Evaluation of CADASTER QSAR Models for the Aquatic Toxicity of (Benzo)triazoles and Prioritisation by Consensus Prediction

Stefano Cassani,[1] Simona Kovarich,[1] Ester Papa,[1] Partha Pratim Roy,[1a] Magnus Rahmberg,[2] Sara Nilsson,[2] Ullrika Sahlin,[3] Nina Jeliazkova,[4] Nikolay Kochev,[5] Ognyan Pukalov,[5] Igor V. Tetko,[6] Stefan Brandmaier,[6] Mojca Kos Durjava,[7] Boris Kolar,[7] Willie Peijnenburg[8,9] and Paola Gramatica[1]

[1]QSAR Research Unit in Environmental Chemistry and Ecotoxicology, DiSTA, University of Insubria, Varese, Italy; [2]IVL Swedish Environmental Research Institute Ltd, Stockholm, Sweden; [3]School of Natural Sciences, Linnaeus University, Kalmar, Sweden; [4]IdeaConsult Ltd, Sofia, Bulgaria; [5]University of Plovdiv, Department of Analytical Chemistry and Computer Chemistry, Plovdiv, Bulgaria; [6]Helmholtz-Zentrum München, German Research Centre for Environmental Health, Munich, Germany; [7]Public Health Institute Maribor, Maribor, Slovenia; [8]National Institute of Public Health and the Environment (RIVM), Laboratory for Ecological Risk Assessment, Bilthoven, The Netherlands; [9]Leiden University, Institute of Environmental Sciences (CML), Department of Conservation Biology, Leiden, The Netherlands

Summary — QSAR regression models of the toxicity of triazoles and benzotriazoles ([B]TAZs) to an alga (*Pseudokirchneriella subcapitata*), *Daphnia magna* and a fish (*Onchorhynchus mykiss*), were developed by five partners in the FP7-EU Project, CADASTER. The models were developed by different methods — Ordinary Least Squares (OLS), Partial Least Squares (PLS), Bayesian regularised regression and Associative Neural Network (ASNN) — by using various molecular descriptors (DRAGON, PaDEL-Descriptor and QSPR-THESAURUS web). In addition, different procedures were used for variable selection, validation and applicability domain inspection. The predictions of the models developed, as well as those obtained in a consensus approach by averaging the data predicted from each model, were compared with the results of experimental tests that were performed by two CADASTER partners. The individual and consensus models were able to correctly predict the toxicity classes of the chemicals tested in the CADASTER project, confirming the utility of the QSAR approach. The models were also used for the prediction of aquatic toxicity of over 300 (B)TAZs, many of which are included in the REACH pre-registration list, and were without experimental data. This highlights the importance of QSAR models for the screening and prioritisation of untested chemicals, in order to reduce and focus experimental testing.

Key words: *aquatic toxicity, applicability domain, (benzo)triazoles, consensus, QSAR, REACH, validation*.

Address for correspondence: *Paola Gramatica, QSAR Research Unit in Environmental Chemistry and Ecotoxicology, DiSTA, University of Insubria, Varese, Italy.*
E-mail: *paola.gramatica@uninsubria.it*

Introduction

Triazoles and benzotriazoles ([B]TAZs) are chemicals under investigation in the EU-FP7 Project, CADASTER (**CA**se Studies on the **D**evelopment and **A**pplication of *In Silico* **T**echniques for **E**nvironmental Hazard and **R**isk Assessment; 1). This project aims to integrate quantitative structure–activity (property) relationship (QSA[P]R) models in risk assessment procedures for the EU Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) system (2), and focuses on four classes of emerging chemicals: flame retardants, perfluorinated compounds, fragrances and (B)TAZs.

(B)TAZs are synthetic molecules characterised by the presence of a simple, or condensed, aromatic heterocyclic ring (2C + 3N atoms). These compounds are structurally highly heterogeneous, and are characterised by their different uses and various mechanisms of action. (B)TAZs have many industrial and domestic uses, i.e. as pesticides, pharmaceuticals (e.g. painkillers, and antimycotic and antidepressant medicines), UV-light stabilisers for plastics, anti-corrosives, dishwashing additives, and as components of liquid aircraft de-icing agents (ADFs) and de-icers for airport runways. As a result, they are produced in large amounts and belong to the High Production Volume chemical category.

[a]*Current address: Guru Ghasidas University, Bilaspur, Koni, India.*

Because of their high production volumes, various applications, high water solubility and polarity, and their resistance to biological and photo degradation, these compounds have become ubiquitous contaminants of the aquatic environment (3), mainly in areas surrounding airports (4). Thus, serious concerns about their potential effects on aquatic organisms have recently arisen.

Within the CADASTER project (1), QSA(P)R models, validated according to the Organisation for Economic Co-operation and Development (OECD) principles for QSAR validation and acceptability in regulation (5), were developed to predict the physicochemical and toxicological properties of (B)TAZs. The models developed for melting point, vapour pressure, water solubility and K_{ow} were helpful for predicting the intrinsic tendency of these compounds toward environmental partitioning, and for identifying the (B)TAZs which are potentially more problematic for the aquatic environment (6). Additionally, other QSAR models were developed for predicting the aquatic toxicities of these compounds to an alga (*Pseudokirchneriella subcapitata*; 7), *Daphnia magna*, and a fish (*Onchorhynchus mykiss*; Cassani *et al*., 2013, submitted for publication). These models were also used for the prediction of the aquatic toxicity of more than 300 (B)TAZs for which no experimental data existed, many of which are included in the ECHA (European Chemical Agency) pre-registration list for REACH.

The work described here is the result of a collaboration among different partners involved in the CADASTER project — Insubria University (UI), Swedish Environmental Research Institute (IVL), Linnaeus University (LnU), Helmholtz-Zentrum München (HMGU) and IdeaConsult Ltd (IDEA). This study aims to develop consensus QSAR predictions to be used in regulatory assessments, for endpoints that are among the key data required for risk assessment of chemicals in REACH (i.e. the toxicity values for algae, zooplankton and fish), on the basis of endpoints listed in the Screening Information Data Sets (SIDS) manual for the assessment of chemicals proposed by the OECD (8).

The various research groups involved in this work developed their own models by different methods — Multiple Linear Regression by Ordinary Least Squares (MLR-OLS), Partial Least Squares Regression (PLSR) and Associative Neural Network (ASNN) — and by means of different theoretical molecular descriptors, calculated by using commercial and freely-available software — DRAGON (9), PaDEL-Descriptor (10) and QSPR-THESAURUS web (11). The predictivity of the models was tested on three evaluation sets composed of data from the literature (fish; 12), or data generated within the CADASTER project (for an alga and a daphnid; 13).

However, since each individual QSAR model, with its specific structural and response domain, may be lacking some relevant overall information, the combination of different modelling approaches into a single prediction by consensus was used to complement the deficiencies of one model with the strengths of another. Additionally, the consensus approach, which was derived by calculating averaged predictions from representative and different individual models, allowed the modelling ability of all the molecular descriptors involved in the models, and therefore the influence of the more-particular structures present in the data sets, to be taken into account. The utility of the consensus approach has already been demonstrated in many QSAR studies (14, 15).

Finally, an additional objective of this paper was to demonstrate the utility of QSAR models for screening large data sets, and to identify compounds which would be potentially more problematic in the aquatic environment. This is helpful for reducing the number of experimental tests required, by allowing testing to focus only on the prioritised compounds, which will generate the most useful data.

Materials and Methods

Data sets for QSAR modelling

The endpoints considered were: EC50 (72 hours) for *P. subcapitata*, EC50 (48 hours) for *D. magna,* and LC50 (96 hours) for *O. mykiss*. Experimental data were collected from the FOOTPRINT Pesticide Properties Database (PPDB; 12), which is a database of physicochemical and (eco)toxicological data on pesticides, developed in the context of the EU-FP6 research project, FOOTPRINT. In the database, each data point has been associated with a score related to data quality, which varies between 1 (worst quality data) and 5 (best quality data). In particular, 1 stands for "estimated data with little or no verification", 2 for "unverified data of unknown source", 3 for "unverified data of known source", 4 for "verified data", and 5 for "verified data used for regulatory purposes". Due to the fundamental relevance of the quality of the input data for the performance of QSAR models, and in order to limit the effects of experimental variability, only data corresponding to the highest quality-scores (i.e. 4 and 5) were used for QSAR modelling. The training sets of experimental data were collected for various (B)TAZs, and also for other azo-aromatic compounds (including diazines, triazines) and similar compounds, to enlarge the response and structural domain of the data set. The data included: EC50 data for *P. subcapitata* on 13 (B)TAZs (including one compound with a non-aromatic triazolyl-ring with a thione group) and 18 additional azo-aromatic compounds (including one

compound with a non-aromatic triazolyl-ring with an oxy group), in a total of 31 training set chemicals; EC50 data for *D. magna* on 39 (B)TAZs (including four compounds with a non-aromatic triazolyl-ring with an oxy group) and 51 additional azo-aromatic compounds (including a non-aromatic diazinone), in a total of 90 training set chemicals; and LC50 data for *O. mykiss* on 28 (B)TAZs (including three compounds with a non-aromatic triazolyl-ring with oxy or thione groups) and 49 additional azo-aromatic compounds (including a non-aromatic diazinone), in a total of 77 training set chemicals. Due to the relatively limited amount of data on fish toxicity, data of quality score 3 were also included in this data set, in order to obtain a sufficiently large data set for QSAR modelling. The EC50 and LC50 values (mol/L) were transformed into pEC50 and pLC50 values, by taking the negative logarithm of the values reported.

Testing protocol for P. subcapitata

A 72-hour toxicity test with *P. subcapitata* (formerly known as *Selenastrum capricornutum*) was performed with 13 compounds selected for testing by the CADASTER partner at the Public Health Institute Maribor, Slovenia (PHI). The test was performed according to *OECD Test Guideline (TG) 201, Freshwater Alga and Cyanobacteria, Growth Inhibition Test* (16). The purpose of this test is to determine the effects of a substance on the growth of freshwater microalgae.

Exponentially growing algae were exposed to the test substance in batch cultures over a period of 72 hours. The test endpoint was inhibition of growth, expressed as the logarithmic increase in biomass (average specific growth rate) during the exposure period. From the average specific growth rates recorded in a series of test solutions, the concentration inducing 50% inhibition of the growth rate was determined and expressed as the E_rC50. In addition, the No Observed Effect Concentration (NOEC) was statistically determined. Test cultures, containing the desired concentrations of the test substance and the desired quantity of algal inoculum, were prepared by diluting aliquots of stock solutions of the test substance and of algal suspension with filtered algal medium. The culture flasks were shaken and placed in the culturing apparatus. The cultures were maintained at a temperature of 20 ± 2°C. The cell concentration in each flask was determined at 24, 48 and 72 hours after the start of the test, by using a Perkin Elmer Victor 3, 1420 Multilabel Counter (Perkin Elmer, Singapore, Republic of Singapore). The pH was measured at the beginning of the test and after 72 hours of exposure. The area where the cultures were incubated received continuous, uniform fluorescent illumination, with a light intensity of approximately 5000lux.

Testing protocol for D. magna

A 48-hour toxicity test with *D. magna* was performed by the CADASTER partner at the National Institute for Public Health and the Environment, The Netherlands (RIVM), on 12 selected compounds according to *OECD TG 202*, Daphnia sp., *Acute Immobilisation Test* (17). The purpose of this test is to determine the effects of a substance on the mobility of daphnids.

Young daphnids, aged less than 24 hours at the start of the test, were exposed to the test substance at a range of concentrations, for a period of 48 hours. Five juveniles were used with 20ml of test solution per 100ml glass beaker. Four beakers were used for each test concentration, with a minimum of 5 test concentrations per chemical, excluding the controls/blanks. Immobilisation was recorded at 48 hours and compared with the control values. The results were analysed in order to calculate the EC50 at 48 hours.

Evaluation set for fish and screening data set

An independent test set of LC50 values for *O. mykiss* became available for 18 (B)TAZs (including one compound with a non-aromatic triazolyl-ring with an oxy group; 12), after the development of the QSARs proposed here. This set was used as an evaluation set for external validation of the models.

In addition, a data set composed of data on 386 (B)TAZs (about 10% with a non-aromatic triazolyl-ring), with or without experimental data for the three studied endpoints, was used to screen and prioritise these compounds on the basis of their overall aquatic toxicity. This data set will be referred to in this paper as the 'screening data set'.

Consensus modelling

Consensus predictions were derived by averaging (arithmetic mean) the predictions obtained by individual models.

The performances of all the individual models and of the consensus model were compared by calculating the root mean squared error (RMSE) based on the following equation:

$$RMSE = \frac{\sum_i (y_i - \hat{y}_i)^2}{n} \qquad \text{[Equation 1]}$$

All the models were tested with the evaluation sets (13 chemicals for an alga, 12 chemicals for *D. magna* and 18 chemicals for a fish). In addition, the UI also validated the models, during their development, by splitting the available data.

Furthermore, a 'qualitative' external validation of the QSARs developed for the alga and daphnid

was performed, by using EC50 data converted into four classes of aquatic toxicity, which were derived on the basis of thresholds of E(L)C50 applied by the EU for the categorisation of chemicals hazardous to the aquatic environment (18; Class 1: very toxic EC50 ≤ 1mg/L; Class 2: toxic EC50 ≤ 10mg/L; Class 3: harmful EC50 ≤ 100mg/L; and Class 4: not harmful EC50 > 100mg/L). The agreement among actual and predicted classes was used to evaluate the consistency of the predictions across the modelling methods.

Finally, to verify the agreement among predictions obtained by the different models for each single (B)TAZ, the Mean Absolute Deviation (MAD) of individual model predictions from the consensus prediction was calculated as follows:

$$MAD = \frac{\sum|\hat{y}_i - \hat{y}_C|}{n} \qquad \text{[Equation 2]}$$

Where the numerator $\hat{y}_i - \hat{y}_C$ is the absolute deviation between individual model predictions ($\hat{y}_i$: obtained by UI, IVL, LNU, HMGU and IDEA) and the consensus predictions ($\hat{y}_C$) for each chemical, and n is the number of individual models considered to generate predictions by consensus.

The higher the MAD value, the higher the disagreement among individual model predictions. This analysis was performed by taking into account predictions for all the 386 (B)TAZs, or predictions only for (B)TAZs included in the applicability domain (AD) of all the individual models.

The CAS Registry numbers and IDs of the studied chemicals, the predicted and experimental data, and the MAD values, can be found in the online supplementary information (Appendix A, Tables S1–S3).

Molecular descriptors and modelling methods

The modelled endpoints, molecular descriptors, and procedures used to develop the QSARs and to check the applicability domains of the models by the different CADASTER partners are described in detail in the following paragraphs, and are summarised in Table S4 of Appendix A in the online supplementary information.

University of Insubria (UI)

The commercial software, DRAGON (9), and the freely-available software, PaDEL-Descriptor v. 2.13 (10), were used to generate input molecular descriptors (more than 600 descriptors in DRAGON and more than 350 descriptors in PaDEL-Descriptor), and MLR-OLS models were separately developed for these two groups of descriptors (i.e. 'DRAGON Model' and 'PaDEL-Descriptor Model'). The selection of the modelling descriptors (by Genetic Algorithm; GA), the validation, and the analysis of the ADs of the models, were performed by using the software QSARINS (19).

In a first modelling step, in order to verify the predictive capabilities of the models (20), all the data sets were split before model development into a training set and a prediction set that were used for external validation. Two different splitting techniques were applied: a) by ordered response; and b) by structural similarity with Kohonen Artificial Neural Networks (K-ANN; 21). Several validation techniques were applied, in order to evaluate model robustness; for example, Leave-one-out (LOO) and Leave-many-out (LMO). Different external validation parameters — Q^2_{ext}–F1 (22), which is included in the OECD guidance (5), Q^2_{ext}–F2 (23), Q^2_{ext}–F3 (24) and the Concordance Correlation Coefficient (CCC; 25, 26) — were calculated and compared, in order to select only models that were verified as externally predictive by all the validation criteria. The Y-scrambling procedure was applied, to verify the absence of chance correlation in each model.

In a second modelling step, the equations of the best models developed for the split data sets, verified for their external predictivity, were recalibrated for all the experimental data (Full models), in order to use all the available information related to the modelled responses and structures. The Full models were then proposed for the consensus analysis.

Outliers for the response were identified as those chemicals with a standardised residual in prediction larger than 2.5σ; outliers for the structure (high leverage compounds) had large hat values (diagonal values of the modelling descriptor hat matrix; outliers > $3p'/n$, where p' is the number of the model variables plus one, and n is the number of training compounds; 20). The plot of standardised residuals *versus* hat values (Williams plot) was always examined in combination with the Insubria graph (plot of hat values *versus* predictions; 7), to verify the responses and structural domains of the models. In particular, the Insubria graph was analysed, in order to verify the structural ADs of the models for those (B)TAZs without experimental data, of which there were over 300.

IVL Swedish Environmental Research Institute (IVL)

The descriptors used for QSAR modelling were calculated with DRAGON v. 6.0 (9). PLS regression was used as the modelling method. Two measures can be considered to determine whether a new substance is within the AD of the model. The first is the distance to the model plane (also called the residual magnitude), and the second is the dis-

tance between the model centre and the projection in the model plane. In the SIMCA software (27), which is used for PLS modelling, the distance of a prediction to the model plane is known as DModXPS (Distance to Model in X space for the Prediction Set), while the statistic DModXPS+ also considers the distance in the model plane. From these distances, and the corresponding distances in the training set, it is possible to calculate a probability that a (new) substance is in the AD.

Linnaeus University (LnU)

With the aim of building a QSAR model that delivers predictions with associated uncertainty, a Bayesian approach was used. In order to simplify the fitting and to address the large descriptor space, a Bayesian model was fitted in two steps. First, Latent Variables were generated from the PLS-components. The number of Latent Variables was set to the smallest number for which the cross-validated predictive error was within 10% of the distance from the smallest to the largest error. Second, uncertainty in predictions was estimated by the use of Markov Chain Monte Carlo sampling, with the Bayesian Lasso as the model algorithm. Bayesian Lasso is an example of regularised regression, which penalises large regression coefficients in the loss function.

The predictive performance of the Bayesian Lasso on PLS Latent Variables was similar to the performance of PLS, since the PLS point predictions were, to a large extent, similar to the posterior mean or mode. Thus, the gain of Bayesian Lasso on PLS was a means of assessing uncertainty in terms of a probability distribution. Due to the small data sets, it was not possible to fit Bayesian Lasso directly onto the raw descriptors. The models predictions were derived by using the monomvr package in R (28). QSARs derived by Bayesian-regularised regression (Bayesian Lasso) were based on descriptors from DRAGON 6.0. According to the recommendations on the CADASTER website, descriptors with a standard deviation of less than 0.01, and with only two values different from the others (seen over the training data set), were removed.

Helmholtz-Zentrum München (HMGU)

Several methods and descriptors were used to develop models by using the On-line CHEMical Database and Modelling Environment (OCHEM) platform (29). This analysis included, among others, ASNN (30) and 11 descriptor packages described in our previous publication (31). Models were developed with each set of descriptors individually, by using a bagging with $n = 64$ models. The default values of the parameters provided at the OCHEM website (29) were used for all the analyses. No descriptor selection was performed. However, descriptors that were highly correlated to one another in the training set ($R > 0.95$), or had almost constant values (fewer than two unique values), were excluded. Therefore, the number of descriptors was different for each model, and ranged from 118 for the alga data set, to 132 for the *Daphnia* data set. The performances of all combinations of methods and descriptors were evaluated by using the prediction of 'out-of-the-bag' molecules. These molecules are not used for the development of models in the respective training sets of the bagging approach (e.g. see Breiman [32]). The ASNN method and ADRIANA Code descriptors (33), which overall contributed with one of the lowest RMSEs across three data sets, were selected for further analysis.

The AD of the ASNN method was estimated by using the standard deviation (SD) of the models in the bagging model (34). As an arbitrary threshold for the AD, we used the SD, which covered 90% of the molecules in the training set. For all calculated coefficients, we estimated 95% confidence intervals based on bootstrapping technique with 1000 replicas.

IdeaConsult Ltd (IDEA)

A full range of 0-D, 1-D and 2-D molecular descriptors available in DRAGON 5.4, were calculated, where the input structures were represented as SMILES linear notations. The initial pool of descriptors was reduced by applying the DRAGON built-in variable exclusion procedure. Three types of descriptors were excluded: constant descriptors (relative standard deviation < 0.01%); near-constant descriptors (all values are equal except one); and highly correlated descriptors (i.e. for each pair of highly correlated descriptors, R > 0.9, the descriptor that has the largest mean correlation coefficient with the rest of the descriptors is removed). Thus, a final set of molecular descriptors was used as an input for MobyDigs (35), which performs a more elaborate variable selection procedure by applying a genetic algorithm. The genetic algorithm was carried out by using the following restrictive conditions:

a) $R^2(x,y) > 0.01$;
b) $R^2(x,x) < 0.95$; and
c) Standardised entropy > 0.05,

where x represents any of the inputted descriptors, and y is the endpoint value subject of modelling. Restriction b) was required, because the DRAGON exclusion procedure did not remove all of the highly correlated pairs. If any descriptor violated one of these conditions, it was sent to a tabu list,

i.e. it was not used in the model development process. Q^2 (LOO cross-validation correlation coefficient) was used as a fitness function (model selection criterion), and a maximum of six variables were allowed in the models developed. A set of models with five and six variables was generated, and model selection was performed by analysing the trade-off between the model accuracy and overfitting by the number of model parameters. In addition, several models were created by choosing the descriptors by expert selection. This selection was based on the variables obtained by means of a genetic algorithm, and also included functions of the descriptor variables (e.g. logarithm, square root, power of two, etc.). Model performance was further verified by leave-one-out validation, bootstrapping and y-scrambling procedures, as well as tests performed with external evaluation data sets. The ADs of the models developed were determined by using the leverage approach as implemented in MobyDigs. Compounds with hat values higher than the average hat (mean of the diagonal values of the hat matrix calculated for the training set) were excluded from the AD.

Principal component analysis

Principal component analysis (PCA) is an explorative, multivariate technique that condenses, by linear combination, the relevant information of a group of variables that describes a system. The result is a smaller number of new, highly informative variables, called Principal Components (PCs). PCs are calculated according to the maximum variance criterion, i.e. each successive component covers the maximum of the variance not accounted for by the previous components. The scores of the objects define their ranking along each PC (36, 37). In this study, the toxicities of over 300 compounds to an alga, *Daphnia* and a fish were predicted on the basis of the QSAR models developed. Predictions falling in the ADs of the models were used to perform PCA. The ranking of the compounds along PC1 defined the overall toxicity of (B)TAZs.

Results and Discussion

In the following paragraphs, the individual models developed by each partner on the same training sets, as well as the consensus models of the three endpoints studied, are presented and discussed.

The modelling method, the type and number of modelling descriptors, the statistical performances and the percentage of chemicals of the screened data set of 386 (B)TAZs that are within the AD of each model, are reported for each endpoint in Tables 1–3, and summarised in the online supplementary information (Appendix B).

The predictivity of the individual models and the accuracy of predictions by consensus were evaluated on common evaluation sets composed of the experimental data determined in the CADASTER project (13; for algal and daphnid models), or on literature data (12; for fish models). In addition, UI performed an *a priori* external validation, during the development of the models, based on the splitting of the original experimental data set.

The ADs of all the models were always verified, in order to evaluate the reliability of the predicted data. In particular, the ADs of UI/IDEA-OLS models and LnU/IVL-PLS models were based on molecular descriptors, which mean that they are reflective of the structural domain, while the ADs of the HMGU models were calculated from predicted values and serve as an indication of the potential accuracy in the prediction.

Finally, a PCA (36, 37) of predicted toxicity values in the three organisms of interest was used to determine the (B)TAZs of highest concern on the basis of their overall aquatic toxicity.

Individual models developed for fish toxicity (O. mykiss) and the consensus model

Six QSAR (MLR, PLS-based and ASNN) models were developed, based on DRAGON, PaDEL and ADRIANA Code descriptors for the endpoint pLC50 in *O. mykiss* (Table 1).

The models developed by IVL, LnU and HMGU on 77 training chemicals were based on hundreds of molecular descriptors. The models developed by IDEA and UI for the same training set were more straightforward, as they were based on six and five descriptors, respectively, which had been selected by GA from the input molecular descriptors. A common evaluation set composed of 18 compounds, taken from the available literature data, was used to evaluate the external predictivity of the models proposed for the consensus approach.

As explained in the *Materials and Methods* section, in the modelling approach followed by UI, the available data set was first split (by response and by structural similarity) into two training sets ($N_{TR} = 54$) that were used to develop the models, and two prediction sets ($N_P = 23$) that were used to test the external predictivity. By means of this procedure, it was possible to select two combinations of four modelling variables (DRAGON or PaDEL descriptors), characterised by good predictive ability ($Q^2_{ext} > 0.74$; CCC > 0.87). Subsequently, two Full models were derived based on these variables, for all the 77 compounds, and proposed for the consensus approach.

The statistical performances for all the Full models are reported in Table 1. Additional detailed information regarding the development and the validation of all the models for fish toxicity is

Table 1: Statistical performance of the individual models selected for the consensus model calculated for pLC50 *Oncorhynchus mykiss* (96 hours)

Model/method/descriptors	N_{TR}	N_{EV}	$N_{Desc.}$	R^2	Q^2	$RMSE_{TR}$	$RMSE_{EX}$	Ext. validation on 18 common compounds	AD on 386 (B)TAZs
UI — OLS DRAGON 5.5	77	18	5	0.82	0.79[a]	0.47	0.43	Q^2ext > 0.84, CCC = 0.92	92%
UI — OLS PaDEL-Descriptor	77	18	5	0.76	0.71[a]	0.55	0.40	Q^2ext > 0.86, CCC = 0.92	97%
IVL — PLS DRAGON 6.0	77	18	503	0.89	0.75[a]	0.37	0.43	Q^2ext = 0.85	73%
LnU — BLASSO-PLS DRAGON 6.0	77	18	243	0.70	—	0.62	0.74	Q^2ext > 0.53	84%
HMGU — ASNN ADRIANA Code	77	18	123	0.6 ± 0.1[b]	0.6 ± 0.1	0.73	0.62	Q^2ext = 0.70 ± 0.30	76%
IDEA — OLS DRAGON 5.4	77	18	6	0.84	0.76[a]	0.45	0.28	Q^2ext > 0.93, CCC = 0.94	91%
Consensus	77	18	—	0.85	—	0.44	0.37	Q^2ext > 0.88, CCC = 0.93	53%

[a]Q^2_{LOO}; [b]95% confidence intervals of all coefficients were calculated by using bootstrapping with n = *1000 replicas.*

reported in the Online Supplementary Information (Appendix B). Table S1 lists all the data predicted by each individual model, as well as by consensus, and the information on the AD.

From the results reported in Table 1, it is evident that all the models have fair fitting ability, which was evaluated with the common training set, with the best results obtained by IVL-PLS model, followed by IDEA and UI models (RMSE values range from 0.37 calculated for the IVL model, to 0.62 calculated for the LnU model). The ASNN model shows lower fitting ability, with RMSE values larger than those for the other models (i.e. 0.73). The RMSE value of the consensus model, 0.44, is lower than that of all the individual models, with the only exception of the IVL model which has the best fitting. The trends of experimental and predicted responses for the training set compounds can be observed in Figure S1, which helps to visualise the methods that overestimate or underestimate any given chemical.

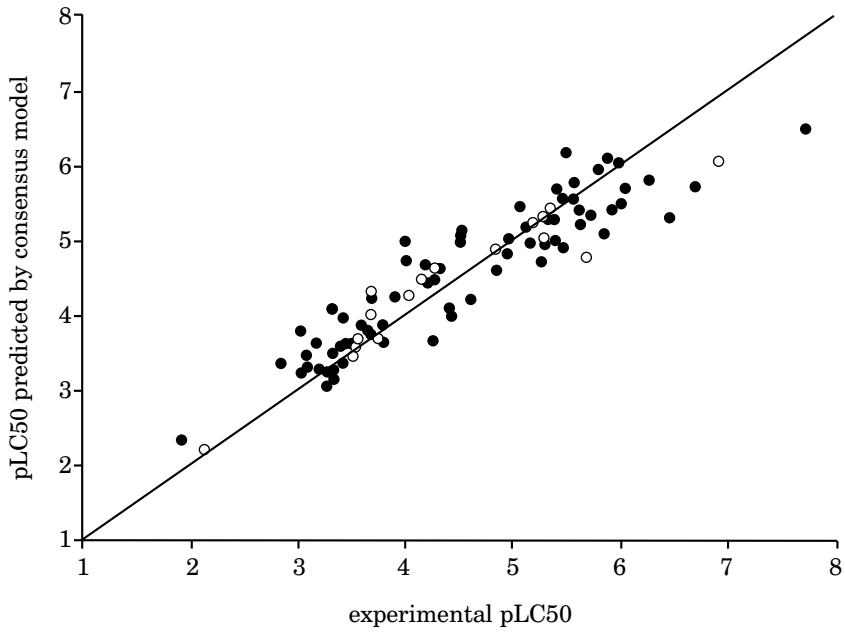
The external validation of the models shows that the QSARs proposed for the consensus have good external predictive ability, in the case of OLS and IVL-PLS models (Q^2_{ext} > 0.84), and satisfying predictive performance calculated for the ASNN model (Q^2_{ext} = 0.7). The LnU model is the only QSAR among those reported in Table 1 with Q^2_{ext} < 0.7. The RMSE value calculated for the consensus model is 0.37, which is lower than the RMSE calculated for the training set. This is an additional demonstration (14, 15) that, despite the difference in variable composition and fitting performance of the individual models in Table 1, their combination provides consensus results in the prediction of external compounds, which are better than the fitting of training chemicals, and, in general, better than the individual models. Only the IDEA model has a lower RMSE value than the consensus ($RMSE_{EX}$ = 0.28). This is probably due to the presence in this model of the descriptor, 'counter of P atoms'. In fact, this descriptor is highly relevant to fit and predict two (B)TAZs (one in the training set, the other in the evaluation set) with a phosphoric group in their structure. However, the specificity of this descriptor improves the performances in prediction of the IDEA model, because it is influenced by the composition of the external evaluation set.

Figure 1 shows the scatter plot of experimental data *versus* pLC50 values predicted by consensus, including also the 18 compounds of the evaluation set. In particular, the evaluation set has a good distribution along the range of values of the training set, and none of the 18 compounds was detected as an outlier.

Individual models developed for algal toxicity (P. subcapitata) *and the consensus model*

Five QSAR (MLR, PLS-based and ASNN) models were developed by UI, IVL, LnU and HMGU, to be used in the consensus approach, starting from DRAGON, PaDEL-Descriptor and ADRIANA Code descriptors for the endpoint pEC50 in *P. subcapitata*. All of these models were externally validated with a common evaluation set composed of 13 compounds, for which experimental data were determined within the CADASTER project. This validation will be addressed in the section, "*External evaluation of algal and daphnid models with new CADASTER experimental data*". The statistical performances of all the models used for the consensus are reported in Table 2. In the modelling approaches followed by IVL, LnU and HMGU with 31 training chemicals, the models were based on hundreds of molecular descriptors, while the UI models were based on only four variables, selected by GA from the input molecular descriptors. In addition, UI split the available data, and developed QSAR models with two independent training sets (N_{TR} = 20 and N_{TR} = 19, respectively), which were evaluated with two prediction sets (N_P = 11 and N_P = 12, respectively). Since the predictive ability of the best split models was considered to be suffi-

Figure 1: Experimental *versus* predicted data from the consensus model for fish toxicity



● = *Fish training set (*n = *77);* ○ = *evaluation set (*n = *18).*

The (B)TAZs of the common evaluation set are labelled differently.

cient (Q^2_{ext} > 0.68; CCC > 0.83), the Full model in Table 2 was calculated for all 31 compounds.

Additional detailed information with regard to the development and the validation of all the algal models, graphs and list of modelling descriptors are summarised in Appendix B.

From the results reported in Table 2, it is evident that the models, even if developed by different methods and by using different descriptors, have satisfactory performances with the training set (RMSE values range from 0.21 calculated for the IVL-PLS model, to 0.58 calculated for the LnU-PLS model). Furthermore, as commonly happens in the consensus approach (14, 15), the RMSE of the consensus model calculated for the training set was lower (i.e. 0.36) than those for all the individual models, with the exception of the IVL model, which had very high fitting power. Figure 2 shows the scatter plot of the experimental data (including the data measured in CADASTER, which will be discussed later) *versus* consensus predictions.

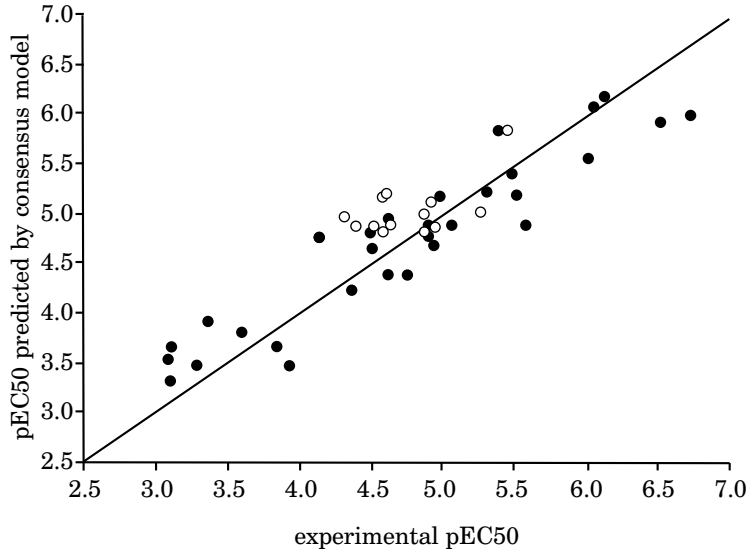
The satisfactory agreement among QSAR predictions can also be seen in Figure S2 (online supplementary information, Appendix A), which shows a

Table 2: Statistical performance of the individual models selected for the consensus model calculated for pEC50 in *Pseudokirchneriella subcapitata* (72 hours)

Model/method/descriptors	N_{TR}	$N_{Desc.}$	R^2	Q^2	$RMSE_{TR}$	AD on 386 (B)TAZs
UI — OLS DRAGON 5.5	31	4	0.85	0.78[a]	0.39	88%
UI — OLS PaDEL-Descriptor	31	4	0.83	0.76[a]	0.41	93%
IVL — PLS DRAGON 6.0	31	375	0.96	0.90[a]	0.21	86%
LnU — BLASSO-PLS DRAGON 6.0	31	242	0.70	—	0.58	85%
HMGU — ASNN ADRIANA Code	31	118	0.70 ± 0.20[b]	0.70 ± 0.20	0.53	96%
Consensus	31	—	0.88	—	0.36	66%

[a]Q^2_{LOO}; [b]*95% confidence intervals of all coefficients were calculated by using bootstrapping with* n = *1000 replicas.*

Figure 2: Experimental versus predicted data from the consensus model for algal toxicity



● = *Alga training set (*n = *31);* ○ = *PHI evaluation set (*n = *13).*

Experimental data measured in CADASTER are labelled differently.

plot of the trend of experimental and predicted responses of the five models for the 31 (B)TAZs in the training set. This figure can be used to visualise the methods which overestimate or underestimate any given chemical.

All the predicted data obtained with each individual model, as well as with the consensus model, and information on ADs, are reported in the Online Supplementary Information (Table S2).

Individual models developed for daphnid toxicity (D. magna) *and the consensus model*

Four linear QSAR models (two OLS-MLR by UI and IDEA, two PLS-based by IVL and LnU) and one non-linear QSAR model (ASNN-based by HMGU) were developed for the endpoint pEC50 in *D. magna*. The statistical performances calculated for the individual models used for the consensus are reported in Table 3.

The models were developed with molecular descriptors calculated by using DRAGON and ADRIANA software. Five or six descriptors were selected by GA and included in the IDEA and UI models, respectively. It should be noted that, while the IDEA model includes a logP descriptor (ALOGP), the UI model is logP-free, in order to guarantee the highest diversity in the description of the structural space. The use of one extra variable in the UI model is justified to compensate for the complexity of the information encoded in the

Table 3: Statistical performances of the individual models selected for the consensus model calculated for pEC50 in *Daphnia magna* (48 hours)

Model/method/descriptors	N_{TR}	$N_{Desc.}$	R^2	Q^2	$RMSE_{TR}$	AD on 386 (B)TAZs
UI — OLS DRAGON 5.5	90	6	0.79	0.75[a]	0.38	89%
IVL — PLS DRAGON 6.0	90	245	0.80	0.74[a]	0.37	57%
LnU — BLASSO-PLS DRAGON 6.0	90	243	0.59	—	0.53	77%
HMGU — ASNN ADRIANA Code	90	132	0.70 ± 0.10[b]	0.70 ± 0.10	0.44	91%
IDEA — OLS DRAGON 5.4	90	5	0.79	0.73[a]	0.38	88%
Consensus	90	—	0.82	—	0.36	48%

[a] Q^2_{LOO}; [b] *95% confidence intervals of all coefficients were calculated by using bootstrapping with* n = *1000 replicas.*

IDEA model by logP, which is a property related to multiple structural features. In contrast, hundreds of descriptors were included in the PLS and ASNN models.

All the models reported in Table 3 were externally validated on the same evaluation set, which in this case, was composed of 12 chemicals tested within the CADASTER project (this topic will be addressed in the next section).

In addition, UI models were originally developed on split data sets (by response and by structural similarity) by using two independent training sets (N_{TR} = 61 and N_{TR} = 60, respectively) and prediction sets (N_P = 29 and N_P = 30, respectively). The actual predictive ability of the best combination of modelling variables selected by GA was satisfying (Q^2_{ext} > 0.71; CCC > 0.84). The Full model developed by UI, in Table 3, was then derived by using these variables for all 90 compounds, and proposed for the consensus approach.

Additional detailed information regarding the development and the validation of all the models developed for *D. magna*, including graphs and a list of modelling descriptors, are summarised in Appendix B.

Also, in this case, models that were based on different descriptors and developed by different methods and different levels of complexity, show similar RMSE values, which range from 0.37 calculated for the IVL-PLS model, to 0.53 calculated for the LnU-PLS model. The satisfactory agreement between these predictions can also be seen in Figure S3. It should be noted that the consensus model has the lowest RMSE value (i.e. 0.36), which means that its fitting ability evaluated with the training set is better than those for each model taken individually. Figure 3 shows the scatter plot of the experimental data *versus* the predictions by consensus, and also includes the CADASTER experimental data used in the common evaluation set and referred to in the next section.

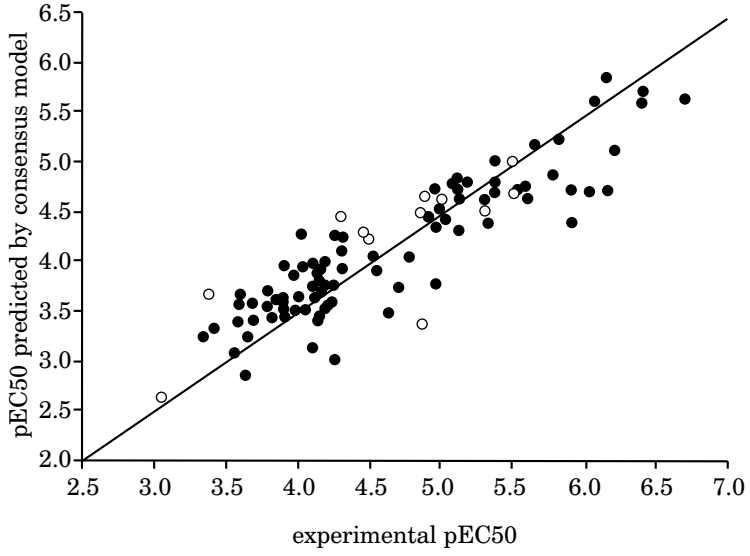
All of the predicted data by each individual model, as well as the consensus model and information on ADs, are reported in the online supplementary information (Appendix A, Table S3).

External evaluation of algal and daphnid models with new CADASTER experimental data

The predictive abilities of each individual algal and daphnid toxicity model and of the consensus models were checked for 13 and 12 (B)TAZs, respectively, which were prioritised and tested within the CADASTER project.

The scatter plot of experimental data *versus* the values predicted in the consensus model for algal toxicity, is shown in Figure 2. The 13 compounds tested in CADASTER by the partner, PHI, are labelled as open circles. As is evident in Figure 2, this evaluation set has a very limited range of

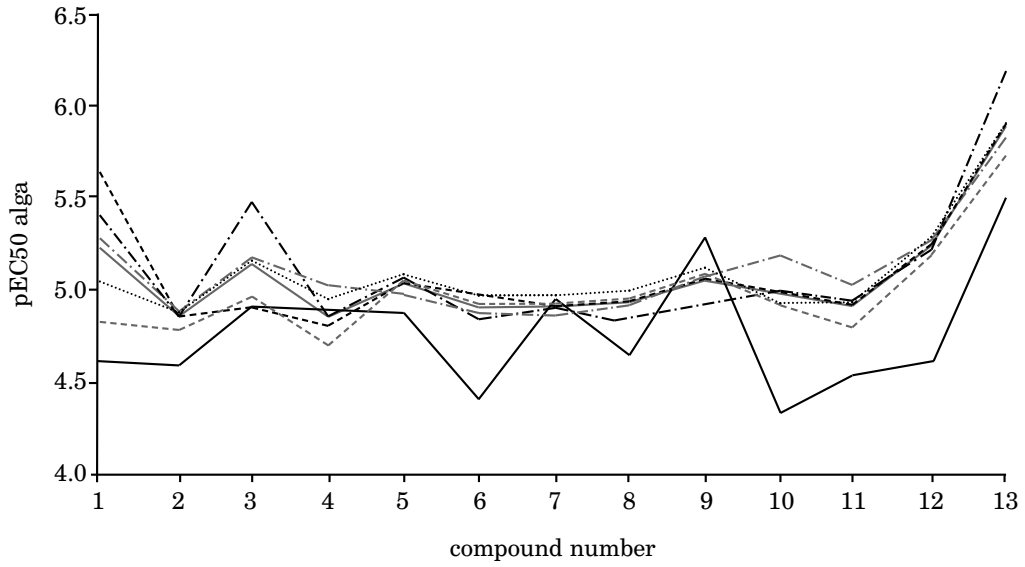
Figure 3: Experimental *versus* predicted data from the consensus model for *Daphnia magna* toxicity



●= Daphnia *training set (*n = *90);* ○= *RIVM evaluation set (*n = *12).*

Experimental data measured in CADASTER are labelled differently.

Figure 4: Comparison of predicted toxicity data from all QSAR models for an alga with experimental data measured in the CADASTER project



——— = *Response;* - - - - - - - - = *UI-DRAGON;* - - - - - - - - = *UI-PaDEL;* ····· = *LNU;* –·–·– = *IVL;* –··–··– = *HMGU;* ——— = *Consensus.*

pEC50 values, i.e. 4.31 to 5.45, compared to the training set, so all of the chemicals exert similar toxicity. This makes the traditional measurements for the evaluation of the external predictivity (Q^2_{ext} and CCC) unsuitable for this comparison, since they are negatively influenced by the distribution of the external data. For this reason, we decided that the best way of comparing model performances for these external predictions was by using the RMSE values (Table 4) and by identifying the classes of toxicity (see below).

As a general comment, all the algal models, even if based on different descriptors and methods, are in reasonable agreement in the prediction of (B)TAZs in the evaluation set (Figure 4). Additionally, the RMSE values calculated for external chemicals are comparable with the RMSE values calculated for the training set (with the exception of the IVL model), which confirms the accuracy of the QSAR predictions for new chemicals. Only four chemicals have errors in prediction larger than 0.5 log units; however, the error of prediction

Table 4: Calculated RMSE and comparison between the training and evaluation set values for an alga and *Daphnia magna* toxicity

***Pseudokirchneriella subcapitata* toxicity measured by PHI (13 compounds)**							
	UI[a]	UI[b]	LnU[a]	IVL[a]	HMGU[c]	IDEA[a]	CONSENSUS
$RMSE_{TR}$	0.39	0.41	0.58	0.21	0.53	—	0.36
$RMSE_{EX}$	0.46	0.32	0.39	0.46	0.44	—	0.40
***Daphnia magna* toxicity measured by RIVM (12 compounds)**							
	UI[a]	UI[b]	LnU[a]	IVL[a]	HMGU[c]	IDEA[a]	CONSENSUS
$RMSE_{TR}$	0.38	—	0.53	0.37	0.44	0.38	0.36
$RMSE_{EX}$	0.53	—	0.6	0.50	0.41	0.62	0.46

[a]*DRAGON descriptors (ver. 5.4 for IDEA; ver. 5.5 for UI; ver. 6.0 for IVL and LnU);* [b]*PaDEL-Descriptor descriptors (ver 2.13);* [c]*ADRIANA Code descriptors.*

remained below 0.7 log units (Appendix B, Table S2).

Three of the test chemicals, namely, myclobutanil (No. 10), epoxiconazole (No. 12) and triazophos (No. 1), include in their structures particular groups and atoms, such as a nitrile group, a phosphate group, an epoxy group and halogen atoms (F, Cl; Appendix A, Figure S4), which are poorly represented (or not present) in the training set used and therefore in the modelling descriptors. In addition, CADASTER experimental data (11.97mg/L; pEC50 4.39) obtained for paclobutrazol (No. 6) vary slightly from the data reported in the literature (7.2mg/L; pEC50 4.61), which are more similar to the QSAR predictions calculated here. Moreover, it is difficult to explain why, according to the CADASTER experiments, paclobutrazol is less toxic than uniconazole-P (No. 8) and diniconazole (No. 9), as they have similar structures (Figure S4) and therefore are predicted by all the QSAR models to have similar toxicities to paclobutrazol.

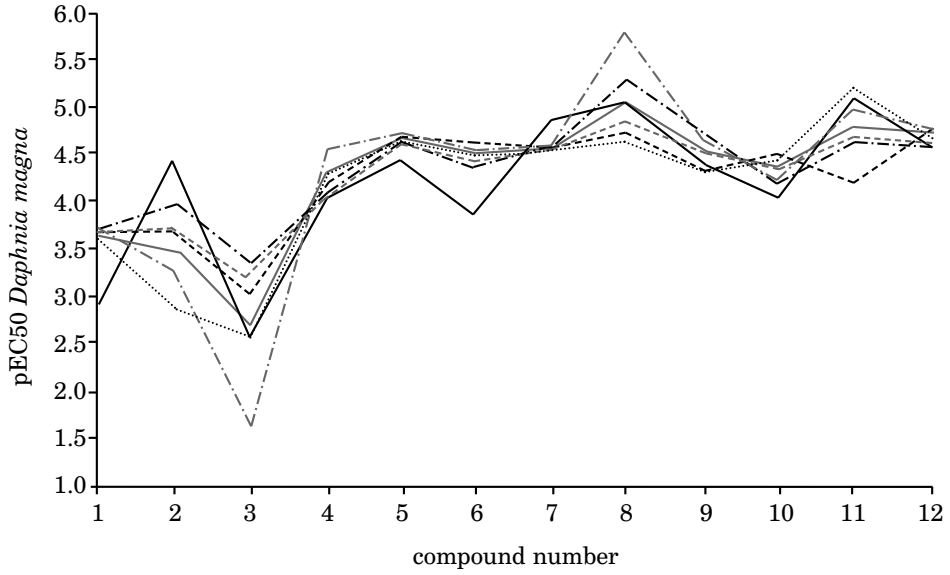
In addition, as explained above, the agreement in predictions across the different modelling methods was assessed by converting experimental and predicted EC50 data into four classes of toxicity (Appendix A, Table S5). All of the models predicted the 13 PHI compounds to be in Class 2 (toxic) with two exceptions: triazophos was predicted to be in Class 1 (very toxic) only by the UI-DRAGON model; and difeconazole, which is the most toxic among the 13 tested compounds, was predicted to be very toxic (in Class 1 instead of experimental Class 2) by all the models. Furthermore, the toxicities of paclobutrazol and myclobutanil, which, according to experimental data measured by PHI, should belong to Class 3 (harmful), were overestimated to be Class 2 (toxic) by all the models. However, it should be noted that, as mentioned above, the toxicity value reported in the literature for paclobutazol (12) — 7.2 mg/L; Class 2 (toxic) — is in agreement with the QSAR predictions calculated here.

The scatter plot of experimental data *versus* the values predicted by the consensus model for *Daphnia* toxicity is shown in Figure 3, where the 12 compounds tested in CADASTER by the partner, RIVM, are labelled as open circles.

The comparison of RMSE values reported in Table 4 shows that all the models are able to predict the RIVM chemicals with satisfactory accuracy (Figure 5). The consensus model has the best RMSE for prediction, after the HMGU model (0.46 and 0.41, respectively).

In contrast to the PHI experimental data for algal toxicity, the RIVM set of data has a larger range of EC50 values. It is important to note that one chemical, guanazole (No. 2) is a strong outlier (Appendix A, Figure S5; residual by consensus: 1.1 log units, Table S3), which is underestimated by all the QSAR models reported in Table 3. This led

Figure 5: Comparison of predicted toxicity data from all QSAR models for a daphnid with experimental data measured in the CADASTER project



——— = *Response;* - - - - - - - - = *UI-DRAGON;* ..... = *LNU;* - - - - - - - - = *IVL;* –·–·–· = *HMGU;* –·–·–·–· = *IDEA;* ——— = *Consensus.*

to an increase in the $RMSE_{EX}$ values for the majority of the models (Table 4).

Moreover, on the basis of the comparison performed on the classification of toxicity, all the QSAR models proposed in Table 3 are in general agreement, and, with the exception of guanazole, which was discussed above, they overestimated (from Class 3 to Class 2) the toxicity of four compounds (Figure 5 and Appendix A, Table S5). Benzotriazole (No. 1) and diclobutrazol (No. 5) were overestimated by all the QSAR models. Paclobutrazol (No. 6) and myclobutanil (No. 9) were classified as toxic by the consensus model, but not by all of the individual QSARs, which are, in these two cases, in disagreement.

Predicted classes of toxicity, derived from predictions calculated by the models reported in Table 2 and Table 3 for the PHI and the RIVM data sets, were then compared with the predictions calculated by ECOSAR (38).

ECOSAR predictions calculated for algal toxicity are in general agreement with individual and consensus predictions, with the exception of three chemicals (No. 3–5 for algal toxicity in Appendix A, Table S5), the toxicities of which were overestimated from Class 2 to Class 1 by ECOSAR. However, it should be noted that overestimation is a minor error in comparison to underestimation, according to the precautionary principle.

The ECOSAR predictions calculated for *Daphnia* are also in general agreement with our predictions, with a few exceptions: tradimefon (No. 4) and cyproconazole (No. 10) were overestimated from Class 3 to Class 2 by ECOSAR; guanazole (No. 2 for *Daphnia* toxicity in Table S5, the outlier referred to above), which was underestimated from Class 2 to Class 3 or Class 4 by all the QSAR models proposed here, and was also underestimated by ECOSAR to Class 4. Based on these results, it is recommended that a deeper analysis of the experimental data that show discrepancies with the data predicted by different methods, should be performed in the future.

Consensus predictions for the aquatic toxicity of 386 (B)TAZs and related compounds

The individual models, reported in Tables 1–3, were used for the prediction of acute toxicity to *P. subcapitata, D. magna* and *O. mykiss* for 386 (B)TAZs and related compounds (the screening data set), with and without experimental data, some of which were also included in the ECHA preregistration list (Tables S1–S3). Since experimental data are not available for the majority of the chemicals considered, particular attention has been paid to the analysis of the AD of the individual models that were used to derive predictions by consensus. The percentage of AD that each model covers is reported in the last columns of Tables 1–3, and the coverage is, in general, satisfactorily high. Thus, the aquatic toxicity endpoints for the majority of these compounds can be reliably predicted by the proposed QSAR models.

Predictions by the consensus model were derived for this screening data set by averaging the predictions obtained in the individual models. MAD values were calculated to quantify the deviation of predictions in individual models from the predictions in the consensus models — range of MAD in *P. subcapitata*: 0.02–2.69, if calculated on the 386 screened compounds, and 0.02–0.52, when considering only the 254 compounds in the AD of all the models; range of MAD in *D. magna*: 0.02–1.5, if calculated on the 386 screened compounds, and 0.02–0.49, when considering only the 186 compounds in the AD of all the models; and range of MAD in *O. mykiss*: 0.05–3.97, if calculated on the 386 screened compounds and 0.05–0.82, when considering only the 205 compounds which fell inside the AD of all the models. The individual predictions and ADs, predictions by the consensus models and MAD values, are reported in Appendix A, Tables S1–S3.

As expected, higher disagreement in predictions among different models is observed for compounds that fall outside the ADs of individual QSARs for which MAD values > 1 log unit were calculated. The predicted values for these chemicals have to be considered as less reliable, since they are extrapolations, and should be used carefully. The following example is valid for all the studied cases: the highest MAD values were calculated for the dye, Direct Orange 41 (ID 22 in Tables S1–S3), which fell outside the AD of all the QSAR models developed here (MAD for *P. subcapitata* toxicity = 2.69; MAD for *D. magna* toxicity = 1.5; and MAD for *O. mykiss* = 3.97).

It is important to note that comparable predictions were obtained for the chemicals included in the AD of all the individual models. The fact that different models, based on different descriptors and/or modelling approaches, led to similar predictions for the newly screened chemicals, adds confidence and reliability to the QSAR predictions obtained by applying the consensus approach.

Prioritisation of (B)TAZs with respect to their overall aquatic toxicity

PCA of predicted toxicity values was applied to characterise the potential overall toxicological profile of (B)TAZs and related compounds in a hypothetical simplified aquatic scenario, by combining the predicted toxicity to the three key organisms studied. This procedure was also useful for identifying potentially hazardous compounds among those analysed. Since only consensus predictions calculated for the chemicals included in the AD of all the individual

models developed for the three organisms were included in the PCA, the potential overall aquatic toxicity was characterised for 128 compounds (including five compounds with a non-aromatic triazolyl-ring with oxy or thione groups; Figure 6).

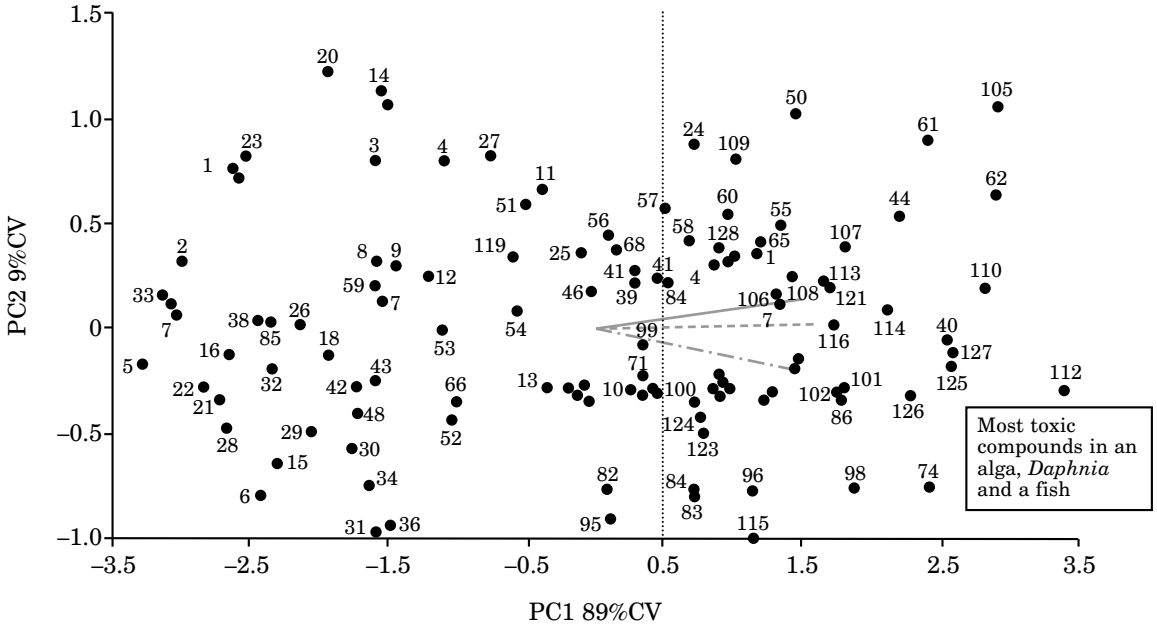
As can be observed in Figure 6, the ranking of chemicals along PC1 (EV = 89%), from left to right, reflects a trend of aquatic toxicity. This ranking separates, on the right, those (B)TAZs (one containing a non-aromatic triazolyl-ring with oxy group) predicted as globally 'more toxic' (i.e. with an overall higher toxicity to the three key organisms in the aquatic scenario analysed), from the less hazardous compounds, on the left. A cut-off was arbitrarily defined on PC1 (PC1 scores > 1.5) to identify the 20 overall most toxic (B)TAZs (one of which containing a non-aromatic triazolyl-ring with oxy group) among more than hundreds of chemicals considered in this analysis (Appendix A, Table S6). For all these hazardous chemicals, the following ranges of toxicity were predicted: *P. subcapitata* EC50 (72 hours) 0.55–2.40 mg/L; *D. magna* EC50 (48 hours) 1.76–16.99mg/L; and *O. mykiss* LC50 (96 hours) 0.36–4.22mg/L. According to EU classification criteria (17), these (B)TAZs can be mainly classified as 'very toxic' when EC(LC)50 ≤ 1mg/L, or 'toxic' when EC(LC)50 ≤ 10mg/L. It should also be noted that, even if five of the prioritised compounds have predicted toxicity values to *D. magna* just above the cut-off value of 10mg/L (and are classified by our models as 'harmful'), the predicted toxicity values to algae and fish are below 2.25mg/L (i.e. toxic or very toxic; Appendix A, Table S6). Thus, the QSAR approach proposed in this paper has been useful to prioritise 20 potentially hazardous (B)TAZs, among over a hundred substances for which experimental tests are necessary to complete the characterisation of their potential toxicological behaviour in water.

Conclusions

In the present study, we propose new QSARs for predicting the aquatic toxicity of (B)TAZs, to the three species that are usually considered when performing risk assessment of chemicals in water: the alga, *P. subcapitata,* the crustacean, *D. magna*, and the fish, *O. mykiss*.

The proposed models have been developed by using different modelling approaches, amount and

Figure 6: Principal Component Analysis (PCA) of predictions from the consensus models for toxicity values in three species for 128 (B)TAZs within the AD of all the individual QSAR models



——— = Daphnia; - - - - - - - = *fish;* —·—·— = *alga.*

····· = *Arbitrary line separating most toxic compounds in alga,* Daphnia *and fish.*

The values shown are for an alga (pEC50 Pseudokirchneriella subcapitata*), a daphnid (pEC50* Daphnia magna*) and a fish (pLC50* Onchorhynchus mykiss*).*

typology of molecular descriptors. The predictions calculated by these individual models, with good fitting performances in all the models and external predictivity in UI models, have been used to perform the consensus approach to improve the predictive ability of the proposed QSARs.

Individual models, as well as results from the consensus approach, have been externally evaluated on experimental data retrieved from the literature or tested within the CADASTER project. The MAD parameter has quantified the agreement between individual predictions and the consensus results, while a comparison among the predicted and actual classes of toxicity demonstrated good agreement among the different QSARs, as well as with the ECOSAR models.

In addition, the QSARs proposed here have been applied to the prediction of the aquatic toxicity of over 300 (B)TAZs without experimental data (some of which are included in the ECHA pre-registration list for REACH), while particular attention was paid to the ADs of the models. As expected, higher disagreement in prediction among different models has been observed for compounds falling outside the ADs of individual QSARs. Predicted values for these chemicals are considered to be less reliable, since they are model extrapolations, and should be used carefully. In contrast, comparable predictions have been obtained for the (B)TAZs included in the ADs of all the models. The fact that different models, based on different descriptors and/or modelling approaches, lead to similar predictions, adds confidence and reliability to QSAR predictions obtained by the consensus approach.

Finally, reliable consensus predictions have been proposed for the assessment of the chemicals studied in a hypothetical, simplified aquatic scenario. Predicted toxicities for algae, daphnids and fish have been combined into a PCA, and the chemicals predicted as being the most toxic have been highlighted for inclusion in a priority list for environmental tests.

This prioritisation and focus on the most dangerous chemicals highlights the fundamental role of QSAR modelling in the screening of compounds that lack experimental data, and in its use by regulators to support weight-of-evidence and non-testing-based approaches for the classification and risk assessment of chemicals. The use of QSARs for REACH has additional relevance, since QSARs can be used to minimise experimental testing, and in particular, testing on animals.

Online Supplementary Information

Appendix A contains Tables S1–S6 and Figures S1–S5. Appendix B includes additional information on individual models by UI, IDEA, LnU and IVL. They are available at www.frame.org.uk.

Acknowledgements

Financial support by the European Union through the project CADASTER FP7-ENV-2007-1-212668 is gratefully acknowledged. Dr Leon van der Wal is acknowledged for his critical comments on the manuscript.

References

1. Anon. (undated). *CAse studies on the Development and Application of* in-Silico *Techniques for Environmental hazard and Risk assessment*. Available at: www.cadaster.eu (Accessed 16.11.12).
2. European Parliament (2006). *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* **L396**, 30.12.2006, 1–849.
3. Wolschke, H., Xie, Z., Möller, A., Sturm, R. & Ebinghaus, R. (2011). Occurrence, distribution and fluxes of benzotriazoles along the German large river basins into the North Sea. *Water Research* **45**, 6259–6266.
4. Cancilla, D., Martinez, J. & Van Aggelen, G. (1998). Detection of aircraft deicing/antiicing fluid additives in a perched water monitoring well at an international airport. *Environmental Science & Technology* **32**, 3834–3835.
5. OECD (2004). *OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure–Activity Relationship Models*, 2pp. Paris, France: 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology. Available at: http://www.oecd.org/dataoecd/33/37/37849783.pdf (Accessed 16.11.12).
6. Bhhatarai, B. & Gramatica, P. (2011). Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. *Water Research* **45**, 1463–1471.
7. Gramatica, P., Cassani, S., Roy, P.P., Kovarich, S., Yap, C.W. & Papa, E. (2012). QSAR modelling is not "push a button and find a correlation": A case study of toxicity of (benzo-)triazoles on algae. *Molecular Informatics* **11–12**, 817–835.
8. OECD (undated). *OECD Cooperative Chemicals Assessment Programme*. Available at: http://www.oecd.org/chemicalsafety/assessmentofchemicals/oecdcooperativechemicalsassessmentprogramme.htm (Accessed 16.11.12).
9. Todeschini, R., Consonni, V., Mauri, A. & Pavan M. (undated). *DRAGON Software*. Milan, Italy: Talete srl. Available at: www.talete.mi.it (Accessed 17.12.12).
10. Yap, C.W. (2011). PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **32**, 1466–1474.
11. CADASTER (undated). *QSPR-THESAURUS Online Platform*. Available at: http://qspr-thesaurus.eu/

login/show.do?render-mode=full (Accessed 16.11.12).
12. University of Hertfordshire (2011). *The PPDB Pesticide Properties Database*. Available at: http://sitem.herts.ac.uk/aeru/footprint/index2.htm (Accessed 16.11.12).
13. Durjava, M., Kolar, B., Arnus, L., Papa, E., Kovarich, S., Sahlin, U. & Peijnenburg, W. (2013). Experimental assessment of the environmental fate and effects of triazoles and benzotriazole. *ATLA* **41**, 65–75.
14. Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Öberg, T., Dao, P., Cherkasov, A. & Tetko, I.V. (2008). Combinational QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis. Journal of Chemical Information & Modeling* **48**, 766–784.
15. Gramatica, P., Pilutti, P. & Papa, E. (2004). Validated QSAR prediction of OH tropospheric degradability: Splitting into training-test set and consensus modeling. *Journal of Chemical Information & Computer Science* **44**, 1794–1802.
16. OECD (2011). *OECD Guidelines for Testing of Chemicals No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test*, 25 pp. Paris, France: Organisation for Economic Co-operation and Development. Available at: http://www.oecd-ilibrary.org/environment/test-no-201-alga-growth-inhibition-test_9789264069923-en (Accessed 26.02.13)
17. OECD (2004). *OECD Guidelines for the Testing of Chemicals No. 202:* Daphnia *sp., Acute Immobilisation Test,* 12pp. Paris, France: Organisation for Economic Co-operation and Development. Available at: http://www.oecd-ilibrary.org/environment/test-no-202-daphnia-sp-acute-immobilisation-test_9789264069947-en (Accessed 16.12.12).
18. European Commission (1991). *Directive 67/548/EEC* (Annex VI). General Classification and Labelling Requirements for Dangerous Substances and Preparations. *Official Journal of the European Union* **L180**, 09.07.1991, 1–79.
19. Chirico, N., Papa, E., Kovarich, S., Cassani, S. & Gramatica, P. (2012). *QSARINS: Software for QSAR MLR Model Development and Validation*, 8pp. Varese, Italy: University of Insubria. Available at: http://dipbsf.uninsubria.it/qsar/news/QSARINS_swGram.pdf (Accessed 15.12.12).
20. Gramatica, P. (2007). Principles of QSAR models validation: Internal and external. *QSAR & Combinatorial Science* **26**, 694–701.
21. Gasteiger, J. & Zupan, J. (1993). Neural networks in chemistry. *Angewandte Chemie International Edition in English* **32**, 503–527.
22. Shi, M., Fang, H., Tong, W., Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C.L. & Sheehan, D.M. (2001). QSAR models using a large diverse set of estrogens. *Journal of Chemical Information & Computer Sciences* **41**, 186–195.
23. Schüürmann, G., Ebert, R.U., Chen, J., Wang, B. & Kühne, R. (2008). External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean. *Journal of Chemical Information & Modeling* **48**, 2140–2145.
24. Consonni, V., Ballabio, D. & Todeschini, R. (2010). Evaluation of model predictive ability by external validation techniques. *Journal of Chemometrics* **24**, 194–201.
25. Chirico, N. & Gramatica, P. (2011). Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *Journal of Chemical Information & Modeling* **51**, 2320–2335.
26. Chirico, N. & Gramatica, P. (2012). Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *Journal of Chemical Information & Modeling* **52**, 2044– 2058.
27. Anon. (undated). *SIMCA*. Umeå, Sweden: Umetrics. Available at: http://www.umetrics.com/simca (Accessed 16.11.12).
28. Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
29. Helmholtz-Zentrum München & eADMET GmbH (undated). *Online Chemical Modeling Environment Project (OCHEM)*. Available at: http://ochem.eu (Accessed 16.12.12).
30. Gasteiger, J. (2006). Of molecules and humans. *Journal of Medicinal Chemistry* **49**, 6429–6434.
31. Sushko, I., Novotarskyi, S., Korner, R., Pandey, A.K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V.V., Tanchuk, V.Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I.I., Palyulin, V.A., Radchenko, E.V., Welsh, W.J., Kholodovych, V., Chekmarev, D., Cherkasov, A., Aires-de-Sousa, J., Zhang, Q.Y., Bender, A., Nigsch, F., Patiny, L., Williams, A., Tkachenko, V. & Tetko, I.V. (2011). Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *Journal of Computer-aided Molecular Design* **25**, 533–554.
32. Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.
33. Tetko, I.V. (2008). Associative neural network. *Methods in Molecular Biology* **458**, 185–202.
34. Sushko, I., Novotarskyi, S., Körner, R., Pandey, A.K., Cherkasov, A., Li, J., Gramatica, P., Hansen, K., Schroeter, T., Müller, K.R., Xi, L., Liu, H., Yao, X., Öberg, T., Hormozdiari, F., Dao, P., Sahinalp, C., Todeschini, R., Polishchuk, P., Artemenko, A., Kuz'min, V., Martin, T.M., Young, D.M., Fourches, D., Muratov, E., Tropsha, A., Baskin, I., Horvath, D., Marcou, G., Muller, C., Varnek, A., Prokopenko, V.V. & Tetko, I.V. (2010). Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *Journal of Chemical Information & Modeling* **50**, 2094–2111.
35. Todeschini, R., Ballabio, D., Consonni, V., Mauri, A. & Pavan, M. (2004). *MobyDigs 1.0*. Milan, Italy: Talete srl. Available at: www.talete.mi.it (Accessed 17.12.12).
36. Jackson, J.E. (1991). *A User's Guide to Principal Components*, 592 pp. New York, NY, USA: Wiley.
37. Anon. (1995). *SCAN: Software for Chemometric Analysis, ver. 1.1 for Windows*. State College, PA, USA: Minitab Inc.
38. US EPA (2012). *Standalone Version 1.11 of ECOSAR*. Washington, DC, USA: US Environmental Protection Agency. Available at: http://www.epa.gov/oppt/newchems/tools/21ecosar.htm (Accessed 16.12.12).