The QSPR-THESAURUS: The Online Platform of the CADASTER Project

Stefan Brandmaier,[1] Willie Peijnenburg,[2,3] Mojca K. Durjava,[4] Boris Kolar,[4] Paola Gramatica,[5] Ester Papa,[5] Barun Bhhatarai,[5] Simona Kovarich,[5] Stefano Cassani,[5] Partha Pratim Roy,[5] Magnus Rahmberg,[6] Tomas Öberg,[7] Nina Jeliazkova,[8] Laura Golsteijn,[9] Mike Comber,[10] Larisa Charochkina,[11] Sergii Novotarskyi,[12] Iurii Sushko,[12] Ahmed Abdelaziz,[12] Elisa D'Onofrio,[5,13] Prakash Kunwar,[13] Fiorella Ruggiu[13] and Igor V. Tetko[12,13,14]

[1]Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Research Unit of Molecular Epidemiology, Institute of Epidemiology II, Munich, Germany; [2]National Institute of Public Health and the Environment, Centre for Safety of Substances and Products (RIVM), Bilthoven, The Netherlands; [3]Leiden University, Institute of Environmental Sciences, Department of Conservation Biology, Leiden, The Netherlands; [4]National Institute for Health, Environment and Food, Maribor, Slovenia; [5]University of Insubria, QSAR Research Unit in Environmental Chemistry and Ecotoxicology, DiSTA, Varese, Italy; [6]IVL Swedish Environmental Research Institute Ltd, Stockholm, Sweden; [7]School of Natural Sciences, Linnaeus University, Kalmar, Sweden; [8]IdeaConsult Ltd, Sofia, Bulgaria; [9]Radboud University Nijmegen, Institute for Wetland and Water Research, Department of Environmental Science, Nijmegen, The Netherlands; [10]Mike Comber Consulting Ltd, Devon, UK; [11]Institute of Bioorganic and Petrochemistry, Kyiv, Ukraine; [12]eADMET GmbH, Garching, Germany; [13]Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, Munich, Germany, [14]Chemistry Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

Summary — The aim of the CADASTER project (**CA**se Studies on the **D**evelopment and **A**pplication of *in **S**ilico* **T**echniques for **E**nvironmental Hazard and **R**isk Assessment) was to exemplify REACH-related hazard assessments for four classes of chemical compound, namely, polybrominated diphenylethers, per and polyfluorinated compounds, (benzo)triazoles, and musks and fragrances. The QSPR-THESAURUS website (http://qspr-thesaurus.eu) was established as the project's online platform to upload, store, apply, and also create, models within the project. We overview the main features of the website, such as model upload, experimental design and hazard assessment to support risk assessment, and integration with other web tools, all of which are essential parts of the QSPR-THESAURUS.

Key words: *CADASTER, hazard assessment, multimedia models, online tools, fate assessment*.

Address for correspondence: *Igor V. Tetko, Institute of Structural Biology, Helmholtz Zentrum München — German Research Centre for Environmental Health (GmbH), Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany*.
E-mail: *itetko@vcclab.org*

Introduction

Implementation of the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation (1) requires demonstration of the safe manufacture and use of chemicals. REACH aims to achieve a proper balance between societal, economic and environmental objectives, and attempts to efficiently use the scarce and scattered information available on the majority of substances. Thereupon, REACH aims to reduce animal testing through the optimised use of *in silico* and *in vitro* information on structurally-related compounds.

The REACH regulation advocates the use of non-animal testing methods, but guidance is needed on how these methods and the resulting data should be used. This includes alternative methods, such as chemical and biological read-across, *in vitro* results, *in vivo* information on analogues, (Quantitative) Structure–Activity Relationships ([Q]SARs), and exposure-based waiving. The concept of Intelligent Testing Strategies for regulatory endpoints has been outlined to facilitate the assessments. However, intensive efforts are needed, to translate the concept into a workable, consensually acceptable and scientifically sound strategy.

The aim of the CADASTER project (2) was to provide the practical guidance to support integrated risk assessment by carrying out hazard assessments for chemicals belonging to four classes:

— *Brominated flame retardants (BFRs)*: a class of hydrophobic chemicals that are incorporated in a variety of consumer products (e.g. electronic

devices, building materials, and textiles) to increase their fire resistance. The focus of CADASTER was on polybrominated diphenyl ethers (PBDEs).

— *Per and polyfluorinated compounds (PFCs)*: a class of synthetic substances widely used in different materials, such as waterproof fabrics, food packaging, non-adhesives, fire-fighting foams, and paints. PFCs studied during the project included both linear and aromatic chemicals, with different carbon-chain lengths, fluorination degrees (per and polyfluorinated compounds) and functional groups (carboxylates, sulphonates, sulphonamides, and alcohols).

— *Substituted musks/fragrances*: a heterogeneous group of chemicals of varying composition, including substituted benzophenones, polycyclic musks, salicylates, cinnamates and other esters with fragrance behaviour, and terpene derivatives.

— *Triazoles and benzotriazoles ([B]TAZs)*: used as components of many pesticides and pharmaceuticals (e.g. painkillers, and antimycotic and antidepressant medicines); they are also abundantly used as components of liquid de-icing agents for aircraft and airport runways, and as UV stabilisers for plastics.

The main goals of the project were to exemplify the integration of information, models and strategies for carrying out safety, hazard and risk assessments for large numbers of substances, and to show how to increase the use of non-testing information for regulatory decisions, whilst meeting the main challenge of quantifying and reducing uncertainty.

In order to achieve this challenging goal and to offer a working platform both for participants within the CADASTER project and for external users, the QSPR-THESAURUS was developed (3). The QSPR-THESAURUS is based on the Online Chemical Modeling Environment (OCHEM; 4). It contains numerous additional features, such as tools for fate, hazard and risk assessment (5), and an interface to the experimental design methods (6–8).

Materials and Methods

The QSPR-THESAURUS is based on OCHEM — "a web-based platform that aims to store data, automate and simplify the typical steps required for QSAR modeling" (4). Several components of OCHEM were reduced in their functionality, whereas new features required by the project, e.g. the possibility to upload models or predicted properties, were implemented (see Table 1 for an overview).

The main motivation was to simplify the OCHEM interface without compromising the functionality required by the project. Further modifications included the implementation of modules for risk and hazard assessment and experimental design, and ensuring support of the QSAR Model Reporting Format (9; QMRF), which is required for the use of models for regulatory purposes. Detailed information about these modules is provided in the following paragraphs.

Database

The database module (Figure 1) is required, in order to submit, collect and annotate experimental records obtained from the literature. It stores data in the original units, tracks users and any modifications they perform to the data, and allows the introduction of new units and properties. The database automatically checks for duplicates, allows the editing of single, or several, records simultaneously, and performs a batch upload of data as Microsoft Excel® and/or SDF files. Furthermore, it permits unlimited export of data as Excel, CSV or SDF files, and each entry requires a literature reference, which allows tracing of the original source to enable a quality check.

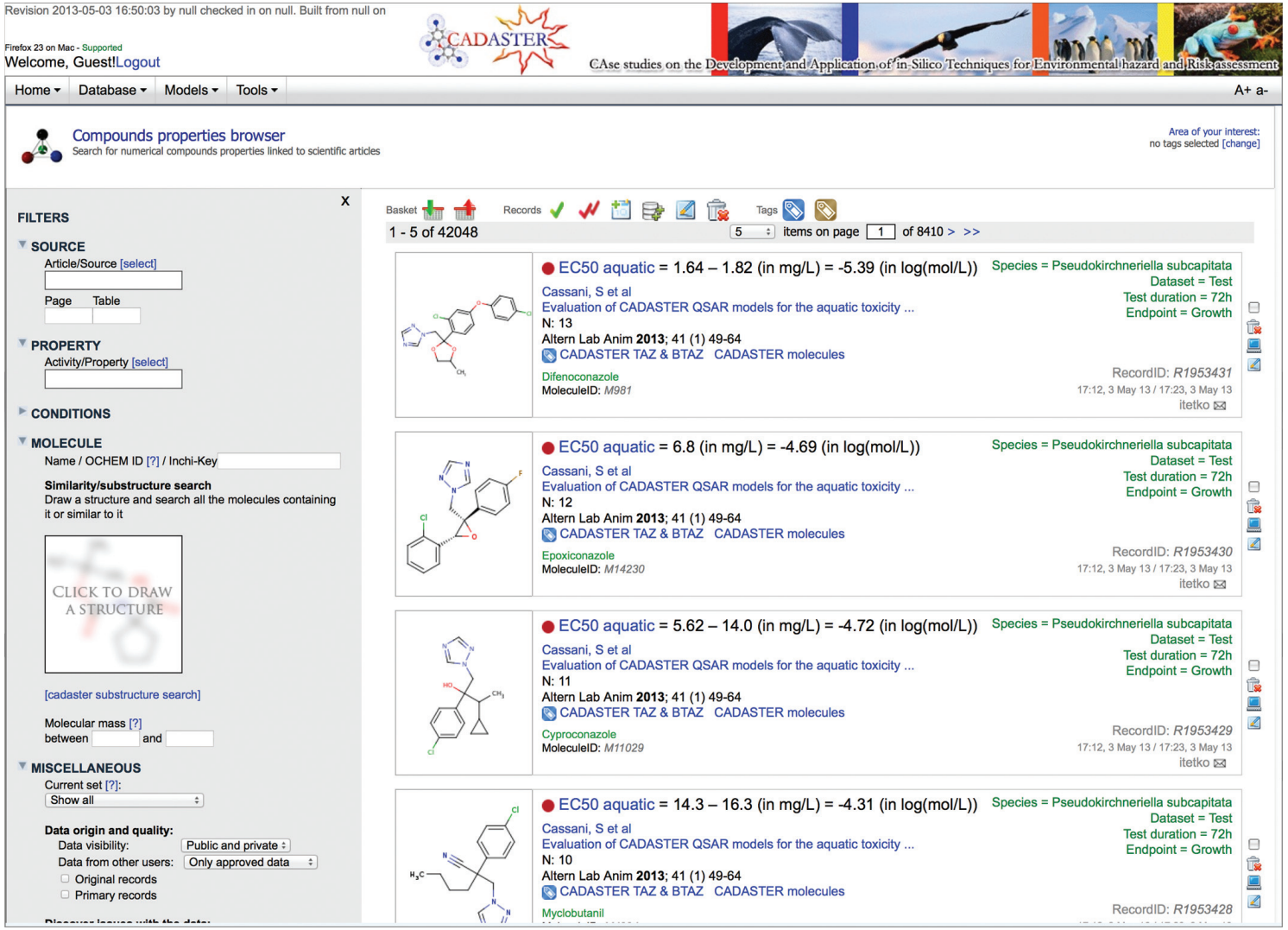
An important feature of the database is the possibility to store the conditions of the experiments. This information is crucial for modelling — in many cases, the result of an experimental measurement is senseless, unless the conditions

Table 1: Differences between OCHEM and the QSPR-THESAURUS

	QSPR-THESAURUS	OCHEM
Predefined views using TAGs	Yes	No
Storage of measured properties	Yes	Yes
Free download of data	Yes	No
Descriptor package	Partial	Full
Export of descriptors	Yes	Yes
Tox alerts	No	Yes
Model development	Yes	Yes
Machine-learning packages	3	14
Model upload	Yes	No
Model application	Yes	Yes
Storage of predicted values	Yes	No
Experimental design	Yes	No
Risk + hazard assessment	Yes	No
QMRF interface	Yes	No

Whereas the compound storage module has not undergone any changes, the variety of provided descriptor sets and the machine learning algorithms were reduced, compared to OCHEM.

Figure 1: The experimental measurements browser window



The left-hand side shows the filtering panel, whereas the measurements are presented on the right. Information about experimental conditions, literature references and a graphical representation are given.

under which the experiment has been conducted are known (e.g. boiling points should be reported together with the pressure). The values stored in the database can be numerical (with units of measurement), qualitative or descriptive (textual).

The QSPR-THESAURUS provides a convenient grouping of data by using tags developed for each of the four classes analysed. Therefore, it allows the users to see only data that are relevant for the class of compound under consideration.

Models

The main purpose of the QSPR-THESAURUS was to store models developed by the project participants. Indeed, it was noted that different research groups used different approaches for the development of models (including proprietary software), which could not be standardised and integrated, due to limitations imposed by the corresponding licensing agreements. The tools for model development include Associative Neural Network (ASNN), as well as Linear Regression and Partial Least Squares approaches. The last two methods were also used as part of the model upload pipeline and experimental design tools. The application of these tools to the creation of models is covered by the CADASTER tutorial (10). This tutorial provides all the steps that are required to reproduce a model, which uses data available in the QMRF Inventory (11).

Model upload

The majority of the models developed by the project participants were linear models, with the exception of those contributed by the German Research Centre for Environmental Health, which were developed by using the ASNN method. Therefore, a tool to upload linear models was developed. To use this tool, users need to provide a training set and (optional) test sets for the model (in the form of a 'basket' of records in QSPR-THESAURUS), as well as a specially prepared Excel spreadsheet containing the names of the descriptors used and the respective linear coefficients. The description of the Excel file and of the model upload procedure are provided on the Wiki page (12).

Applicability domain calculation

All the uploaded, or developed, models include an estimation of the accuracy of the prediction. The estimation of the accuracy of neural network models, which is also published on the website of the CADASTER project, is based on the concept of 'distance to a model' (DM), which was proposed and conceptualised by the project participants (13, 14). Several DMs are supported: the standard deviation of an ensemble of models (STDEV), the correlation in the space of models (CORREL; 15), and leverage. The DMs are calibrated against the accuracy of models for the training set, by using a cross-validation procedure (16).

The estimated accuracy of the predictions as a function of the respective DM is visualised on the accuracy-averaging plot (Figure 2). The DMs are used to estimate the prediction accuracies for new molecules. The same methodology was extended by the project participants for classification models (17).

The Williams plot (see Figure 3), which reports the standardised residuals in the y-axis and the hat values (h) in the x-axis, is used to demonstrate the response and structural applicability domain (AD) of linear models. The leverage threshold (h*) is used to identify predictions that are outside the AD of the linear models in the descriptor space. The predictions outside the AD in the response space are defined according to the distribution of experimental values in the training set.

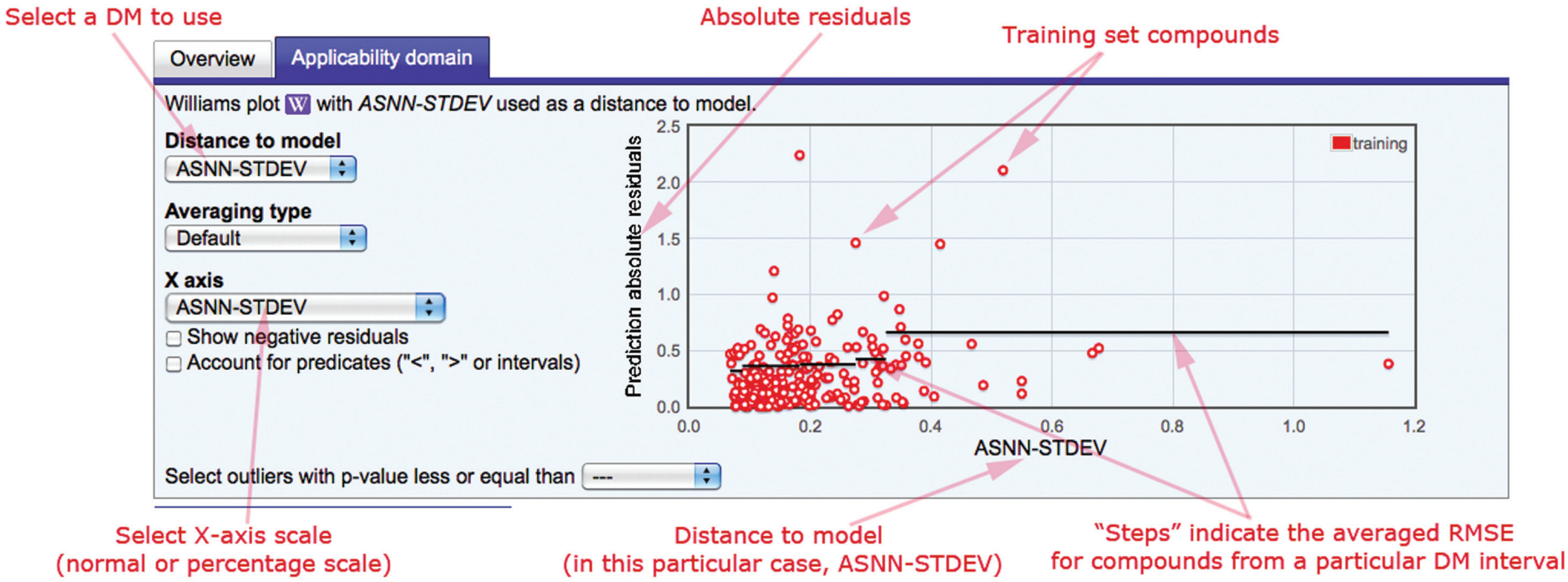
In addition, the Insubria graph (18), which is a modified version of the Williams plot, is provided for linear model predictions (see Figure 4). On this plot, the y-axis corresponds to the predicted value (rather than the standardised residual, as in the Williams plot).

Reporting of models

The CADASTER webpage allows users to provide information about their models, by using the QMRF, which is a standardised way to provide information about QSAR models. This provides users with exhaustive information about the model according to principles set out by the Organisation for Economic Co-operation and Development (OECD; 19), and is a requirement for the use of models for regulatory purposes.

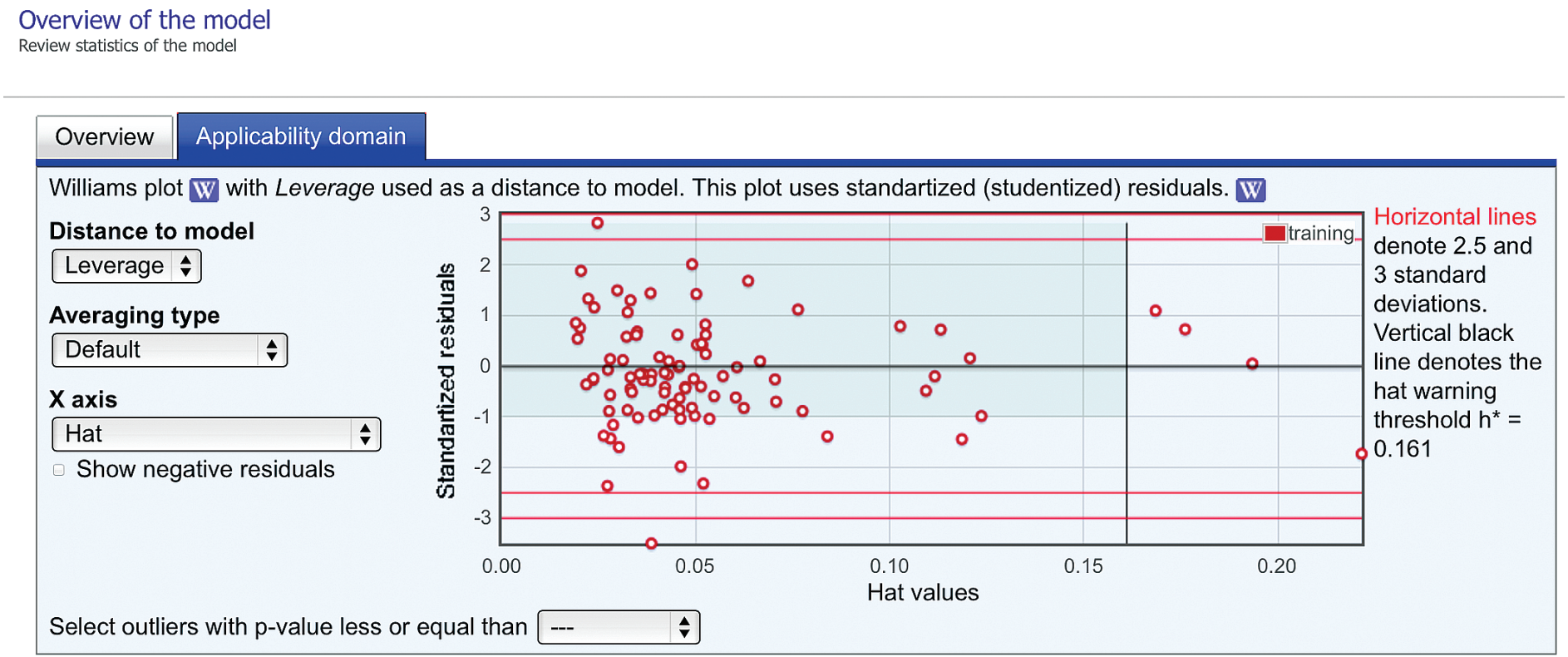
The implemented QMRF enables the user to examine and edit previously uploaded QMRFs, and/or upload or create new ones from scratch. It is also possible to clone existing QMRFs, in order to use them as templates for new ones. This is especially helpful when several models refer to the same publication, as they share substantial information about the authors, used data, methodology, etc. The QMRFs can be exported and imported as XML files, which are fully compatible with the AMBIT editor and the JRC database (11). The QMRF can also store additional information about the model as images. An online viewer was developed and integrated in QSPR-THESAURUS to display it for the published models.

Figure 2: The graphical representation of the AD-estimation in the QSPR-THESAURUS



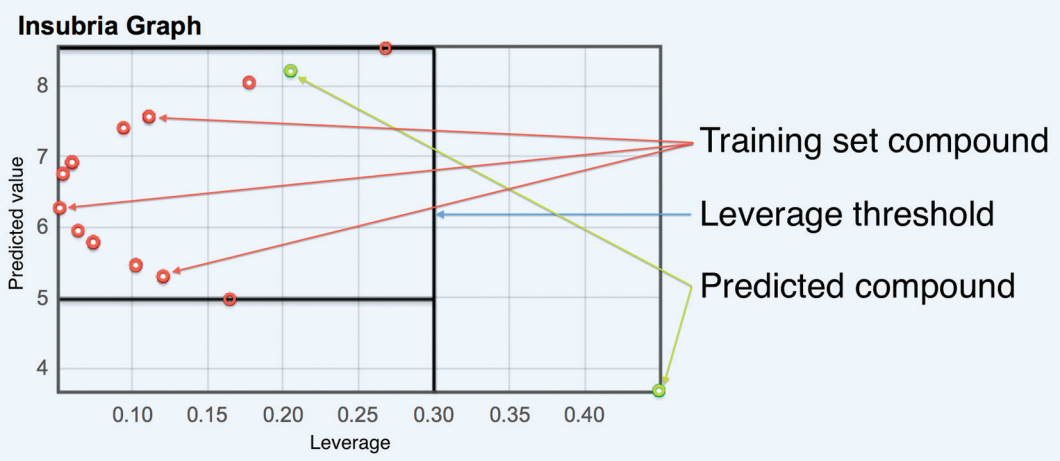
The distance to model is displayed on the x-axis, whereas the y-axis displays the absolute residuals.

Figure 3: The Williams plot



Horizontal lines denote 2.5 and 3 standard deviations. The vertical line denotes the hat warning threshold h.*

Figure 4: The Insubria graph is a variation of the Williams plot for prediction results on chemicals without experimental values



Use of the web tool for application of the models

The QSPR-THESAURUS browser allows users to select the model of interest, which can be used to predict compounds by employing one of the following options:

— upload an SDF file with structures;

— provide SMILES or name of the compound;

— draw the structure on an interactive editor; or

— select a previously prepared basket or tag.

The predictions are shown in the browser of calculated values, together with information on whether they are within the AD of the model. The predicted values can be exported as Excel, SDF or CSV files.

Application of the models by using the standalone tools

While the use of a web tool could be sufficient for many users, others may need to integrate such calculations as part of their workflow and automated data processing engines. In order to facilitate such use, Simple Object Access Protocol (SOAP) web services were developed, which enable an automatic prediction of molecules on the QSPR-THESAURUS website. Several examples were developed, e.g. based on Java and Perl, to show how the web services could be used to develop standalone tools. Detailed information and sample implementations can be found elsewhere (20).

Experimental design

The QSPR-THESAURUS offers online tools for experimental design, which include the D-Optimal design (21) and its adaptive variation, the PLS-Optimal approach (2, 6, 8). The latter approach works on PLS latent variables, and provides property-specific compound selection. The experimental design is accessible from the 'Tools' dropdown menu. The Wiki page (22) describes the steps required to perform the experimental design for new compounds. The calculation results can be downloaded in a number of different formats.

Hazard assessment

The Species Sensitivity Distribution (SSD) approach (23, 24) was implemented, based on R code developed by the partners at the School of Natural Sciences, Linnaeus University (25, 26). It treats sensitivities of observed species as random samples from the ecosystem. The approach fits an SSD to logEC50 values for different species, in order to derive the hazardous concentration, considering uncertainty, from the sample size (27). Uncertainty in the hazardous concentration, based on point predictions of QSAR models, is derived as the non-central Student's *t*-distribution. The QSPR-THESAURUS provides a number of models for aquatic toxicity, developed by the project participants, to be used for the SSD calculations (28–30). The users can also provide predicted values calculated with their favourite models.

Environmental fate assessment

The estimation of the Predicted Environmental Concentration (PEC) is performed by using SimpleBox (31), which was developed by RIVM and is provided as a Visual Basic program for Excel. The project dedicated substantial effort

toward its integration and the development of a server, which currently runs on a Windows virtual machine (28). SimpleBox requires a number of physicochemical properties as input, and models to predict these properties were made available on the QSPR-THESAURUS website according to our recent publication (32). Monte-Carlo simulations, which are based on the uncertainty of estimated properties, are run to produce a distribution of PEC values, and calculate percentiles to be used for risk assessment. Once these parameters are entered, the calculation is started and the results are shown. In addition, detailed results of the sampling can be downloaded as a CSV file.

Risk assessment

The risk assessment tool was developed as a result of the integration of the SSD with the Environmental Fate assessment tool. It was used to exemplify the risks of environmental pollutants following the release of a chemical substance into the environment.

Results

Collection of data

A data search on all endpoints of relevance was performed for the environmental risk and hazard assessment of the groups of chemicals included in the case studies. Physicochemical properties, environmental fate parameters, and aquatic and terrestrial ecological effect parameters were included. Ecological effect parameters of interest were, among others, the available toxicity data. This task was carried out by means of a literature search, supplemented with searches of existing databases on risk and hazard assessment parameters, including IUCLID and AQUIRE. Thereupon, additional data were collected from industry sources (DuPont and the Research Institute for Fragrance Materials, Inc.) and regulatory agencies.

Only limited amounts of data were found for compounds of the four classes under investigation. Therefore, the search was widened to include data on chemicals that were considered relevant for modelling — e.g. flame retardants that are not PBDEs, polyfluoro-compounds in addition to perfluoro-chemicals, PFCs, etc. This expansion of the data was essential to permit studies making use of read-across techniques (33), as well as QSAR model development, to be performed.

Subsequent to the evaluation of the available experimental data and (Q)SAR models, new data were generated on endpoints and chemicals for which insufficient data were available (34). Test compounds were selected according to structural coverage, toxicity patterns, physicochemical properties, REACH-relevance, and the availability of analytical techniques. Toxicity as well as fate and behaviour testing were performed at the National Institute for Health, Environment and Food, Maribor, and at RIVM. The experiments performed included bioaccumulation testing of polybrominated diphenylethers with the aquatic oligochaeta *Tubifex tubifex*, and toxicity testing of perfluoroalkylated substances, and their transformation products, with lettuce (*Lactuca sativa*) and a green alga (*Pseudokirchneriella subcapitata*). Thereupon, testing of perfluoroalkylated substances was performed with two cladoceran species (*Daphnia magna* and *Chydorus sphaericus*), as well as with zebrafish (*Danio rerio*) embryos. Toxicity testing of substituted musks/fragrances was performed with the green alga (*P. subcapitata*) and with *D. magna*, and their readily biodegradability was tested according to OECD guideline 301 D (closed bottle test). Toxicity testing of substituted (benzo)triazoles was performed with *D. magna*, zebrafish (*D. rerio*) embryos and the green alga (*P. subcapitata*); substituted (benzo)triazoles were also tested to assess their readily biodegradability using the same OECD 301 D closed bottle test.

The results of these studies are summarised elsewhere (29, 34–40). The data obtained were uploaded to the website of the CADASTER project. The database provides a collection of 5,440 experimental records for 120 properties, collected from more than 544 sources for four classes of emerging chemicals, and which were also used in publications by the CADASTER project partners (29, 30, 32, 41–56).

Development of QSAR and QSPR models

A preliminary analysis of the literature showed that the majority of the published QSARs had not been externally validated, and/or did not specify their ADs. Thus, they did not fulfil the OECD principles for QSAR validation for regulatory applicability (19) and were of limited utility for the specific classes of compound studied under this project.

Following the development of the website, the models developed and published by the project members were uploaded to the QSPR-THESAURUS site. We observed that models developed by using 3-D descriptors were difficult to reproduce, since the procedures used for structure optimisation (e.g. the use of different initial conformations that can be found by using molecular mechanics approaches or generated manually) could not be exactly reproduced, especially for flexible molecules with many degrees of freedom. Two

approaches were implemented to address this problem and ensure reproducibility of the predictions.

Firstly, we extended the QSPR-THESAURUS to also include the possibility of uploading values calculated by the models, as well as leverage values and information, regardless of whether a molecule is within or outside the structural AD of a model. This allowed the reproduction of predictions for compounds that were described in publications exactly as they were published. The uploaded calculated values were linked to the models they referred to.

Secondly, we implemented a structural database (57) by using the BOINC framework (58). Users can submit compounds of interest, which are then optimised with the semi-empirical tool MOPAC (59), which was used by the partners. The descriptors calculated from the optimised structures allow the development of 3-D models, which are reproduced on the website of the project.

The models available on the QSPR-THESAURUS website cover both environmental toxicity and physicochemical properties for the classes of chemical compounds analysed. The toxicity models focused on EC50 (and LC50, IC50) for various tests and organisms, concerning various effects such as immobility or growth. Most of the models were developed by using linear methods including OLS (55%) and PLS (6%). Other models were developed by using the ASNN method. The majority of the uploaded models were contributed to by the QSAR Research Unit in Environmental Chemistry and Ecotoxicology at the University of Insubria and by the German Research Centre for Environmental Health.

The models published on the QSPR-THESAURUS website were integrated with OpenTox ToxPredict (http://toxpredict.org) by using web services. The information provided through this portal includes predicted values, the accuracy of prediction, as well as whether the given prediction is inside the AD of the model. This validated the developments of the QSPR-THESAURUS web services, and provided an important means of disseminating project-related information.

Evaluation of the experimental design approaches

Several experimental design approaches were developed. The main focus within these approaches, which were presented in several studies, was to show the benefits of stepwise, adaptive approaches (6, 8, 60) for endpoints, as they allow successive testing phases. The consideration of the incrementally-accumulating information about the target property was shown to be conducive for all of the endpoints examined. The combination of PLS latent variables with the D-Optimal design was shown to significantly improve the predictive quality of the developed models, as compared to the predictive quality of models derived from classical experimental design approaches. The same observations were made when applying a similarity-based approach to selected descriptors and the utilisation of predicted properties to span the chemical space.

Risk assessment by using the web tool

The risk assessment tool developed within the project provided a practical guidance to QSAR-integrated risk assessment, by exemplifying the integration of information, models and strategies for carrying out safety, hazard and risk assessments for large numbers of substances. It was used to provide a case study with respect to the prioritisation of polybrominated diphenylethers, according to their impact on the environment (28). The air emission scenario was used, based on the SimpleBox implementation, while the SSD was estimated by using models developed to estimate environmental toxicity toward two fish species (rainbow trout and fathead minnow) and the water flea *D. magna*.

The results of this study suggest that the potential for environmental impact increases with the number of bromine atoms in the compound (28). However, it is also known that for PBDEs the "lower brominated mixtures are more toxic than are the higher congeners" (61). In fact, the majority of predictions for compounds with a large number of Br atoms, i.e. > 5–6, were out of the AD of the respective models for both toxicity and physicochemical properties predictions. Any inferences based on such predictions could be biased, and thus could lead to potentially wrong conclusions. This problem was not taken into consideration in our previous study. Therefore, we have implemented on the website an alert mechanism, which indicates the number of predictions that are out of the AD of models. The presence of such alerts will warn the users about the potential problems when using the web tool.

This outcome demonstrates the importance of considering the AD in order to make correct conclusions, as well as the usefulness and practicability of the developed approach of *in silico* risk assessment.

Conclusion

The QSPR-THESAURUS incorporates certain features that were particularly necessary for the successful execution of the CADASTER project and for the exemplification of the use of *in silico* models for

risk assessment. The compound properties database provides open and free access to the data collected and measured during the project, which are organised according to the four chemical classes analysed. It also provides online access to models developed by the project participants. The case study developed from PBDE exemplifies the use of *in silico* predictions for fate, hazard and risk assessment. The instructions on how to apply, upload, or develop new models by using the tools made available are provided on the website. The public availability of the models and the data will allow external users to easily access and make use of the outcomes of the project. Thus, it will contribute to its promotion (62) and a wider acceptance of computational methods in REACH and environmental sciences.

Acknowledgements

This study was partly supported by the EU through the CADASTER project (FP7-ENV-2007-212668) and the FP7 MC ITN project Environmental ChemOinformatics (ECO; Grant Agreement No. 238701), and the GO-Bio 1B BMBF project iPRIOR (Grant Agreement No. 315647). The authors thank Dr Pantelis Sopasakis for his contributions, and for programming some of the parts of the interface.

References

1. 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.
2. Peijnenburg, W. & Tetko, I.V. (2013). Exemplification of the implementation of alternatives to experimental testing in chemical risk assessment — case studies from within the CADASTER project. *ATLA* **41**, 13–17.
3. CADASTER (undated). *QSPR-THESAURUS Online Platform*. Available at: http://qspr-thesaurus.eu/ (Accessed 04.01.14).
4. Sushko, I., Novotarskyi, S., Körner, 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.
5. Aldenberg, T. & Rorije, E. (2013). Species sensitivity distribution estimation from uncertain (QSAR-based) effects data. *ATLA* **41**, 19–31.
6. Brandmaier, S., Sahlin, U., Tetko, I.V. & Öberg, T. (2012). PLS-Optimal: A stepwise D-optimal design based on latent variables. *Journal of Chemical Information & Modeling* **52**, 975–983.
7. Brandmaier, S., Tetko, I.V. & Öberg, T. (2012). An evaluation of experimental design in QSAR modelling utilizing the K-medoid clustering. *Journal of Chemometrics* **26**, 509–517.
8. Brandmaier, S. & Tetko, I.V. (2013). Robustness in experimental design: A study on the reliability of selection approaches. *Computational & Structural Biotechnology Journal* **7**.
9. JRC (2014). *QSAR Reporting Formats and JRC (Q)SAR Model Inventory*. Ispra, Italy: Institute for Health and Consumer Protection (JRC-IHCP). Available at: http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/QRF (Accessed 04.01.14).
10. CADASTER (undated). *Tutorial on Model Upload to QSPR-THESAURUS Web Using JRC Models*. Available at: http://www.cadaster.eu/node/118 (Accessed 05.01.14).
11. JRC (undated). *(Q)SAR Model Reporting Format Inventory*. Available at: http://qsardb.jrc.it/qmrf/ (Accessed 05.01.14).
12. Anon. (2012). *Model Uploader*. Available at: http://wiki.cadaster.eu/index.php/Model_uploader (Accessed 05.01.14).
13. Sushko, I., Novotarskyi, S., Körner, R., Pandey, A.K., Kovalishyn, V.V., Prokopenko, V.V. & Tetko, I.V. (2010). Applicability domain for *in silico* models to achieve accuracy of experimental measurements. *Journal of Chemometrics* **24**, 202–208.
14. Tetko, I.V., Sushko, I., Pandey, A.K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D. & Varnek, A. (2008). Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: Focusing on applicability domain and overfitting by variable selection. *Journal of Chemical Information & Modeling* **48**, 1733–1746.
15. Tetko, I.V., Bruneau, P., Mewes, H-W., Rohrer, D.C. & Poda, G.I. (2006). Can we estimate the accuracy of ADME–Tox predictions? *Drug Discovery Today* **11**, 700–707.
16. Roy, P.P., Kovarich, S. & Gramatica, P. (2011). QSAR model reproducibility and applicability: A case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-)triazoles. *Journal of Computational Chemistry* **32**, 2386–2396.
17. 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.

18. Gramatica, P., Cassani, S., Roy, P.P., Kovarich, S., Yap, C.W. & Papa, E. (2012). QSAR modeling is not 'push a button and find a correlation': A case study of toxicity of (benzo-)triazoles on algae. *Molecular Informatics* **31**, 817–835.
19. 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 05.01.14).
20. Anon. (2011). *StandaloneAccess*. Available at: http://wiki.ochem.eu/w/StandaloneAccess (Accessed 05.01.14).
21. de Aguiar, P.F., Bourguignon, B., Khots, M.S., Massart, D.L. & Phan-Than-Luu, R. (1995). D-optimal designs. *Chemometrics & Intelligent Laboratory Systems* **30**, 199–210.
22. Anon. (2012). *Welcome to OCHEM*. Available at: http://wiki.ochem.eu/w/Main_Page (Accessed 05.01.14).
23. Posthuma, L., Traas, T.P. & Suter, G.W.I. (2002). *Species Sensitivity Distributions in Ecotoxicology*, 616pp. Boca Raton, FL, USA: Lewis Publishers.
24. Aldenberg, T. & Slob, W. (1993). Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. *Ecotoxicology & Environmental Safety* **25**, 48–63.
25. Sahlin, U. (2013). Uncertainty in QSAR predictions. *ATLA* **41**, 111–125.
26. Sahlin, U., Golsteijn, L., Iqbal, M.S. & Peijnenburg, W. (2013). Arguments for considering QSAR uncertainty in hazard and risk assessments. *ATLA* **41**, 91–110.
27. Aldenberg, T. & Jaworska, J.S. (2000). Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicology & Environmental Safety* **46**, 1–18.
28. Tetko, I.V., Sopasakis, P., Kunwar, P., Brandmaier, S., Novotarskyi, S., Charochkina, L., Prokopenko, V. & Peijnenburg, W.J.G.M. (2013). Prioritisation of polybrominated diphenyl ethers (PBDEs) by using the QSPR-THESAURUS web tool. *ATLA* **41**, 127–135.
29. Cassani, S., Kovarich, S., Papa, E., Roy, P.P., Rahmberg, M., Nilsson, S., Sahlin, U., Jeliazkova, N., Kochev, N., Pukalov, O., Tetko, I., Brandmaier, S., Durjava, M.K., Kolar, B., Peijnenburg, W. & Gramatica, P. (2013). Evaluation of CADASTER QSAR models for the aquatic toxicity of (benzo)triazoles and prioritisation by consensus prediction. *ATLA* **41**, 49–64.
30. Cassani, S., Kovarich, S., Papa, E., Roy, P.P., van der Wal, L. & Gramatica, P. (2013). *Daphnia* and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies Quantitative Activity–Activity modelling. *Journal of Hazardous Materials* **258–259**, 50–60.
31. Den Hollander, H., Van Eijkeren, J. & Van de Meent, D. (2004). *SimpleBox 3.0: Multimedia Mass Balance Model for Evaluating the Fate of Chemicals in the Environment. Report No. 601200003*, 155pp. Bilthoven, The Netherlands: National Institute for Public Health and the Environment (RIVM).
32. Sarfraz Iqbal, M., Golsteijn, L., Öberg, T., Sahlin, U., Papa, E., Kovarich, S. & Huijbregts, M.A.J. (2013). Understanding Quantitative Structure–Property relationships uncertainty in environmental fate modeling. *Environmental Toxicology & Chemistry* **32**, 1069–1076.
33. Rorije, E., Aldenberg, T. & Peijnenburg, W. (2013). Read-across estimates of aquatic toxicity for selected fragrances. *ATLA* **41**, 77–90.
34. CADASTER (undated). *Deliverables*. Available at: http://ecoitn.eu/node/110 (Accessed 05.01.14).
35. Ding, G-H., Frömel, T., van den Brandhof, E-J., Baerselman, R. & Peijnenburg, W.J.G.M. (2012). Acute toxicity of poly- and perfluorinated compounds to two cladocerans, *Daphnia magna* and *Chydorus sphaericus*. *Environmental Toxicology & Chemistry* **31**, 605–610.
36. Ding, G., Wouterse, M., Baerselman, R. & Peijnenburg, W.J.G.M. (2012). Toxicity of polyfluorinated and perfluorinated compounds to lettuce (*Lactuca sativa*) and green algae (*Pseudokirchneriella subcapitata*). *Archives of Environmental Contamination & Toxicology* **62**, 49–55.
37. Ding, G. & Peijnenburg, W.J.G.M. (2013). Physicochemical properties and aquatic toxicity of poly- and perfluorinated compounds. *Critical Reviews in Environmental Science & Technology* **43**, 598–678.
38. Mansouri, K., Consonni, V., Durjava, M.K., Kolar, B., Öberg, T. & Todeschini, R. (2012). Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by QSAR modeling. *Chemosphere* **89**, 433–444.
39. Durjava, M.K., 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.
40 Mojca Durjava, M.K., Kolar, B. & Peijnenburg, W. (2012). *Overview of New Data Generated (Deliverable 2.5)*. Available at: http://www.cadaster.eu/sites/cadaster.eu/files/data/deliverable/public/Deliverable2.5.pdf (Accessed 14.03.14).
41. Bhhatarai, B. & Gramatica, P. (2010). Per- and polyfluoro toxicity (LC50 inhalation) study in rat and mouse using QSAR modeling. *Chemical Research in Toxicology* **23**, 528–539.
42. Bhhatarai, B., Teetz, W., Liu, T., Öberg, T., Jeliazkova, N., Kochev, N., Pukalov, O., Tetko, I.V., Kovarich, S., Papa, E. & Gramatica, P. (2011). CADASTER QSPR models for predictions of melting and boiling points of perfluorinated chemicals. *Molecular Informatics* **30**, 189–204.
43. Bhhatarai, B. & Gramatica, P. (2011). Prediction of aqueous solubility, vapor pressure and critical micelle concentration for aquatic partitioning of perfluorinated chemicals. *Environmental Science & Technology* **45**, 8120–8128.
44. Bhhatarai, B. & Gramatica, P. (2011). Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. *Water Research* **45**, 1463–1471.
45. Bhhatarai, B. & Gramatica, P. (2011). Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse. *Molecular Diversity* **15**, 467–476.
46. Papa, E., Kovarich, S. & Gramatica, P. (2009). Development, validation and inspection of the applicability domain of QSPR models for physico-chemical properties of polybrominated diphenyl ethers. *QSAR & Combinatorial Science* **28**, 790–796.
47. Papa, E., Kovarich, S. & Gramatica, P. (2011). On the use of local and global QSPRs for the prediction of physico-chemical properties of polybrominated

diphenyl ethers. *Molecular Informatics* **30**, 232–240.
48. Öberg, T. & Liu, T. (2011). Extension of a prediction model to estimate vapor pressures of perfluorinated compounds (PFCs). *Chemometrics & Intelligent Laboratory Systems* **107**, 59–64.
49. Papa, E., Kovarich, S. & Gramatica, P. (2010). QSAR modeling and prediction of the endocrine-disrupting potencies of brominated flame retardants. *Chemical Research in Toxicology* **23**, 946–954.
50. Kovarich, S., Papa, E. & Gramatica, P. (2011). QSAR classification models for the prediction of endocrine disrupting activity of brominated flame retardants. *Journal of Hazardous Materials* **190**, 106–112.
51. Kovarich, S., Papa, E., Li, J. & Gramatica, P. (2012). QSAR classification models for the screening of the endocrine-disrupting activity of perfluorinated compounds. *SAR & QSAR in Environmental Research* **23**, 207–220.
52. Papa, E., Luini, M. & Gramatica, P. (2009). Quantitative Structure–Activity Relationship modelling of oral acute toxicity and cytotoxic activity of fragrance materials in rodents. *SAR & QSAR in Environmental Research* **20**, 767–779.
53. Papa, E., Kovarich, S. & Gramatica, P. (2013). QSAR prediction of the competitive interaction of emerging halogenated pollutants with human transthyretin. *SAR & QSAR in Environmental Research* **24**, 333–349.
54. 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.
55. 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.
56. Golsteijn, L., Iqbal, M.S., Cassani, S., Hendriks, H.W.M., Kovarich, S., Papa, E., Rorije, E., Sahlin, U. & Huijbregts, M.A.J. (2014). Assessing predictive uncertainty in comparative toxicity potentials of triazoles. *Environmental Toxicology & Chemistry* **33**, 293–301.
57. CADASTER (undated). *MOPAC*. Available at: http://cadaster.eu/mopac/ (Accessed 06.01.14).
58. Anderson, D.P. (2004). BOINC: A System for Public-resource Computing and Storage. *Proceedings Fifth IEEE/ACM International Workshop on Grid Computing, Pittsburgh, PA, USA*, pp. 4–10.
59. Stewart, J.J.P. (1990). MOPAC: A semiempirical molecular orbital program. *Journal of Computer-Aided Molecular Design* **4**, 1–103.
60. Brandmaier, S., Novotarskyi, S., Sushko, Y. & Tetko, I.V. (2013). From descriptors to predicted properties: Experimental design by using applicability domain estimation. *ATLA* **41**, 33–47.
61. Birnbaum, L. & Staskal, D. (2004). Brominated flame retardants: Cause for concern? *Environmental Health Perspectives* **112**, 9–17.
62. Tetko, I. (2012). The perspectives of computational chemistry modeling. *Journal of Computer-Aided Molecular Design* **26**, 135–136.