Consensus QSAR Models: Do the Benefits Outweigh the Complexity?

Mark Hewitt, Mark T. D. Cronin, Judith C. Madden, Philip H. Rowe, Clara Johnson, Anrdrea Obi, and Steven J. Enoch*

School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, England

Received January 17, 2007

This study has assessed the use of consensus regression, as compared to single multiple linear regression, models for the development of quantitative structure−activity relationships (QSARs). To provide a comparison, four data sets of varying size and complexity were analyzed: silastic membrane flux, toxicity of phenols to *Tetrahymena pyriformis*, acute toxicity to the fathead minnow and flash point. For each data set, a genetic algorithm was used to develop a model population and the performance of consensus models was compared to that of the best single model. Two consensus models were developed, one using the top 10 models, and the other using a subset of models chosen to provide maximal coverage of model space. The results highlight the ability of the genetic algorithm to develop predictive models from a large descriptor pool. However, the consensus models were shown to offer no significant improvements over single regression models, which are as statistically robust as the equivalent consensus models. Consensus models developed from a selection of the best QSARs were shown not to be superior to a selection of diverse in "model space" QSARs. For the data sets analyzed in this study, and in light of the Organization for Economic Cooperation and Development principles for the validation of QSARs, the increase in model complexity when using consensus models does not seem warranted given the minimal improvement in model statistics.

INTRODUCTION

(Quantitative) structure−activity relationships [(Q)SARs] are being utilized extensively in a wide range of scientific disciplines, including chemistry, biology, environmental sciences and toxicology. Currently, a major driving force behind the success and growth of (Q)SARs for the prediction of toxicity is the implementation of the European Union REACH[1] (Registration, Evaluation and Authorization of CHemicals) legislation. It is anticipated that, under the REACH legislation, (Q)SARs will be used more extensively to reduce the need for in vivo toxicological assessment of existing chemicals. Specifically within the REACH legislation it is anticipated that (Q)SARs will be used to fill data gaps for regulatory purposes.[2−4]

Due to the interest in the regulatory utilization of (Q)SARs, there has been an increased awareness of the requirement for good practice in their application to make predictions of toxicity. This is coupled with an explosion within (Q)SAR of a plethora of new statistical approaches to develop models. Efforts to evaluate (Q)SARs for their usefulness resulted in the Organization for Economic Cooperation and Development (OECD) Principles for the Validation of (Q)SARs.[5] Enshrined within the principles is an assessment of model transparency and (if possible) mechanistic interpretation. Model transparency and mechanistic interpretation are possible, for instance, with (multiple) linear regression (MLR) analysis based on readily comprehensible descriptors. This is at odds, however, with many philosophies of model building. Published (Q)SAR models now apply a variety of statistical approaches, varying in complexity, in order to identify significant chemical or physical properties relating to a biological or chemical response. These techniques range in complexity from regression analysis through to nonlinear neural networks and support vector machines. There is no doubt that there is a place for the complete range of methods, although much knowledge is still required about the strengths and weaknesses of each method and how they may be applied successfully in a framework for REACH and other regulatory environments.

When a large descriptor set is used in the development of a (Q)SAR, appropriate selection methods to reduce the number of descriptors are vital to create an appropriate model.[6] Traditionally, modelers have favored simple transparent modeling techniques such as multiple linear regression, with descriptor selection being performed in a stepwise fashion (based on probabilities to include or exclude a particular variable). However, although transparent, these methods do not perform as well with nonlinear data or for the identification of a relevant subset of descriptors from a large number. In addition, where a large descriptor pool is being used, stepwise regression within the MINITAB software is not suitable as it is only able to handle a maximum of 1000 descriptors.[7] In such cases, more sophisticated methods, such as genetic algorithm selection techniques, are utilized. However such methods are frequently seen as being nontransparent, or as being "black boxes". In terms of application to REACH endpoints and ultimately risk assessment, they may provide a model without a clear indication of how it was derived. This makes interpretation of the model, and hence its application, difficult, which in turn is leading to resistance to the use of models derived in this way.

* Corresponding author phone: + 44 151 231 2066; fax: + 44 151 231 2170; e-mail: S.J.Enoch@ljmu.ac.uk.

10.1021/ci700016d CCC: $37.00 © 2007 American Chemical Society
Published on Web 07/07/2007

Table 1. Summary of Data Sets Used in the Study

dataset	no. compounds	data set	endpoint	structural makeup	number of calculated descriptors	reference(s)
1	256	membrane flux	silastic membrane flux (logJ)	heterogeneous structures	1389	24−26
2	250	*T. pyriformis*	log IGC_{50} (conc. mmol/L)	substituted phenol derivatives	1418	27
3	605	fathead minnow	log LC_{50} (conc. mol/L)	heterogeneous structures	1537	28
4	399	flash point	flash point °C	heterogeneous structures	1446	29−31

Table 2. Summary of Calculated Descriptors and Associated Software

Software	Calculated Descriptors
TSAR V3.3	Molecular mass; molecular surface area; molecular volume; inertia moments; ellipsoidal volume (Ellvol); total dipole; dipole moments; log P; molecular refractivity; kier simple and valence-corrected molecular connectivity indices: zero order, 1st and 6th order path, 3rd to 4th order cluster, 4th order path/cluster, 3rd to 6th order ring; kappa shape indices; shape flexibility; rotatable bonds (Σ rot); Randic, Balaban and Wiener topological indices; sum of E-state indices; number of halogen atoms (Σ Hal); number of H-bond donor (N_{HBD}) and acceptor (N_{HBA}) centres; ring counts; group counts (methyl, ethyl (ΣC2H5), amino, hydroxyl, phenyl, acid anhydride, acid chloride, aldehyde, alpha haloketone, beta haloketone, beta lactone, chain C=N, cyano, hydrazide, hydrazine, isocyanate, isothiocyanate, nitro, polycyclic, polyethene, sulphonyl, thiourea); total energy; electronic energy; nuclear energy; surface area; mean polarizability; total molecular charge; heat of formation (DeltaHf); ionization potential; LUMO; HOMO; quadpoles; octupoles
HYBOT V2.1.0.706	Alpha; max(Ca); max(Cd); max(Q+); max(Q-); Sum(Ca); Sum(Cd); Sum(Q+), Sum(Q-); SumC; Sum(Q+)/Alpha; Sum(Q-)/Alpha; Sum(Ca)/Alpha; Sum(Cd)/Alpha; Sum(C)/Alpha
Dragon Professional V5.3	Constitutional descriptors; topological descriptors; walk and path counts; connectivity indices; information indices; 2D autocorrelations; edge adjacency indices; Burden eignenvalues; topological charge indices; eigenvalue-based indices Randic molecular profiles; geometrical descriptors; RDF descriptors; 3D-MoRSE descriptors; WHIM descriptors GETAWAY descriptors; functional group counts; atom-centred fragments; charge descriptors; molecular properties
ACD Labs V9.08	Log P

Genetic algorithms (GAs) have been applied in descriptor selection to build a variety of models. They are suitable for the development of models from a large pool of molecular descriptors (typically over 1000).[8−11] A GA for descriptor selection is a stochastic search method that mimics biological evolution by undertaking a survival of the fittest approach. This involves a population of solutions, consisting of individuals known as chromosomes. Once a chromosome has been created, its fitness is assessed using a scoring function defined by the user, typically either the cross-validated coefficient of determination (r_{cv}^2) or predictivity for an external test set (q_{ext}^2). This process is repeated for every individual in the population with the chromosomes then being ranked. The best solutions are chosen to survive unchanged, while the remainder are subjected to crossover and mutation, the outcome of which mimics biological evolution to produce a new population. This cycle is then repeated for a specified number of iterations.[12] This leads to a final population (of descriptors) that is better suited to the modeled endpoint than the individuals they were created from, just as in natural selection.[10] It has been suggested that MLR applied to descriptors selected by a GA provides better models than those selected by stepwise regression.[9,13] The use of a GA is also beneficial as a GA searches a population of solutions in parallel, potentially avoiding local minima and suboptimal solutions.

In an effort to derive more predictive models than the traditional single regression model, consensus models are being utilized. These involve the application of a genetic algorithm to create a number of unique individual models, and to use an average of the predictions made. Consensus modeling has been employed in different (Q)SAR applications.[11,14−16] The theory proposes that an individual QSAR model may overemphasize some physicochemical aspects of the molecules in question, underestimate others, and ignore many important features completely. Thus, it seems reasonable to anticipate that a consensus (Q)SAR model, derived from averages of predictions from individual models, may provide better statistical fit and predictive ability as compared to the individual models. This has been seen in previous studies that have shown a superiority of consensus approaches over the use of individual models.[9,11,14,17−21] It is important to note that there are alternative methods to develop a consensus model other than simply using the simple mean prediction. For example, it is possible to use a leverage-weighted mean, whereby the most predictive model has the greatest contribution. It is also important to note that the models within a consensus model can still make poor predictions, whereby the mean of these is equally incorrect. For this reason, outlier analysis is still required to ascertain "bad" predictions.

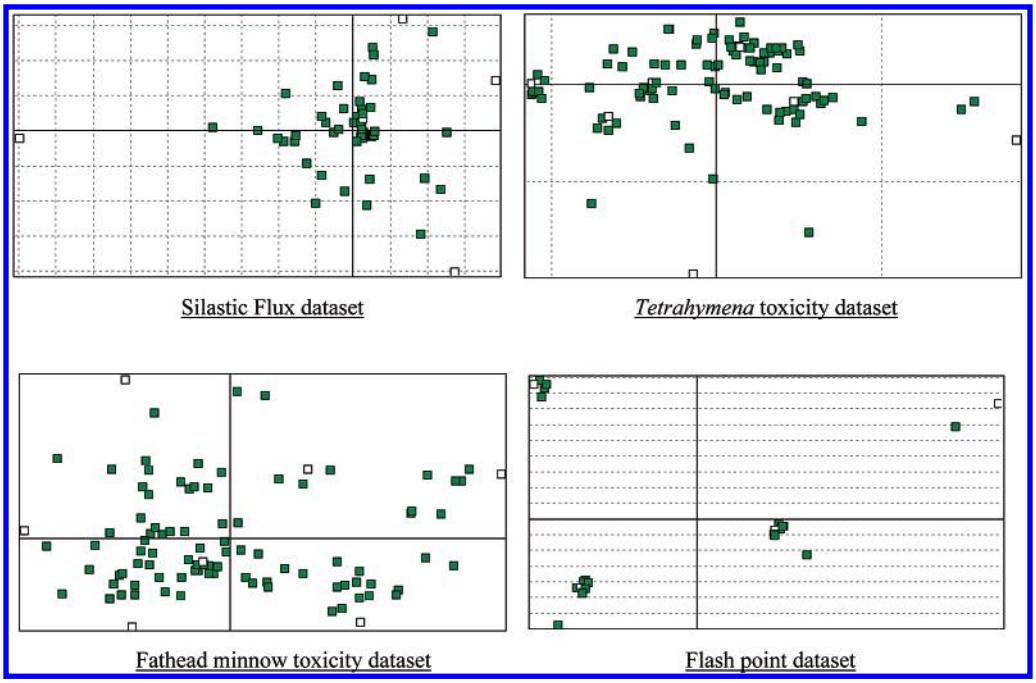

Figure 1. Principle coordinate analysis plots visualizing the global model space of the GA model population for each data set (□ represents models included in the diverse consensus model).

Consensus modeling has also been shown to diminish the effects of noisy data, in that all individual models contain varying amounts of noisy data and the averaging of these noisy data leads to more reliable predictions.[22] They would seem particularly appropriate to model large, heterogeneous data sets where the mechanism of action is not known or may be considered to be quite "general", for example, narcosis, passive diffusion, or a physicochemical property. Consensus models are often coupled to a genetic algorithm for the identification of significant physicochemical descriptors.

Despite their utility for model development, a full assessment of the capabilities of consensus models has not been performed in light of the requirements of the OECD validation principles. Recent studies by Gramatica et al.[9,17] have compared the use of single regression and consensus models and state the importance of chemical and model space. Gramatica et al.[9,17] suggest that, in order to develop a more robust consensus model, individual models covering the full expanse of model space must be used.

As previously mentioned, consensus models are thought to provide more stable predictions based on an average principle, and for this to be effective, a consensus model should contain individual models from different areas of model space. It has been suggested that the Hamming distance is an effective way of mapping model space.[9,17,23] Hamming distances are used to give a distance between two points in a multidimensional geometric space. In this application, they give the distance between two models in descriptor space.

The aim of this investigation, therefore, was to assess consensus regression models for (Q)SAR in a direct comparison with using the single best multiple linear regression model generated by the GA in each case. In addition, the importance of covering maximum model space was investigated using a second consensus model that was developed to best represent the available global model space. The endpoints modeled were flux through a silastic membrane, flash point, acute toxicity to the fathead minnow, and toxicity of phenols to *Tetrahymena pyriformis*. A diverse set of descriptors was calculated in each case to assist the development of consensus models. Consensus models were then compared directly with single multiple linear regression models. The models were then analyzed to see if consensus modeling offered greater statistical fit and predictivity to attempt to explain why better fit was obtained. It should be stressed at the outset that the purpose of this study was not modeling *per se*, most endpoints are well modeled, but to investigate the role of consensus modeling and the use of a GA in order to develop MLR models.

METHODS

Data Sets. Four data sets covering a range of physical, biological and toxicological properties were obtained from the literature and are summarized below and in Table 1.

Data set one contains data for flux through an artificial membrane. Such flux measurements are important as they can be related to the flux of compounds through the skin. These data were generated by calculating the steady-state flux of compounds through a polydimethylsiloxane membrane at 30 °C and were compiled by Cronin et al.[24] from the original references from Chen et al.[25,26]

Data set two contains toxicity data for *Tetrahymena pyriformis*. Phenol derivatives were assessed and their toxicity reported as the concentration (mmol/L) resulting in 50% growth inhibition (log IGC_{50}) after 40 h. These data were collected from the literature.[27]

Data set three contains aquatic toxicity data for the fathead minnow (*Pimephales promelas*). Data were recorded as 96-h median lethal concentration [LC_{50} (mol/L)] values. Data were collected from the literature.[28]

Data set four contains flash point data (°C). This is the only physical property modeled in this study and is of great

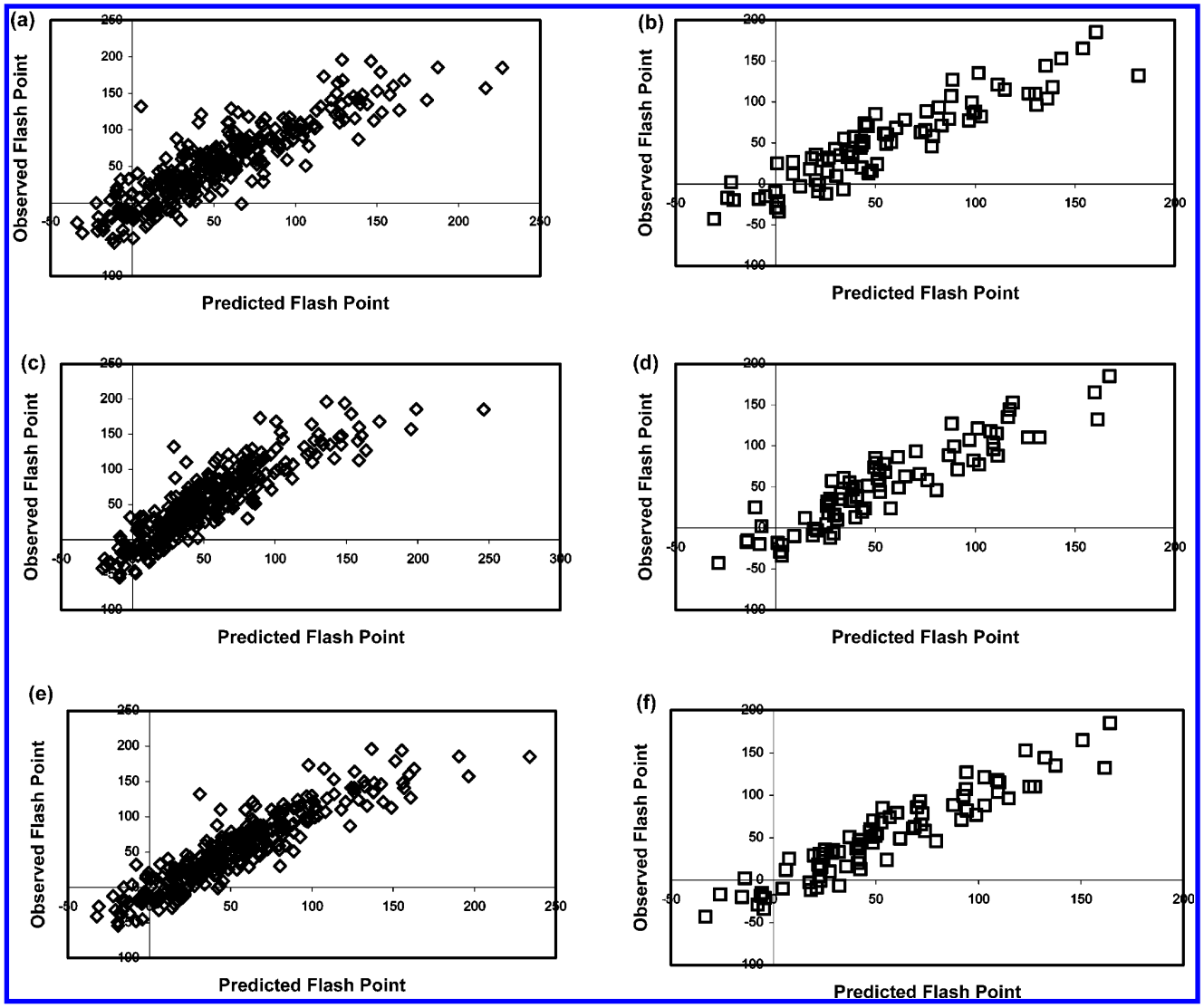

Figure 2. Plot of observed versus predicted flash point. Using SR (a,b), DC (c,d), and C10 (e,f) models. ♦ = Test set. ◇ = Training set.

Table 3. Summary of Model Statistics Using Single GA-MLR (SR), Diverse Consensus (DC), and Top 10 Consensus (C10) Models

data set	model type	r_{cv}^2	$RMSE_{Train}$	q_{ext}^2	$RMSE_{Test}$
silastic membrane	SR	80.08	0.49	77.49	0.51
flux	DC	80.90	0.48	79.00	0.49
	C10	80.30	0.49	79.05	0.49
Tetrahymena pyriformis	SR	64.14	0.50	70.79	0.46
toxicity	DC	63.79	0.50	74.99	0.43
	C10	65.24	0.49	73.55	0.43
fathead minnow	SR	70.29	0.70	60.95	0.82
toxicity	DC	70.77	0.70	59.61	0.84
	C10	71.45	0.69	59.28	0.84
flash point	SR	77.92	24.31	84.11	20.08
	DC	78.31	24.10	84.01	20.15
	C10	82.54	21.62	88.39	17.17

importance for the industrial handling and storage of chemicals and also in risk assessment. The flash point data set was taken from Tetteh et al.,[29] who originally sourced them from two chemistry handbooks.[30,31] Flash point was measured using a variety of standard techniques.

For further details of the methodological aspects of the biological and physicochemical tests, the reader is referred to the original literature sources.

Descriptors. For each compound included in these analyses, the 3D structure was generated from Simplified Molecular Input Line Entry System (SMILES) notations using TSAR version 3.3 (Accelrys Inc, Oxford, England). A large number of descriptors were calculated using TSAR; HYBOT contained within the MOLPRO software package, version 2.1.0.706 (Dr. Sergei V. Trepalin, 1997−2000); DRAGON professional version 5.3 (R. Todeschini, Milano Chemometrics and QSAR research group, 2005); and ACD/ChemSketch version 9.08 (Advanced Chemistry Development Inc, 2006). The exact number of descriptors for each data set differed due to differences in chemical structures within the data sets (such as the presence of certain functional groups or ring systems), nonsuitable descriptors (i.e., all values being zero

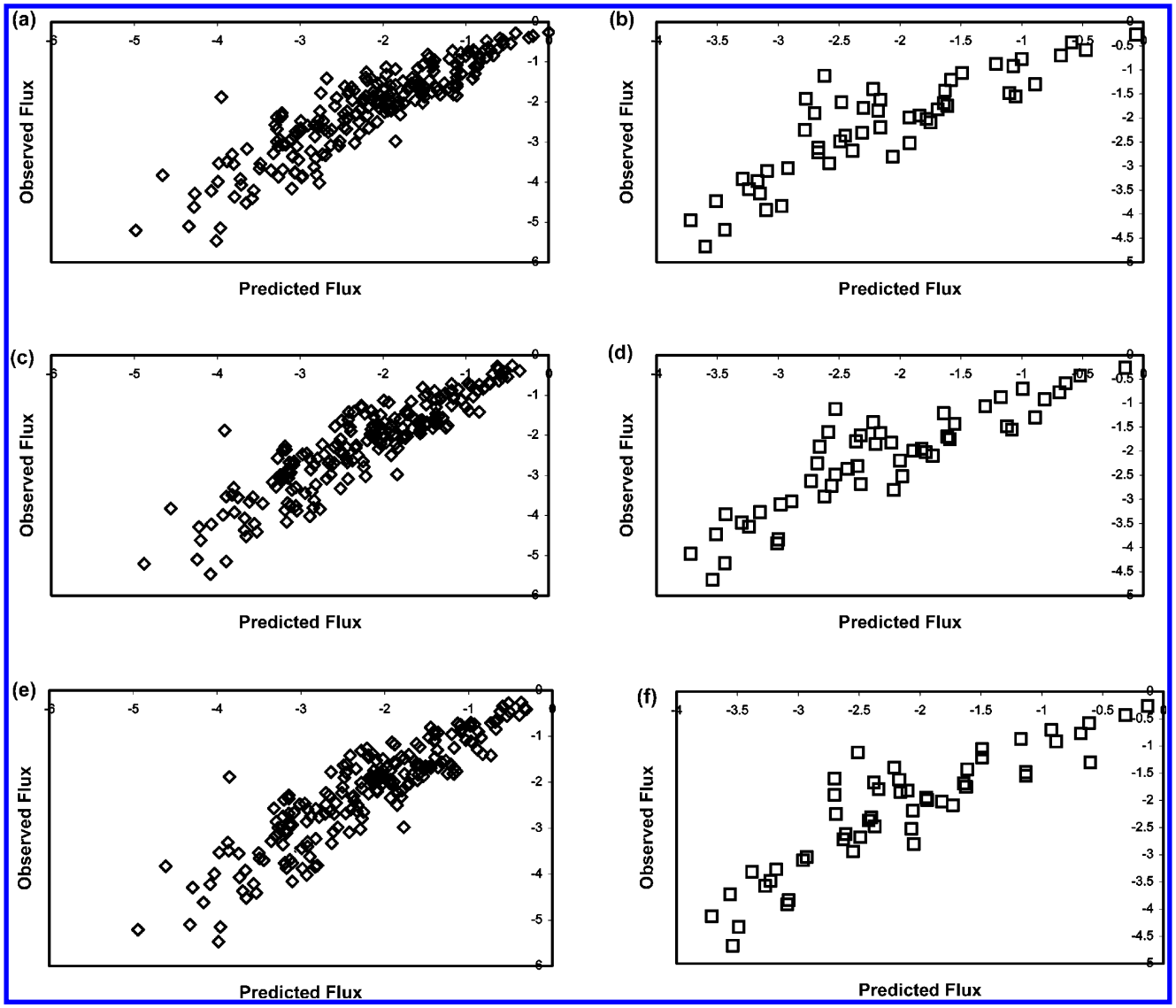

Figure 3. Plot of observed versus predicted silastic membrane flux. Using SR (a,b), DC (c,d), and C10 (e,f) models. ♦ = Test set. ◇ = Training set.

or identical), and missing descriptors (i.e., could not be generated for certain compounds due to software limitations). A list of descriptors/descriptor classes is given in Table 2.

Statistical Analysis - Development and Optimization of Multiple Linear Regression Models and Consensus Models Using Genetic Algorithm Descriptor Selection. Consensus models were developed using the MOBYDIGS software.[32] This contained two elements, the GA and the consensus modeling function. For each of the four data sets, a single input file was generated for MOBYDIGS that contained all endpoint data and pooled descriptor values. Any missing values were coded for by "−999", as required by MOBYDIGS.

Training/Test Split. Each data set was split using a 4:1 training/test split to form an external test set. Compounds were ranked according to their endpoint, and every fifth compound was labeled as a test compound and removed from the training set. This method produced test sets that fairly represented the data.

Model Optimization. To determine the optimum number of descriptors to include in each model, the genetic algorithm was used to build models containing an increasing number of descriptors (starting with one and with a maximum of 10). The model statistics [r_{cv}^2 (cross-validated correlation coefficient), q_{ext}^2 (external validation correlation coefficient), and F (Fisher statistic)] were compared in order to ascertain the optimum number of descriptors to include in each model. When an additional descriptor resulted in an insignificant increase in model fit and predictivity, or a decrease in the F statistic, the addition of descriptors was stopped. All subsequent consensus models were derived using the optimum number of descriptors.

GA Configuration. The GA was set up using a population size of 100, the software's maximum (default of 50). This population of models was ranked by the GA according to their predictivity using q^2_{ext} values. The maximum number of descriptors allowed in each model was set at that value found to be optimum as detailed previously. The Tabu list in MOBYDIGS was enabled and set at the default setting whereby descriptors enter the Tabu list when they have a fourth-order moment value greater than eight. As a result constant, or near-constant, descriptors are removed from the analysis. The GA was then allowed to run for 100 000 cycles in order to generate a stable model population.

Consensus Modeling. Following the use of the GA, a large number of models was created and could be then used to perform consensus modeling. MOBYDIGS is limited to using a maximum of 10 models to make consensus predictions. This study investigated consensus predictions made using the top 10 (C10) models as identified by the GA (MOBYDIGS default). A second consensus model [diverse consensus (DC)] was made using a number of diverse models chosen to best cover the available model space (see following sections). In addition, the predictions using the top single MLR model (SR model) found by the GA were also recorded in order to assess the potential benefits of consensus modeling over using the single best model. When these three methods were used, predictions were then made for all training and test compounds.

Statistical Analysis. *Assessment of Model Space and Development of Diverse Consensus Models.* As previously discussed by Gramatica et al.,[9,17] sufficient coverage of model space is of great importance in order to create a robust consensus model. As with any QSAR, in order to understand a consensus prediction fully, together with its limitations, the coverage of model space should be investigated.

To create a consensus model which contained maximum diversity, a method proposed by Todeschini et al.[23] was utilized. This method is based on Hamming distances calculated from molecular descriptors, enabling the most dissimilar models to be identified (greatest Hamming distance). Using principle coordinate analysis (PCA), a two-dimensional plot of model space is generated allowing the distribution of models within the model space to be visualized (see Figure 1). To ensure that the models generated by the GA were of comparable quality, only models with r_{cv}^2 and q_{ext}^2 values within 10% of the top model were included in the model space analysis. In order to generate a consensus model that best models this global model space, a number of individual models were chosen from different clusters of models on the PCA plot. This is, however, a purely arbitrary process and highlights the difficulties both in defining a cluster of models and in the choice of model to use from each cluster. For the purposes of this investigation, a model was chosen at random from each obvious cluster, and if no clear clusters existed, models were taken from the outskirts of model space and from its approximate center. The chosen models were used to create a DC model that best represents the global model space.

Assessment of Model Performance. In order to assess model performance, it was necessary to investigate statistical fit and predicitivity. In order to assess each of these, the following statistics were generated:

r_{cv}^2 = cross-validated correlation coefficient

q_{ext}^2 = correlation coefficient for external test set

RMSE = root mean square error

RMSE was calculated for each model using eq 1 shown below:

$$\mathrm{RMSE} = \sqrt{\frac{\sum (\mathrm{Pred} - \mathrm{Obs})^2}{N - 1}} \qquad (1)$$

where

Pred = predicted value

Obs = observed value

N = number of compounds in the data set

r_{cv}^2 for the training data and q_{ext}^2 for the test data were also calculated for each model using eq 2:

$$q_{ext}^2 = 1 - \frac{\sum (\mathrm{Pred} - \mathrm{Obs})^2}{\sum [\mathrm{Obs} - \bar{X}(\mathrm{Obs})]^2} \qquad (2)$$

where

$\bar{X}$(Obs) = the mean observed value

These statistics can be used as an indication of model quality and stability.

RESULTS AND DISCUSSION

Four diverse data sets were used to assess whether (Q)SARs derived from a consensus analysis of a population of multiple linear regression models were superior to the best performing single regression model. Model quality was assessed in terms of statistical fit for the training data and predictivity of an external test set for each data set. The performance of differing (Q)SAR methodologies is of particular interest given the REACH legislation[33] and the implementation of such methods with regard to the OECD principles for validation.[5]

Consensus and Single Regression Models. Two consensus QSAR models were constructed for each data set studied. The first of these was the MOBYDIGS default best 10 model consensus analysis (C10). This consisted of taking the average prediction from the top 10 models from the model population. The second method was similar to that utilized in previous studies[9,17] in which a diverse subset of the model population was used to create the consensus prediction (DC). The diverse subset was selected from a reduced set of the final population of models; with only models within 10% of the best model included in the reduced set (both r_{cv}^2 and q_{ext}^2 were within 10% of the best model).

Construction of the DC model involved using the Hamming distance between the individual models in a population to assess the global model space. Utilizing the Hamming distance between individual models allowed clusters of related models to form, the DC model was then constructed by selecting a representative sample of models from the clusters. PCA as implemented in the MOBYDIGS software package enabled the global model space and the clusters of individual models to be visualized, leading to the four plots shown in Figure 1.

It is clear from Figure 1 that the four endpoints studied exhibit differing levels of individual model clustering within the available global model space. In cases where discreet clusters form, for example, in the flash point data set, it is relatively easy to select an appropriate number of individual models to make the DC model. However, when the individual models do not form discreet clusters, or when the clusters overlap to some degree, it becomes much more difficult to assess which (and how many) of the individual

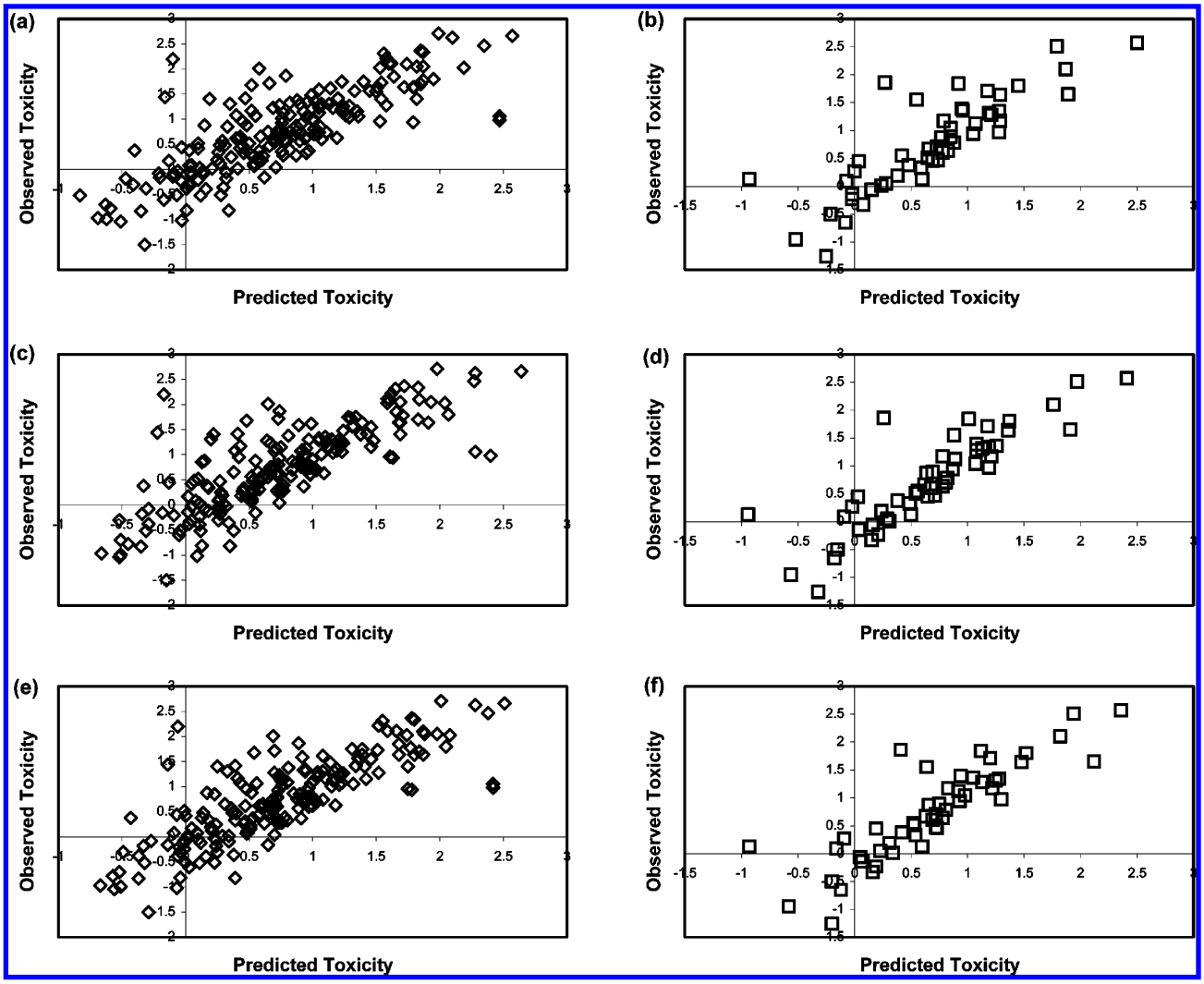

Figure 4. Plot of observed versus predicted *Tetrahymena pyriformis* toxicity. Using SR (a,b), DC (c,d), and C10 (e,f) models. ♦ = Test set. ◇ = Training set.

models should be used to construct the DC model. This is highlighted by the global model space of the remaining data sets. The problem of how to select the optimum number of individual models to form a DC model has been discussed previously,[9,17] with it being suggested that an algorithmic method would be preferable.

Comparison of the performance of the two consensus analyses (C10 and DC) revealed that neither method consistently outperformed the other. In three of the four data sets, the two methods exhibited approximately equal levels of statistical fit and predictivity. Only for the flash point data was any improvement observed between the two methods, with the C10 model moderately outperforming the DC model (Table 3).

In addition to consensus models, the best single regression model (SR model) for each data set was analyzed. The SR model was simply the individual model in the final population that had the best r_{cv}^2 and q_{ext}^2 values. For each of the data sets studied, the SR model was comparable to the DC model, with only the C10 model for the flash point data set exhibiting any improvement in statistical fit and predictivity compared to the SR model (Table 3). Equations 3−6 show the SR models for each of the endpoints studied; these equations in combination with the equations for the DC and C10 models (Supporting Information, Table S1) highlight the quality and stability of all the models developed in this study.

Silastic membrane flux:

$$\log J = -5.002(0.207)\text{Sum(C)/Alpha} - 0.741(0.039)\text{X3sol} + 1.356(0.143)$$

$$r_{adj}^2 = 80.5,\ r_{cv}^2 = 80.1,\ s = 0.487,\ F = 420,\ q_{ext}^2 = 77.5 \quad (3)$$

Toxicity to *Tetrahymena pyriformis*:

$$\log \text{IGC}_{50} = 0.481(0.029)\log P - 0.759(0.219)\text{GATS1v} + 0.233(0.03)\text{C026} - 0.469(0.176)$$

$$r_{adj}^2 = 65.1,\ r_{cv}^2 = 64.2,\ s = 0.489,\ F = 124,\ q_{ext}^2 = 70.8 \quad (4)$$

Toxicity to fathead minnow:

$$\log \text{LC}_{50} = -0.286(0.034)\text{nDB} - 0.778(0.026)\text{ALOGP} + 1.05(0.07)$$

$$r_{adj}^2 = 70.6,\ r_{cv}^2 = 70.3,\ s = 0.698,\ F = 489,\ q_{ext}^2 = 70.3 \quad (5)$$

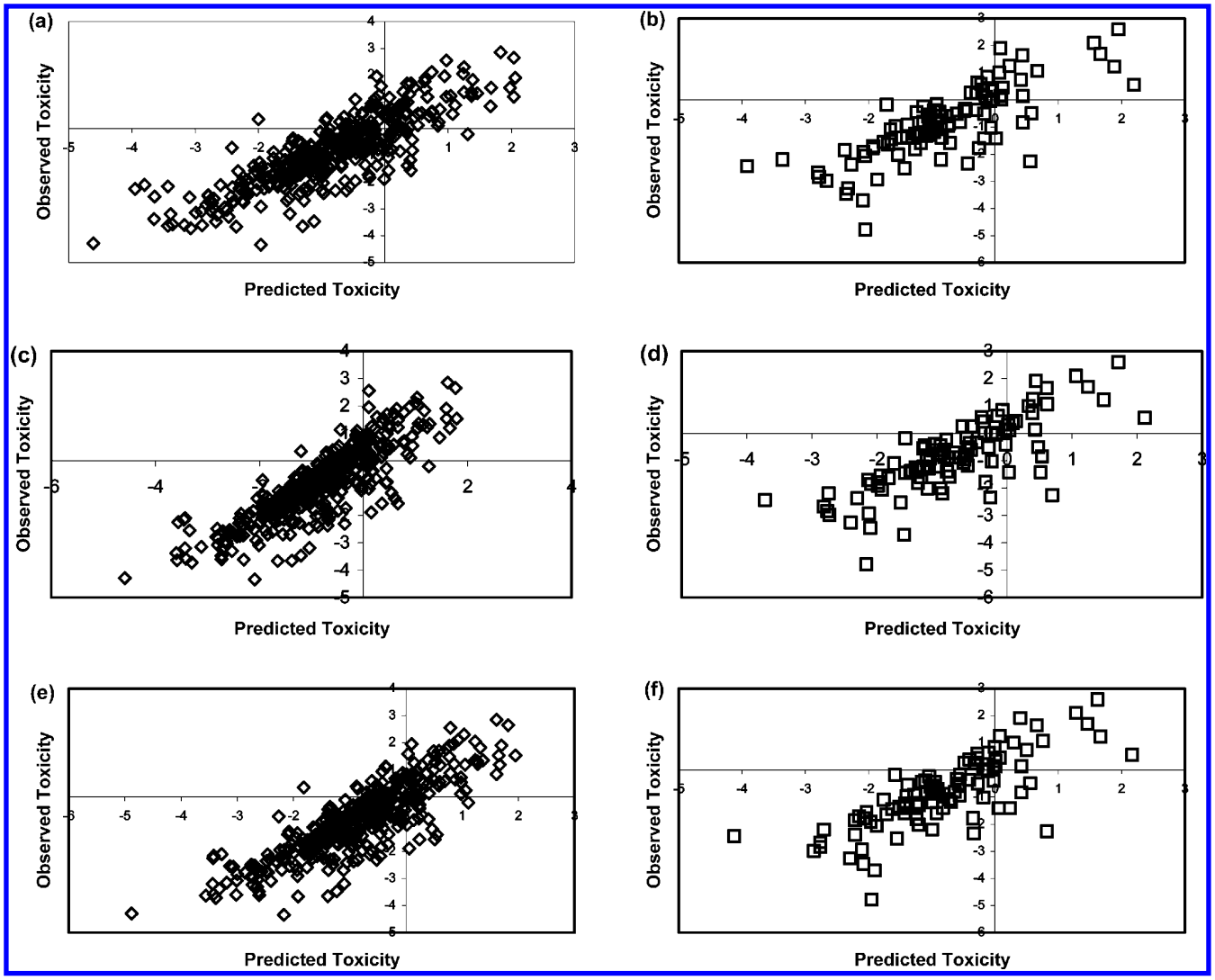

Figure 5. Plot of observed versus predicted fathead minnow toxicity. Using SR (a,b), DC (c,d), and C10 (e,f) models. ♦ = Test set. ◇ = Training set.

Flash point:

$$\text{flash point} = -24.7(1.69)\max(\text{Cd}) + 22.4(0.69)\text{G2} - 69.6(3.85)$$

$$r_{\text{adj}}^2 = 78.3,\ r_{\text{cv}}^2 = 77.9,\ s = 24.1,\ F = 575,\ q_{\text{ext}}^2 = 84.1 \quad (6)$$

The improvements in the models for the flash point data set can be seen visually by plotting the predicted values against the actual values (Figure 2). These plots highlight the improvement in statistical fit (Figure 2a,c,e) and predictivity (Figure 2b,d,f) in the SR model (Figure 2a and b) compared to the DC (Figure 2c and d) and C10 models (Figure 2e and f). In addition, they also show the similarity between the SR and DC models visually. Similar plots for the remaining data sets, visually confirming the similarity of C10, DC, and SR models can be seen in Figures 3−5.

The results of this study are in contrast to those found previously utilizing consensus MLR methods,[9,17] in that for three of the four data sets no improvement in either statistical fit or predictivity was observed when comparing either the DC or C10 models to the SR model. Only for the flash point data did a consensus analysis offer an improved model; again in contrast to previous studies, it was the C10 rather than the DC model (Table 3).

The 4% improvement in r_{cv}^2 and q_{ext}^2 values observed for the C10 model for the flash point data is in keeping with that observed in previous studies. Studies have shown that only relatively small improvements should be expected when applying a consensus analysis. For example, Gramatica et al.[9] observed that for the prediction of OH tropospheric degradation of volatile organic compounds consensus modeling improved r^2 and q_{ext}^2 by between 1 and 3%.

In terms of the OECD principles relating to model simplicity and interpretability, the results indicate that for the data sets studied there is no consistent benefit in terms of statistical fit or predictivity for either consensus method investigated, compared to the nonconsensus SR models. In addition, development of a population of QSAR models from which the consensus model is derived requires a stochastic search method and many hundreds of descriptors. The use of a large number of molecular descriptors results, inevitably, in the use of descriptors that are difficult to interpret mechanistically; this is in conflict with the OECD requirements.

In conclusion, this study has investigated the benefits, in terms of statistical fit and predictivity, of consensus QSAR

modeling using four varied data sets. It has been demonstrated that, for the data sets studied, there appears to be no consistent significant benefits in terms of statistical fit and predictivity in the consensus models compared to single regression models. The results also showed that both consensus methods (DC and C10) produced models of equal quality.

ACKNOWLEDGMENT

This work was supported, in part, by the European Union Sixth Framework ReProTect Integrated Project (LSHB-CT-2004-503257) and the European Union CAESAR Specific Targeted Research Project (SSPI-022674-CAESAR).

Supporting Information Available: Summary of model equations, silastic membrane data set, *Tetrahymena pyriformis* data set, fathead minnow data set, and flash point data set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) *White Paper on a Strategy for a Future Chemical Policy (COM(2001)-88final)*; Commission of the European Communities: Brussels, Belgium, 2001.
(2) Worth, A. P.; Cronin, M. T. D. Report on the Workshop on the Validation of QSARs and Other Computational Prediction Models. *ATLA, Altern. Lab. Anim.* **2004**, *32*, 703−706.
(3) Simon-Hettich, B.; Rothfuss, A.; Steger-Hartmann, T. Use of Computer-Assisted Prediction of Toxic Effects of Chemical Substances. *Toxicology* **2006**, *224*, 156−162.
(4) Worth, A. P.; Bassan, A.; De Bruijn, J.; Saliner, A. G.; Netzeva, T.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Eisenreich, S. The Role of the European Chemicals Bureau in Promoting the Regulatory Use of (Q)SAR Methods. *SAR QSAR Environ. Res.* **2007**, *18*, 111−125.
(5) *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, France, 2004.
(6) Ghafourian, T.; Cronin, M. T. D. The Impact of Variable Selection on the Modelling of Oestrogenicity. *SAR QSAR Environ. Res.* **2005**, *16*, 171−190.
(7) Minitab for Windows Statistical Software: *MINITAB*, version 14.12.0; Minitab Inc.: State College, PA, 2004.
(8) Cho, S. J.; Hermsmeier, M. A. Genetic Algorithm Guided Selection: Variable Selection and Subset Selection. *J. Chem. Inf. Model.* **2002**, *42*, 927−936.
(9) Gramatica, P.; Giani, E.; Papa, E. Statistical External Validation and Consensus Modeling: A QSPR Case Study for Koc Prediction. *J. Mol. Graphics Modell.* **2007**, *25*, 755−766.
(10) Ersin, B.; Peter, S.; Rebecca, H.; Yun-De, X.; Aaron, J. C.; Jeffrey, D. S. Genetic Algorithms and Self-Organizing Maps: A Powerful Combination for Modeling Complex QSAR and QSPR Problems. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 483−493.
(11) Ganguly, M.; Brown, N.; Schuffenhauer, A.; Ertl, P.; Gillet, V. J.; Greenidge, P. A. Introducing the Consensus Modeling Concept in Genetic Algorithms: Application to Interpretable Discriminant Analysis. *J. Chem. Inf. Model.* **2006**, *46*, 2110−2124.
(12) Hasheminia, H.; Akhavan Niaki, S. T. A Genetic Algorithm Approach to Find the Best Regression/Econometric Model among the Candidates. *Appl. Math. Comput.* **2006**, *183*, 337−349.
(13) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267−281.
(14) Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of Prediction Confidence and Domain Extrapolation of Two Structure-Activity Relationship Models for Predicting Estrogen Receptor Binding Activity. *Environ. Health Perspect.* **2004**, *112*, 1249−1254.
(15) Baurin, N.; Mozziconacci, J. C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Model.* **2004**, *44*, 276−285.
(16) van Rhee, A. M. Use of Recursion Forests in the Sequential Screening Process: Consensus Selection by Multiple Recursion Trees. *J. Chem. Inf. Model.* **2003**, *43*, 941−948.
(17) Gramatica, P.; Pilutti, P.; Papa, E. Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training−Test Sets and Consensus Modelling. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794−1802.
(18) Mario, L.; Vinothini, S. In Silico Prediction of Aqueous Solubility, Human Plasma Protein Binding and Volume of Distribution of Compounds from Calculated pKa and AlogP98 Values. *Mol. Diversity* **2003**, *7*, 69−87.
(19) Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Livingstone, D. J.; Ford, M. G.; Whitley, D. C. Selecting Screening Candidates for Kinase and G Protein-Coupled Receptor Targets Using Neural Networks. *J. Chem. Inf. Model.* **2002**, *42*, 1256−1262.
(20) Sutherland, J. J.; Weaver, D. F. Development of Quantitative Structure−Activity Relationships and Classification Models for Anticonvulsant Activity of Hydantoin Analogues. *J. Chem. Inf. Model.* **2003**, *43*, 1028−1036.
(21) Sussman, N. B.; Arena, V. C.; Yu, S.; Mazumdar, S.; Thampatty, B. P. Decision Tree SAR Models for Developmental Toxicity Based on an FDA/TERIS Database. *SAR QSAR Environ. Res.* **2003**, *14*, 83−96.
(22) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity. *Mutagenesis* **2004**, *19*, 365−377.
(23) Todeschini, R.; Consonni, V.; Pavan, M. A Distance Measure Between Models: A Tool for Similarity/Diversity Analysis of Model Populations. *Chemom. Intell. Lab. Syst.* **2004**, *70*, 55−61.
(24) Cronin, M. T. D.; Dearden, J. C.; Gupta, R.; Moss, G. P. An Investigation of the Mechanism of Flux Across Polydimethylsiloxane Membranes by Use of Quantitiative Structure−Permeability Relationships. *J. Pharm. Pharmacol.* **1998**, *50*, 143−152.
(25) Chen, Y.; Yang, W. L.; Matheson, L. E. Prediction of Flux Through Polydimethylsiloxane Membranes Using Atomic Charge Calculations. *Int. J. Pharm.* **1993**, *94*, 81−88.
(26) Chen, Y.; Vayumhasuwan, P.; Matheson, L. E. Prediction of Flux Through Polydimethylsiloxane Membranes Using Atomic Charge Calculations: Application to an Extended Data Set. *Int. J. Pharm.* **1996**, *137*, 149−158.
(27) Cronin, M. T. D.; Aptula, A. O.; Duffy, J. C.; Netzeva, T. I.; Rowe, P. H.; Valkova, I. V.; Schultz, T. W. Comparative Assessment of Methods to Develop QSARs for the Prediction of the Toxicity of Phenols to *Tetrahymena pyriformis*. *Chemosphere* **2002**, *49*, 1201−1221.
(28) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting Models of Toxic Action From Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.* **1997**, *16*, 948−967.
(29) Tetteh, J.; Suzuki, T.; Metcalfe, E.; Howells, S. Quantitative Structure−Property Relationships for the Estimation of Boiling Point and Flash Point Using a Radial Basis Function Neural Network. *J. Chem. Inf. Model.* **1999**, *39*, 491−507.
(30) Dean, J. A. *Lange's Handbook of Chemistry*; McGraw-Hill: New York, 1987.
(31) Dean, J. A. *Handbook of Organic Chemistry*; McGraw-Hill: New York, 1987.
(32) *Mobydigs - Software for Multilinear Regression Analysis and Variable Selection by Genetic Algorithm*, version 1.0 for Windows; Milano Chemometrics and QSAR Research Group: Milano, Italy, 2006.
(33) EC, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation 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. *Official Journal of the European Union, L: Legislation (English Edition)* **2006**, L 396/1 of 30.12.2006.

CI700016D