicol. Chem. 22, 1653–1665.

Worth, A.P., Bassan, A., De Bruijn, J., Saliner, A.G., Netzeva, T., Patlewicz, G., Pavan, M., Tsakovska, I., Eisenreich, S., 2007. The role of the European chemicals bureau in promoting the regulatory use of (Q)SAR methods. SAR QSAR Environ. Res. 18, 111–125.

Yuan, H., Wang, Y., Cheng, Y., 2007. Local and global quantitative structure–activity relationships modeling and prediction for baseline toxicity. J. Chem. Inf. Model. 47, 159–169.