Public (Q)SAR services, integrated modeling environments, and molecular repositories on the web: state of the art and perspectives for future development

Igor V. Tetko[a,b], Uko Maran[c], Alexander Tropsha[d]

[a] Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, b. 60w, D-85764 Neuherberg, Germany

[b] BigChem GmbH, Ingolstädter Landstraße 1, b. 60w, D-85764 Neuherberg, Germany

[c] Institute of Chemistry, University of Tartu, Ravila 14A, Tartu 50411, Estonia

[d] Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599, USA

* Address for correspondence
Dr. Igor V. Tetko,
Institute of Structural Biology, Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany
itetko@vcclab.org
Tel.: +49-89-3187-3575
Fax: +49-89-3187-3585

Keywords: model repositories, on-line modeling environments, web-based models, QSAR, QSPR, chemoinformatics

This is a pre-print of article "Tetko, I. V.; Maran, U.; Tropsha, A., Public (Q)SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development. *Mol. Inform.* **2017**, 36, DOI: 10.1002/minf.201600082"

Abstract

Thousands of (Quantitative) Structure-Activity Relationships (Q)SAR models have been described in peer-reviewed publications; however, this way of sharing seldom makes models available for the use by the research community outside of the developer's laboratory. Conversely, on-line models allow broad dissemination and application representing the most effective way of sharing the scientific knowledge. Approaches for sharing and providing on-line access to models ranges from web services created by individual users and laboratories to integrated modeling environments and model repositories. This emerging transition from the descriptive and informative, but "static", and for the most part, non-executable print format to interactive, transparent and functional delivery of "living" models is expected to have a transformative effect on modern experimental research in areas of scientific and regulatory use of (Q)SAR models.

1. Introduction

The development of the World Wide Web has significantly changed the way research results are delivered and used to fuel new research. The development of distributed and cloud technologies have made data collections (including "Big Data for chemicals such as Pubchem or ChEMBL), standalone applications and remote web-service based tools easily accessible and usable. This evolution of both chemical data and tool sharing technologies puts pressure on (Q)SAR researchers to make their models available, executable, and transparent for wider user communities.

In spite of significant growth of the QSAR modeling field in recent years,[1, 2] there have also been noticeable publications reporting misuses and irreproducibility of (Q)SAR modeling results.[3-5] Indeed, not well-reasoned decisions about the selection of parameters for the modeling algorithm or poor and insufficient pre-processing of biological data and chemical structures is likely to result in non-reproducible models. These issues have been widely recognized in the scientific community.[6-11] The need and vision for wider application of (Q)SAR models in regulatory decision support has led to the establishment of OECD principles of QSAR model validation,[12] which were developed to emphasize best practices for model documentation and promote the use of (Q)SAR models in fulfilling needs of REACH legislation. One of the OECD main principles requires the use of "an unambiguous algorithm" as the prerequisite for the regulatory model acceptance. While this principle is very important, its practical implementation can be non-trivial.

Even when there is a detailed scientific description of the model, factors like pre-processing of molecules, e.g., ionization or standardization of representation of chemical groups could affect the results of QSAR modeling. An example can be provided by recent studies of structure-toxicity relationships for nitro-aromatic compounds.[13] Each compound was manually inspected

in order to create a curated dataset. During this process, five different representations of nitro groups were identified (Figure 1). Obviously, the difference in one or two bonds may appear to be insignificant in the context of the entire compound, but in reality, those inconsistencies in the representation of the same functional group would lead to different descriptors of the same molecule and, in some cases, to poor QSAR modeling results. Indeed, it was shown[14] for two different datasets that when instances of all five different nitro group representations were distributed between training and test sets the external model predictive accuracy R^2_{ext} was less than 0.6 or close to zero for the two datasets vs. 0.9 or 0.5, respectively, when the standardized representation of the nitro group was used.

Additional factors such as the choice of 3D molecular conformation, as well as natural variability of modeling steps, e.g., due to selection of a seed number to initialize neural networks weights, could influence the modeling results. Moreover, even implementation of the same chemical descriptor with different programs can result in dramatic differences in model performance. Good and vivid example is provided by the prediction of octanol-water partition coefficient (logP). Indeed, two implementations of logP calculator, MLOGP, by ADMET Predictor (SimulationsPlus, Inc.) and Dragon (Kode s.r.l.) resulted in 0.73 log unit error differences, RMSE 1.07 *vs*. 1.8, respectively, for the Pfizer in house data collection of more than 95k compounds.[15] The MLOGP model for logP is fairly simple involving only 13 parameters in a linear equation and it may be considered "an unambiguous algorithm" according to the OECD rules; however, in this case it was not true, because the definition of the individual terms in the equation was ambiguous and thus contributed to the variability in the prediction results. Since there is no computational implementation of MLOGP by the original authors, one can never be sure that any of its application corresponds to the algorithm used in the original model.

Moreover, it introduces another level of uncertainty: prediction of any properties of the new compound with any model relying on MLOGP as a descriptor is only valid when MLOGP for this new compound is calculated with the same program (and even its particular version) that was used during the model development step. This example also makes clear that QSAR models require very thorough and exact documentation accompanied with full data used in the development in order to be reproducible and also for the independent model evaluation purposes.

The above discussion highlights the critical importance of providing explicit documentation and adhering to best practices for model development and validation when making QSAR models publicly available. Models can be distributed as standalone tools and/or as publicly accessible web implementations; both approaches allow for wide dissemination and independent evaluation of the models. Both ways are currently widely used by academic community. In this review, we discuss state-of-the-art in sharing "live" (Q)SAR models as web services. The implementations of (Q)SAR models on the web range from very simple web sites and services offering predictions with one or several models to the sophisticated integrated modeling environments, which allows model development, storage and prediction to smart model repositories that give access to full model data. Sometimes such web resources are managed by individual research groups, and sometimes they are products of collaborative research projects conducted by large consortia. Irrespective of the scale and origin of such publicly available tools, these efforts represent a substantial departure from traditional *descriptive* publication of models towards making models publicly accessible on-line, interactive, and usable.

2. Research group centric model collections

Collections of (Q)SAR models have been established by many research groups and laboratories working in the fields of computational chemistry and drug discovery. Even one model published and made publicly, or even commercially, available on the web improves accessibility. Often, while starting with a single model, such efforts result in model collections that incorporate new developments and frequently allow using models for the predictions. Such collections usually contain tools developed by one single group or/and a group of collaborators. Moreover, these web sites frequently contain not only tools per se, but also datasets, publications or pre-prints, etc. Several examples of such web sites and their main functionalities are discussed below.

ChemDB portal [http://cdb.ics.uci.edu]: Enables different predictions ranging from molecular properties such as water solubility (AquaSol module) to reaction outcomes and protein targets.[16] The site also provides other tools, such as chemical search, prediction of 3D conformations, analysis of functional groups and others.

ScreenDB [http://infochim.u-strasbg.fr/webserv/VSEngine.html]: This project was developed as a result of cooperative In Silico Design and Data Analysis (ISIDA)[17] effort including model development module, which incorporates multilinear regressions, k-nearest neighbors technique, neural networks, and support vector machines approaches to build structure-property models, and a knowledge base to store models. Presently the web portal enables prediction of twelve physico-chemical properties, AMES toxicity,[18] estrogen receptor models developed within CERAPP project,[19] as well as several target binding affinities.

PROTOX [http://tox.charite.de/tox**]**: This web site is designed for the prediction of rodent oral toxicities of small molecules.[20] The prediction is based on the similarity analysis of small

molecules against different toxicophores collected and constantly improved by the authors and recently tested on Tox21 prediction challenge.[21]

RS WebPredictor [http://reccr.chem.rpi.edu/software.html]. The web-service is an application of a predictor for cytochrome P450-mediated sites of metabolism. In addition to this tool, the Rensselaer Exploratory Center for Cheminformatics Research (RECCR) also provides a number of online tools to create (Q)SARs using support vector regression (SVR), partial least squares (PLS), Kernel-PLS and Support Vector Machines (SVM).

Way2Drug includes **PASS-Online, GUSAR-Online and BBB Predictor** [http://way2drug.com/total_plus]. PASS-Online, predicts more than 4000 different end-points. GUSAR-Online includes (Q)SAR models on acute rat toxicity (4 models), ecotoxicity (4 models), anti-target activity (32 models) and it allows consensus prediction capability. Other options are blood-brain-barrier predictor (BBB Predictor), *in silico* prediction of sites of metabolism (SOMP), as well as an access to several other tools. Some of the tools require registration and login (e.g., PASS on-line) while others are freely available on-line.

GUSAR@NIH [http://cactus.nci.nih.gov/chemical/apps/cap]: It is a web service, which includes a total of 25 models many of which are collaborative effort and are overlapping with those available at Way2Drug.

admetSAR [http://lmmd.ecust.edu.cn/admetsar1]: It is (Q)SAR based ADMET properties prediction web service including five regression and 26 classification models.[22] It also allows to search molecules by names, structure and/or perform a similarity search.

The aforementioned web sites indicate a wide and growing diversity of on-line models and tools on the web. Each of these web sites has an individual design and functionality, which is determined by the research activities of the respective groups and their collaborators. The

provided list is representative but by no means, exhaustive. Many more web sites are most likely available elsewhere.

---

3. Model collections from (Q)SAR oriented projects

Many research projects in the EU have been initiated with the goal of improving the quality of (Q)SAR models and their use in regulatory decision support. Those projects have relied on Web portals as very efficient dissemination channels and tools for sharing of models and results developed within the projects. Below, we provide an overview of several web sites, which collect models developed within completed or on-going research projects.

VCCLAB [http://www.vcclab.org]: Virtual Computational Chemistry LABoratory (VCCLAB) web site provides calculation of molecular descriptors, machine modeling tools as well as on-line access to several models for the calculation of logP (logarithm of octanol-water partition coefficient) and water solubility. The web site was launched in 2002 as a result of INTAS project.[23] Since that time about a hundred thousand users (counted by unique IPs) performed more than 2 million calculations. The site was developed using Java applets and in recent years this seriously limited its functionality due to the increasing security-related limitations of modern browsers.

QSPR-Thesaurus [http://www.qspr-thesaurus.eu]: The web site was developed within the FP7 CADASTER project [http://www.cadaster.eu].[24] It is based on a branch of OCHEM platform[25] and offers collection of data and models contributed by the project participants. The models are accompanied by the estimation of an accuracy of predictions and their applicability domain. Like the OCHEM web site, it allows upload of the user-supplied data and makes models available.

iPRIOR [http:/iprior.ochem.eu]: This site was also developed as a branch of OCHEM platform. The site was used[26] to develop tens of thousands of models for *in vivo* and *in vitro* toxicity data from the ToxCast project.[27]

COSMOS KNIME WebPortal [http://knimewebportal.cosmostox.eu]: COSMOS is an EU FP7 funded project [http://www.cosmostox.eu] aiming to integrate *in silico* models for the prediction of human repeated dose toxicity of cosmetics. This project is part of the overarching efforts in the EU to optimize safety of the cosmetic products without the use of animals. Currently, this project makes available six models for biokinetics, absorption and nuclear receptor binding, with the access to models provided to registered users; the registration is free.

eTOXsys [http://www.e-tox.net]: eTOX is an Innovative Medicines Initiative partnership between the European Community and the European Federation of Pharmaceutical Industries and Associations (EFPIA). This partnership has developed a drug safety database from the pharmaceutical industry legacy toxicology reports and public toxicology data to better predict the toxicological profiles of small molecules in early stages of the drug development. The partnership maintains the eTOXsys web-service that allows access to 74 models developed within eTOX[28] for different types of endpoints: physicochemical properties, ADME, transport (binding/inhibition), carcinogenicity, genotoxicity, organ toxicity, safety pharmacology. Models are developed within the eTOXlab model building environment making use of various built-in machine-learning methods such as Partial Least Squares Regression (PLS-R), Partial Least Squares Discriminant Analysis (PLS-DA), Fractional Factorial Design (FFD) variable selection, etc. Strong emphasis is also placed on the reporting of prediction results and working with confidential chemical data. At the time of writing the eTOXsys web-service was not available for the external evaluation.

Not always can research groups providing web services for accessible models secure continuing support and quite often support for online services ceases within a few years after the project ends. Two examples of such web sites include VEGA-QSAR and ToxPredict.

VEGA-QSAR [http://www.vega-qsar.eu]: This site agglomerates results of several EU projects (CALEIDOS, ORCHESTRA, ANTARES, CAESAR) and says to provide on-line access to models developed within those projects. At the time of writing of this review, the access to models on-line was not available. Project, however, provides a standalone application that is updated frequently.

ToxPredict [http://toxpredict.org]: This web site was developed within FP7 OpenTox project.[29] The web site provided access to several dozens of models generated by the collaborators of the project as well as by external providers, including those available on the QSPR-Thesaurus web site.[24] Unfortunately, this web site is no longer active.

4. (Q)SAR models in integrated modeling environments

As described in many reviews,[30-32] (Q)SAR models have been most commonly provided in the form of standalone software that could be open, or restricted, or commercial. Good examples are EPI Suite™ from the US EPA,[33] VEGA-QSAR,[34] ChemProp [http://www.ufz.de/index.php?en=6738], ToxTree,[35] QSARINS-Chem,[36] the OECD QSAR Toolbox [http://www.qsartoolbox.org]. During the past decade, with the advent of the web, as well as the distributed and cloud computing technologies serious attempts have been made to move model development and subsequent access to the models into two-in-one solutions, i.e., integrated modeling environments on the web. The Internet based tool for model development using Polynomial Neural Networks (PNN)[37] first made available in 2000 from the web site of the Neuroheuristic Laboratory of the University of Lausanne was probably the first documented implementation of distributed QSAR modeling efforts in chemistry (Figure 2). In 2001 it was complemented with Java Applet to predict solubility and lipophilicity.[38] These developments were extended with new tools to develop models,[39, 40] calculate descriptors[41] and analyze data,[42, 43] some of which have been publicly available at the VCCLAB[23] since 23 March, 2002. VCCLAB provided access to tools developed and running in the laboratories of six partners from five countries in Europe. OpenMolGRID[44, 45] was started in 2002; it adapted distributed computing and grid technologies for (Q)SAR model development and deployment in the field of drug design, in particularly in predicting cytotoxicity and other ADMET endpoints, with the emphasis on sharing and reproducibility of models. In 2006, Chemomentum,[46] the successor of OpenMolGRID, integrated 24 different tools for predictive modeling as distributed computing solutions with focus on (Q)SAR and predictive toxicology towards effective use of models in REACH for decision support as one of three application lines. Since 2008, Chembench[47] and

OCHEM[25] have provided public (for registered users; registration is free) access to several QSAR modeling tools and growing collection of bioactivity and ADME/Tox models. Below the active integrated modeling environments on the web are described.

CHEMBENCH [http://chembench.mml.unc.edu]: The Chembench is one of the first publicly accessible, for registered users, integrated portal that was started in 2008 and described in an application note in 2010.[47] Chembench is designed to integrate translational cheminformatics research conducted both in the Tropsha group over a period of more than twenty years as well as in collaborating laboratories elsewhere. The current ChemBench system consists of four modules: Dataset, Modeling, Prediction, and My Bench. The functionality contained under Dataset allows users to upload, store, and standardize chemical structures. My Bench enables the analysis and visualization of chemical structures, to examine the distribution of activities, and generate a heat map to check for obvious relationships between global compound similarity and activity. Through Modeling module, users can select a dataset (either an uploaded dataset or a provided benchmark) and build a (Q)SAR model. Several methods for model development (e.g., support vector machines, k-nearest neighbors, and random forest) are available for the use. The Prediction module allows users to predict new compounds' activities using one or more of the models built on Chembench by the user or by Chembench developers; compound structures could be drawn within the Chembench environment or uploaded in the SDF format. Chembench implements best practices of QSAR modeling and validation[11] and models developed within Chembench are fully compliant with the OECD guidelines on QSAR model validation.[12] Chembench has been extensively used for the development of QSAR models, as well as a teaching tool.

OCHEM [http://ochem.eu]: OCHEM, which became publicly available in 2008, provides a full spectrum of model development tools, allows upload of previously published linear models or storage and archiving of results of other modeling experiments.[25] This platform was used to develop models for datasets incorporating hundreds of thousands of molecules.[48-50] OCHEM currently stores 105 published models developed and contributed by different teams across the world. Some are consensus models and include up to seventeen individual sub-models, as is the case for the predictor of the compound solubility in DMSO.[48] All individual submodels (N=390) can be also accessed on the web site (Figure 3). The model development methods include inspired by thalamo-cortical organization of brain[51] associative neural networks,[52] and their library mode approach,[53] support vector machines,[54] partial least squares, linear regression analysis, fast stage wise multivariate linear regression,[55] k nearest neighbors as well as several Weka algorithms.[56] OCHEM also includes other useful tasks, such as calculation of molecular descriptors, virtual screening of compounds collections as well as comparison of sets of molecules[57] using SMARTS patterns, namely toxicological alerts,[58] functional groups,[59] frequent hitters[60, 61] and others. The predicted Matched Molecular Pairs[62] allow interpreting of models as well as identification of experimental errors. The OCHEM models report confidence intervals and applicability domain[63] for new predictions and thus help user to interpret the calculated results. The web site supports the development of models with several different properties (also combinations of regression and classification properties) simultaneously.[64] Neural network models support the development of models with intervals and ranges. OCHEM can use externally uploaded user-provided descriptors including measured data (e.g., *in vitro* measurements), support conditions (e.g., temperature can be specified as a condition for boiling point), automatic unit conversion, support mixtures[65] as well as use predictions by models as

descriptors for developing new models.[64] The users can also upload results of previously developed models, namely calculated values and descriptors. If the uploaded descriptors are available, the uploaded linear models can be applied to new molecules. OCHEM provides 25 descriptor blocks, including several commercial software packages such as Dragon,[41] Adriana,[66] ChemAxon and Mera/Mersy.[67]

Models available on OCHEM were frequently top-performing within different benchmarking exercises, e.g., prediction of metal complexation,[68] environmental toxicity,[69] readily biodegradability,[57] endocrine disruptors,[19] logP,[15, 70, 71] AMES toxicity,[72] CYP450 inhibition,[73] etc., as well as contributed the top-rank model for the EPA ToxCast[74] and the best overall balanced accuracy for twelve targets of the NIH Tox21[75] challenges. The OCHEM is widely used for educational purposes by a number of Universities across the world. It was used as a part of educational process in the FP7 Marie Curie Initial Training Network "Environmental ChemOinformatics" (ECO) [http://www.eco-itn.eu][76] and will be also used in Horizon2020 Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorates, "Big Data in Chemistry" ("BIGCHEM") [http://bigchem.eu].[77] It has more than 3000 registered users and it has performed more than 32 million calculations. Being backed-up by more than 1500 CPU-cores, OCHEM can easily handle models requiring thousands of CPU-hours by using parallel calculations with up to 48 cores. OCHEM efficiently supports sparse data format, which made it possible to develop models with 300k molecules and more than 700k individual descriptors (i.e., with full matrix of >200,000,000,000 entries).[50] All these features make OCHEM a powerful framework for (Q)SAR analysis of the emerging Big Data in chemistry and chemogenomics.[77]

CHARMMing [http://www.charmming.org]: This web site was previously developed as a Web-based front-end to the CHARMM molecular simulation package.[78] Since 2015 it has been extended with tools to develop (Q)SAR models using several machine learning algorithms for both regression and classification tasks.[79] The web site (that requires free registration) uses 2048-bit Morgan fingerprints calculated using RDkit [http://www.rdkit.org/docs]. The users can upload data as SD file, select the target property and submit calculations. There is no option to specify or optimize parameters of machine learning methods. Unexpectedly, the authors of this manuscript failed to develop a model using uploaded data. Thus, while this web site could become an interesting tool in the future, it still requires significant efforts to achieve this status.

5. (Q)SAR model repositories

(Q)SAR model web collections described in previous sections focused on models developed and made available by the same group, collaborators within a project, or in integrated modeling environments. They rely on internal protocols and standards, employed by different contributors. Model repositories provide solutions where modeling is left to the developers such that the repositories are focusing on systematic storage and delivery of models and also on predictions when architecture allows this. The (Q)SAR model repositories require the development of standards for organizing both models and underling data. There have been several efforts in organizing (Q)SAR model information where emphasis and strategies have been very different. For example QSAR Model Reporting Format (QMRF) [http://eurl-ecvam.jrc.ec.europa.eu/databases/jrc-qsar-model-database-and-qsar-model-reporting-formats] originates directly from the aforementioned OECD principles[12] and from the need to allow model developer to suggest their models for the legislative use in the framework of REACH. QMRF organizes model information in the text-rich form in a single XML file with accompanying SDF files for the chemical structures. QSAR-ML[6] also uses single XML file focusing on QSAR data sets and their exchange, without the mathematical definition of the model, which is left for the software used for model development. MIAQMR-ML collects minimum information about a QSAR model representation also into the XML format according to the six rules of model provenance, description, data reference, development reliability and predictivity.[80] These rules are derived from the OECD QSAR model validation principles[12] and MIAQMR-ML is used by Syngenta Ltd. in frame of Model and Data Farm (MADFARM) that also has web interface.[81] QSAR DataBank or QsarDB archive format[82] uses the 'collection of files' approach that systematically names and groups files containing information about a

QSAR model: molecular and experimental data, mathematical representation of models and full provenance information of model and its components. QsarDB archive format is fully machine-readable, keeps data and model information as the original developer designed it and allows easy reuse of the model(s). Detailed review about formats for (Q)SAR model organization can be found elsewhere.[83] Models delivered in QMRF and QsarDB archive format are organized in repositories, which are briefly reviewed below.

QMRF Inventory [http://qsardb.jrc.it]: This site was developed in response for a need to organize QSAR models documented in QMRF format for assessment of various endpoints within REACH. The site uploads reviewed and accepted QMRF-s and accompanying SDF files. The authors can either provide links to their models (in case when the model is available somewhere, e.g., as public or free software tools) and reference to articles, which provide more detailed scientific descriptions. In practice, the textual information in QMRF and original publication overlaps, and is extended in some cases. The web site itself does not store models in directly usable form but instead it provides information, which could be sufficient either to reproduce models or to locate them. The web-based interface allows users to search for QMRF documents via QMRF number, free text, pre-defined lists of endpoints, algorithms, software, structures and authors of the models. Inventory is accompanied with the QMRF Editor that supports the development of QMRF XML files. The web site allows chemical authorities and companies who need to predict various endpoints for the registration of their compounds in REACH to locate models and their providers. At the time of writing, the QMRF Inventory included 109 (Q)SAR models.

QsarDB Repository [http://qsardb.org]: The QsarDB repository[84] was developed for efficient electronic organization of (Q)SAR models, that could be easily accessible, reproducible

and also discoverable by other users and in the form that allows independent validation. QsarDB repository is essentially a large collection of QDB archives[82] that include information about model development, original data used for the model development and full mathematical representation of models. Providing mathematical representation in machine readable form is essential, because (Q)SAR models are *de-facto* mathematical models Mathematical models used in (Q)SAR transform input (chemical) data according to some rules, while coefficients and weights provide the resulting values. The implementation of these transformations can be in the form of programming language or be developer specific, but their results should be the same. The problem of re-use of the developed models has been a target for the Data Mining Group [http://dmg.org] that developed a standard for representing mathematical models in XML format, the Predictive Model Markup Language (PMML). If models are expressed in this format, they can be understood and executed by other users in exactly the same way. This feature is used by QsarDB for the mathematical representation of (Q)SAR models. It was made to work with our current knowledge as to how to handle the experimental chemical structure data by converting the structure to molecular descriptors. This synergy of the mathematical model representation and the original data, i.e., primary information, facilitates reproducibility and provides a possibility for the independent evaluation. QsarDB always calculates all statistical characteristics, i.e., secondary information, on the fly. QsarDB repository improves the conventional (Q)SAR model publishing *via* assigning unique resource identifiers, i.e. DOI-s, for predictive models, allowing both easy access to the models and citing of models. The data and mathematical representation of the model in QDB archive is organized together with the provenance information, the references to the literature or the original data source(-s), what modeling workflow was used, and how descriptors were calculated. QsarDB repository along

with data collects (Q)SAR specific metadata from the well-organized and correctly populated QDB archive (or asks metadata from model provider) that helps to locate (Q)SAR models and leaves longer meta-information about model development and its scientific interpretation for the scientific publication or for any other published media, where original authors of the models can discuss in detail all constraints and assumptions during the model development. QsarDB open repository allows easy establishment of supplementary material for conventional PDF publications and makes these publications interactive. Linking with scientific publications can be established through DOI links via manually inserting them into the publication text and/or automatically, as in the case of Taylor & Francis, who provides automatic links to external repositories as supplementary information (currently available for 29 QDB archives).

In addition to basic information about stored models (Figure 4), the QsarDB repository has a smart component that allows detailed visualization and analysis of the model content *via* the QDB Explorer module and the use of models for prediction, through the QDB Predictor module. All models in QsarDB are readily executable and can be used for the prediction purposes directly from the repository, by using (i) built-in slide-bar feature for descriptors, (ii) inserting in-house calculated molecular descriptor values via web-form or (iii) directly from structure for cases when descriptor calculator is implemented in QsarDB (currently available for models with CDK[85] descriptors). The latter option is also available as a web-service. Both QDB Explorer and Predictor allow visualization of applicability domain and simple indication if prediction is within the domain of the model. QDB Predictor module allows saving prediction results into specialized reporting formats required by the regulatory authorities. QsarDB allows full text search of QDB archives and also structure/sub-structure search of compounds in the models. At the time of writing, the QsarDB repository included 410 unique (Q)SAR models. QsarDB

provides access to a variety of model types supported by the PMML. Presently, the following regression and classification model types have been tested and application examples provided: (multi-) linear regression (MLR), decision tree (DT), neural network (NN), random forest (RF), support vector machine (SVM), k-nearest neighbors (k-NN), and ensemble (consensus) model. At the time of writing this text, repository included models for 70 different endpoints, including 38 specifically listed and required by the REACH. The design of QsarDB archive format and the repository follow the OECD principles[12] for predictive model documentation and it allows easy access to both data and mathematical representation of the model thereby providing full transparency for any users, including evaluators or regulators.

6. Summary and Outlook

More than fifty years of active studies in the (Q)SAR field has resulted in great number of studies and large number of models.[84] Many new chemicals have been discovered and developed for therapeutic or industrial use with the active use of (Q)SAR models, and these significant achievements of the QSAR modeling field cannot be overlooked.[86] However, until recently, published models could not be really accessed regardless of how detailed their description was in respective scientific papers or monographs. The traditional textual publishing format has restricted if not prevented the practical application of many QSAR models by users outside of the developer's lab or company. Simple cause for this is that conventional PDF/HTML articles in vast amount of the cases do not describe models with enough detail to be reproduced and reused. The quality documentation is also not required by the publishers. With simple multi-linear regression and small datasets it was relatively easy to provide description sufficient to understand and reproduce the model. However, with the growing transition to the use of machine learning algorithms and big datasets model development has become much more serious undertaking that requires specialized and standardized data formats and software solutions relying on web technologies for proper model and data representation.

(Q)SAR as research filed and more importantly its methodological framework has matured tremendously during past decades.[87, 88] Various artificial intelligence and machine learning approaches have been successfully utilized to improve the coding of (Q)SARs. This has been guided by the explosion of experimental information becoming available for small molecules and increasing chemical complexity needed to be analyzed and modeled as the approaches to how the molecular structure can be characterized and what molecular descriptors can be calculated for numerical characterization of the structure have become very diverse as well. Many different

software can be used for the development of (Q)SARs.[30, 31, 77] For example, as the analysis of the QsarDB repository[84] indicates, thirty six programs are commonly used for the calculation of molecular descriptors and twenty programs are implemented for model development. These developments indicate that (Q)SAR methodological framework is very diverse and requires the practitioner to have larger spectra of knowledge in multiple fields. Diversity of tools and expert knowledge required to build and validate models makes (Q)SAR model development a sophisticated scientific art. At the same time one must admit that the times when (Q)SAR model developer and user were the same person is over for the most part; today the model user community is much larger than that of developers, and therefore users require the ease of access and ability to routinely employ models as part of their experimental studies.

As we discussed herein, there is a highly significant emerging trend in making models work for the users. This trend is enabled by the developments in web technologies and the willingness of model developers to make the products of their intellectual labor available to wider community. This is also pushed by the funding agencies' data sharing requirements. However, the ease of use that comes with web based model distribution puts additional pressure on the developers in that the models must be developed using rigorous best practices, validation protocols and must be well documented. This feature must be in place to ensure that users would truly benefit from, rather than misled by, the available models so the model sharing has the transformative rather than detrimental effect on the users' ability to discover or regulate new compounds. Further developments should focus on how to make (Q)SAR models even more transparent and easily understandable. Considering the above-described complexity of (Q)SAR methodological framework this will be the true challenge for sharing, distribution and using of models for many application scenarios.

The availability of (Q)SAR models on the web has been rapidly growing and is becoming an accepted and popular way for knowledge dissemination. Many individual laboratories as well as projects use the web as an important tool to spread information about their research. In this way the developed models can be delivered to the end-users as they are created. However there are drawbacks. The support of such individual developments requires a lot of efforts, which and can be handled by the relatively large and established groups of specialists. The same is true and even more pertinent for the web sites developed within the projects. While the research group web sites are frequently supported thanks to the research funding, the support for the web sites developed within the project is frequently terminated after the end of the project. Without continuity such web sites lose their functionality with time due to problems with bug fixes and updates of the operation system, change of the software version, and licensing issues. However even in these cases it is important to keep the developed models alive and preserved to extend life of such models and to accommodate models and data for further use. Considering the fairly large and growing number of (Q)SAR publications,[1, 84] efforts on model sharing on the web discussed above compile only a tiny fraction of knowledge available in the scientific literature in the form of descriptive, predictive, and potentially useful but unavailable models. In the future, the requirements by the publisher[89] to make models available not only in print but also on the web can be one of the possible measures to support model sharing. This can be mandatory or voluntary. Web technologies already provide support for it and example of such realizations have been described above.

Diversity of (Q)SAR methodological frameworks offers a lot of creativity to model developers. As a rule, the selection of tools and approaches depends on the problem at hand. Creativity, however, must consider best practices of model development and model

documentation. Despite of the fifty plus years of active methodological development and dissemination via traditional publications, very little attention has been given to preserving comprehensive information concerning (Q)SAR models. Previously, with fairly small datasets and mostly multi-linear models this has not been an issue. However, with the experimental data explosion in chemistry, the need for much more effective and impactful (Q)SAR data organization and dissemination is timely and one of the essential development areas for coming years. We most likely do see new developments in this area with (i) individual models as focused services, (ii) integrated modeling web environments where model development and storage solution are integrated and (iii) model repositories, in providing access to the vast amount of models.

As already discussed, model documentation is important for the preserving and reuse of the published models. From this point, a scientific article must contain sufficient information to re-build, i.e., reproduce the published model and also include information what constraints one must consider when applying the model. Therefore, to ensure the reproducibility of the model described in the publications, better documentation of data and model information is needed, where emerging advanced web technologies can contribute substantially in future. In deploying the model, the next level of reproducibility is required that is related to the workflow. In this case, built-in workflow components (like descriptor calculation together with geometry optimization) can be used within prediction services. During the model development process many approximations are often introduced and those approximations are too easily forgotten or not documented, which can contribute to difficulties with reproducing and reusing the models. The standardized and well documented model development protocols could contribute to solving this problem. This is particularly important for the regulatory decision support scenarios where

once used prediction must be backtracked and therefore models must be preserved for longer time.

An important aspect of the reuse of models and to a certain extent also of the model reproducibility is the availability of molecular descriptor calculation software. Whereas algorithms for molecular descriptor calculations are usually available in the scientific literature, their software realizations are mainly commercial. As a result, only small portion of models is developed with open molecular descriptor calculators, which contribute another difficulty with the reuse of published model. In the future, massive use (and to certain extent also reuse) of (Q)SAR-s in web applications could truly benefit from the open, well organized, and standardized descriptor calculators.

Importantly, the aforementioned problems do not exist for models developed with integrated modeling environments, such as Chembench or OCHEM, and repositories, like QsarDB. Models published within integrated modeling environments are applied using exactly the same protocols as those the models were developed with therefore allowing the exact reproducibility of the results. Models and the associated data used for model development published within repositories are well documented and detail of model representation allows recalculation and reproduction of models independent of the original modeling setup and environment.

Due to the explosion of the experimental information, the data pre-treatment for modeling and model deployment is gaining importance and most likely we will see new web solutions in the future. Data pre-treatment includes at least three components, data and structure curation, structure standardization and data filtering for making dataset fit for the modeling purposes.[10] Integrated web modeling environments described herein already include rigorous protocols for data curation and standardization,[11, 14, 90] where they are applied both to chemicals used for

model development and also in deployment in prediction services. Despite of the available software solutions for data pre-treatment, all steps in these protocols must be well documented, because of reproducibility issues and because they expose constraints to the use of model and very often also to the models domain of applicability.

Publishing of QSAR models on the web is changing the way we disseminate and use the published results.[91] Indeed, instead of looking at the model statistics we can now interactively explore models, apply them to new compounds and use in our next study within minutes. There is no doubt in our minds that public availability of models should become a requirement for publishing of (Q)SAR results in the future, which will promote a better use of computational chemistry and chemoinformatics methods in chemistry and Life Sciences in general.

Acknowledgements

U.M. is grateful for financial support from the Estonian Ministry of Education and Research (grant IUT34-14). A.T. acknowledges the support from the NIH grant GM 096967 and expresses his sincere gratitude to Prof. Diane Pozefsky for leading the technical design of the Chembench project.

Conflict of interests

I.V.T. is CEO and founder of BigChem GmbH, which licenses the OCHEM.[25] The other authors declared that they have no actual or potential conflicts of interests.

References

[1] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, Baskin, II, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, *J. Med. Chem.* **2014**, *57*, 4977-5010.

[2] I. Baskin, D. Winkler, I. V. Tetko, *Expert Opin. Drug Discov.* **2016**, *11*, 785-795.

[3] J. C. Dearden, M. T. Cronin, K. L. Kaiser, *SAR QSAR Environ. Res.* **2009**, *20*, 241-266.

[4] T. R. Stouch, J. R. Kenyon, S. R. Johnson, X. Q. Chen, A. Doweyko, Y. Li, *J. Comput. Aided. Mol. Des.* **2003**, *17*, 83-92.

[5] S. R. Johnson, *J. Chem. Inf. Model.* **2008**, *48*, 25-26.
[6] O. Spjuth, E. L. Willighagen, R. Guha, M. Eklund, J. E. Wikberg, *J. Cheminform.* **2010**, *2*, 5.
[7] P. Gramatica, *QSAR Comb. Sci.* **2007**, *26*, 694-701.
[8] E. Benfenati, M. Clook, S. Fryday, A. Hart, in *Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes*, Elsevier, Amsterdam, **2007**, pp. 1-57.
[9] I. V. Tetko, D. J. Livingstone, A. I. Luik, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826-833.
[10] V. Ruusmann, U. Maran, *J. Comput. Aided. Mol. Des.* **2013**, *27*, 583-603.
[11] A. Tropsha, *Mol. Inf.* **2010**, *29*, 476-488.
[12] Guidance document on the validation of (quantitative)structure-activity relationships [(Q)SAR] models. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2
[13] A. G. Artemenko, E. N. Muratov, V. E. Kuz'min, N. N. Muratov, E. V. Varlamova, A. V. Kuz'mina, L. G. Gorb, A. Golius, F. C. Hill, J. Leszczynski, A. Tropsha, *SAR QSAR Environ. Res.* **2011**, *22*, 575-601.
[14] D. Fourches, E. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **2010**, *50*, 1189-1204.
[15] R. Mannhold, G. I. Poda, C. Ostermann, I. V. Tetko, *J. Pharm. Sci.* **2009**, *98*, 861-893.
[16] C. A. Azencott, A. Ksikes, S. J. Swamidass, J. H. Chen, L. Ralaivola, P. Baldi, *J. Chem. Inf. Model.* **2007**, *47*, 965-974.
[17] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. V. Tetko, G. Marcou, *Curr. Comp. Aid. Drug Design* **2008**, *4*, 191-198.
[18] I. Sushko, S. Novotarskyi, R. Korner, A. K. Pandey, A. Cherkasov, J. Li, P. Gramatica, K. Hansen, T. Schroeter, K. R. Muller, L. Xi, H. Liu, X. Yao, T. Oberg, F. Hormozdiari, P. Dao, C. Sahinalp, R. Todeschini, P. Polishchuk, A. Artemenko, V. Kuz'min, T. M. Martin, D. M. Young, D. Fourches, E. Muratov, A. Tropsha, I. Baskin, D. Horvath, G. Marcou, C. Muller, A. Varnek, V. V. Prokopenko, I. V. Tetko, *J. Chem. Inf. Model.* **2010**, *50*, 2094-2111.
[19] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A. M. Richard, C. M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E. B. Wedebye, F. Grisoni, G. F. Mangiatordi, G. M. Incisivo, H. Hong, H. W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N. G. Nikolov, O. Nicolotti, P. L. Andersson, Q. Zang, R. Politi, R. D. Beger, R. Todeschini, R. Huang, S. Farag, S. A. Rosenberg, S. Slavov, X. Hu, R. S. Judson, *Environ. Health Perspect.* **2016**, *124*, 1023-1033.
[20] M. N. Drwal, P. Banerjee, M. Dunkel, M. R. Wettig, R. Preissner, *Nucleic Acids Res.* **2014**, *42*, W53-58.
[21] M. N. Drwal, V. B. Siramshetty, P. Banerjee, A. Goede, R. Preissner, M. Dunkel, *Frontiers Environ. Sci.* **2015**, *3*.
[22] F. Cheng, W. Li, Y. Zhou, J. Shen, Z. Wu, G. Liu, P. W. Lee, Y. Tang, *J. Chem. Inf. Model.* **2012**, *52*, 3099-3105.
[23] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, V. V. Prokopenko, *J. Comput. Aided. Mol. Des.* **2005**, *19*, 453-463.
[24] S. Brandmaier, W. Peijnenburg, M. K. Durjava, B. Kolar, P. Gramatica, E. Papa, B. Bhhatarai, S. Kovarich, S. Cassani, P. P. Roy, M. Rahmberg, T. Oberg, N. Jeliazkova, L. Golsteijn, M. Comber, L. Charochkina, S. Novotarskyi, I. Sushko, A. Abdelaziz, E. D'Onofrio, P. Kunwar, F. Ruggiu, I. V. Tetko, *Altern. Lab. Anim.* **2014**, *42*, 13-24.
[25] I. Sushko, S. Novotarskyi, R. Korner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. I. Baskin, V. A. Palyulin, E. V. Radchenko, W. J.

Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q. Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I. V. Tetko, *J. Comput. Aided. Mol. Des.* **2011**, *25*, 533-554.

[26] A. Abdelaziz, Y. Sushko, S. Novotarskyi, R. Korner, S. Brandmaier, I. V. Tetko, *Comb. Chem. High Throughput Screen.* **2015**, *18*, 420-438.

[27] R. S. Judson, K. A. Houck, R. J. Kavlock, T. B. Knudsen, M. T. Martin, H. M. Mortensen, D. M. Reif, D. M. Rotroff, I. Shah, A. M. Richard, D. J. Dix, *Environ. Health Perspect.* **2010**, *118*, 485-492.

[28] F. Sanz, P. Carrió, O. López, L. Capoferri, D. P. Kooi, N. P. E. Vermeulen, D. P. Geerke, F. Montanari, G. F. Ecker, C. H. Schwab, T. Kleinöder, T. Magdziarz, M. Pastor, *Mol. Inf.* **2015**, *34*, 477-484.

[29] B. Hardy, N. Douglas, C. Helma, M. Rautenberg, N. Jeliazkova, V. Jeliazkov, I. Nikolova, R. Benigni, O. Tcheremenskaia, S. Kramer, T. Girschick, F. Buchwald, J. Wicker, A. Karwath, M. Gutlein, A. Maunz, H. Sarimveis, G. Melagraki, A. Afantitis, P. Sopasakis, D. Gallagher, V. Poroikov, D. Filimonov, A. Zakharov, A. Lagunin, T. Gloriozova, S. Novikov, N. Skvortsova, D. Druzhilovsky, S. Chawla, I. Ghosh, S. Ray, H. Patel, S. Escher, *J. Cheminform.* **2010**, *2*, 7.

[30] A. A. Toropov, A. P. Toropova, I. Raska, Jr., D. Leszczynska, J. Leszczynski, *Comput Biol Med* **2014**, *45*, 20-25.

[31] O. Nicolotti, E. Benfenati, A. Carotti, D. Gadaleta, A. Gissi, G. F. Mangiatordi, E. Novellino, *Drug Discov. Today* **2014**, *19*, 1757-1768.

[32] A. A. Lagunin, R. K. Goel, D. Y. Gawande, P. Pahwa, T. A. Gloriozova, A. V. Dmitriev, S. M. Ivanov, A. V. Rudik, V. I. Konova, P. V. Pogodin, D. S. Druzhilovsky, V. V. Poroikov, *Nat. Prod. Rep.* **2014**, *31*, 1585-1611.

[33] U.S. Environmental Protection Agency EPI Suite v 4.11. http://www.epa.gov/opptintr/exposure/pubs/episuitedl.htm

[34] E. Benfenati, A. Manganaro, G. Gini, in *CEUR Workshop Proceedings, Vol. 1107*, **2013**, pp. 21-28.

[35] G. Patlewicz, N. Jeliazkova, R. J. Safford, A. P. Worth, B. Aleksiev, *SAR QSAR Environ. Res.* **2008**, *19*, 495-524.

[36] P. Gramatica, S. Cassani, N. Chirico, *J. Comput. Chem.* **2014**, *35*, 1036-1044.

[37] I. V. Tetko, T. I. Aksenova, V. V. Volkovich, T. N. Kasheva, D. V. Filipov, W. J. Welsh, D. J. Livingstone, A. E. P. Villa, *SAR QSAR Environ. Res.* **2000**, *11*, 263-280.

[38] I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva, A. E. Villa, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 246-252.

[39] T. I. Aksyonova, V. V. Volkovich, I. V. Tetko, *Syst. Anal. Model. Sim.* **2003**, *43*, 1331-1339.

[40] I. V. Tetko, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717-728.

[41] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, WILEY-VCH, Weinheim, **2000**.

[42] I. V. Tetko, A. Facius, A. Ruepp, H. W. Mewes, *BMC Bioinformatics* **2005**, *6*, 82.

[43] D. C. Whitley, M. G. Ford, D. J. Livingstone, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160-1168.

[44] S. Sild, U. Maran, A. Lomaka, M. Karelson, *J. Chem. Inf. Model.* **2006**, *46*, 953-959.

[45] S. Sild, U. Maran, M. Romberg, B. Schuller, E. Benfenati, in *Lecture Notes in Computer Science, Vol. 3470*, **2005**, pp. 464-473.

[46] B. Schuller, B. Demuth, H. Mix, K. Rasch, M. Romberg, S. Sild, U. Maran, P. Bała, E. Del Grosso, M. Casalegno, N. Piclin, M. Pintore, W. Sudholt, K. K. Baldridge, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 4854 LNCS*, **2008**, pp. 82-93.

[47] T. Walker, C. M. Grulke, D. Pozefsky, A. Tropsha, *Bioinformatics* **2010**, *26*, 3000-3001.

[48] I. V. Tetko, S. Novotarskyi, I. Sushko, V. Ivanov, A. E. Petrenko, R. Dieden, F. Lebon, B. Mathieu, *J. Chem. Inf. Model.* **2013**, *53*, 1990-2000.

[49] I. V. Tetko, Y. Sushko, S. Novotarskyi, L. Patiny, I. Kondratov, A. E. Petrenko, L. Charochkina, A. M. Asiri, *J. Chem. Inf. Model.* **2014**, *54*, 3320-3329.

[50] I. V. Tetko, D. Lowe, A. J. Williams, *J. Cheminform.* **2016**, *8*, 2.
[51] A. E. Villa, I. V. Tetko, P. Dutoit, Y. De Ribaupierre, F. De Ribaupierre, *J. Neurosci. Methods* **1999**, *86*, 161-178.
[52] I. V. Tetko, *Methods Mol. Biol.* **2008**, *458*, 185-202.
[53] I. V. Tetko, P. Bruneau, *J. Pharm. Sci.* **2004**, *93*, 3103-3110.
[54] C. Cortes, V. Vapnik, *Machine Learn.* **1995**, *20*, 273-297.
[55] N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, N. S. Zefirov, *Dokl. Chem.* **2007**, *417*, 282-284.
[56] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *SIGKDD Explorations* **2009**, *11*.
[57] S. Vorberg, I. V. Tetko, *Mol. Inf.* **2014**, *33*, 73-85.
[58] I. Sushko, E. Salmina, V. A. Potemkin, G. Poda, I. V. Tetko, *J. Chem. Inf. Model.* **2012**, *52*, 2310-2316.
[59] E. S. Salmina, N. Haider, I. V. Tetko, *Molecules* **2016**, *21*, 1.
[60] K. Schorpp, I. Rothenaigner, E. Salmina, J. Reinshagen, T. Low, J. K. Brenke, J. Gopalakrishnan, I. V. Tetko, S. Gul, K. Hadian, *J. Biomol. Screen.* **2014**, *19*, 715-726.
[61] J. K. Brenke, E. S. Salmina, L. Ringelstetter, S. Dornauer, M. Kuzikov, I. Rothenaigner, K. Schorpp, F. Giehler, J. Gopalakrishnan, A. Kieser, S. Gul, I. V. Tetko, K. Hadian, *J. Biomol. Screen.* **2016**, *21*, 596-607.
[62] Y. Sushko, S. Novotarskyi, R. Korner, J. Vogt, A. Abdelaziz, I. V. Tetko, *J. Cheminform.* **2014**, *6*, 48.
[63] I. V. Tetko, P. Bruneau, H. W. Mewes, D. C. Rohrer, G. I. Poda, *Drug Discov. Today* **2006**, *11*, 700-707.
[64] A. Varnek, C. Gaudin, G. Marcou, I. Baskin, A. K. Pandey, I. V. Tetko, *J. Chem. Inf. Model.* **2009**, *49*, 133-144.
[65] I. Oprisiu, S. Novotarskyi, I. V. Tetko, *J. Cheminform.* **2013**, *5*, 4.
[66] J. Gasteiger, *J. Med. Chem.* **2006**, *49*, 6429-6434.
[67] V. Potemkin, M. Grishina, *Drug Discov. Today* **2008**, *13*, 952-959.
[68] I. V. Tetko, V. P. Solov'ev, A. V. Antonov, X. Yao, J. P. Doucet, B. Fan, F. Hoonakker, D. Fourches, P. Jost, N. Lachiche, A. Varnek, *J. Chem. Inf. Model.* **2006**, *46*, 808-819.
[69] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, *J. Chem. Inf. Model.* **2008**, *48*, 1733-1746.
[70] I. V. Tetko, H. P. Varbanov, M. Galanski, M. Talmaciu, J. A. Platts, M. Ravera, E. Gabano, *J. Inorg. Biochem.* **2016**, *156*, 1-13.
[71] I. V. Tetko, I. Jaroszewicz, J. A. Platts, J. Kuduk-Jaworska, *J. Inorg. Biochem.* **2008**, *102*, 1424-1437.
[72] I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, V. V. Kovalishyn, V. V. Prokopenko, I. V. Tetko, *J. Chemom.* **2010**, *24*, 202-208.
[73] S. Novotarskyi, I. Sushko, R. Korner, A. K. Pandey, I. V. Tetko, *J. Chem. Inf. Model.* **2011**, *51*, 1271-1280.
[74] S. Novotarskyi, A. Abdelaziz, Y. Sushko, R. Korner, J. Vogt, I. V. Tetko, *Chem. Res. Toxicol.* **2016**, *29*, 768-775.
[75] A. Abdelaziz, H. Spahn-Langguth, K. Werner-Schramm, I. V. Tetko, *Frontiers Environ. Sci.* **2016**, *4*, 2.
[76] I. V. Tetko, K. W. Schramm, T. Knepper, W. J. Peijnenburg, A. J. Hendriks, J. M. Navas, I. A. Nicholls, T. Oberg, R. Todeschini, E. Schlosser, S. Brandmaier, *Altern. Lab. Anim.* **2014**, *42*, 7-11.
[77] I. V. Tetko, O. Engkvist, U. Koch, J. L. Reymond, H. Chen, *Mol. Inf.* **2016**.
[78] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, *J. Comp. Chem.* **1983**, *4*, 187-217.

[79] I. E. Weidlich, Y. Pevzner, B. T. Miller, I. V. Filippov, H. L. Woodcock, B. R. Brooks, *J. Comput. Chem.* **2015**, *36*, 62-67.
[80] A. Palczewska, X. Fu, P. Trundle, L. Yang, D. Neagu, M. Ridley, K. Travis, *Int. J. Inform. Manag.* **2013**, *33*, 567-582.
[81] X. Fu, A. Wojak, D. Neagu, M. Ridley, K. Travis, *J. Cheminform.* **2011**, *3*, 24.
[82] V. Ruusmann, S. Sild, U. Maran, *J. Cheminform.* **2014**, *6*, 25.
[83] S. Sild, G. Piir, D. Neagu, U. Maran, in *Big Data in Predictive Toxicology Vol. in press* (Eds.: D. Neagu, A. Richards), Royal Society of Chemistry, **2016**.
[84] V. Ruusmann, S. Sild, U. Maran, *J. Cheminform.* **2015**, *7*, 32.
[85] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493-500.
[86] W. M. Berhanu, G. G. Pillai, A. A. Oliferenko, A. R. Katritzky, *ChemPlusChem* **2012**, *77*, 507-517.
[87] T. Fujita, D. A. Winkler, *J. Chem. Inf. Model.* **2016**, *56*, 269-274.
[88] J. Dearden, *Int. J. Quant. Struct. Prop. Relat.* **2016**, *1*, 1-14.
[89] G. Melagraki, A. Afantitis, *Comb. Chem. High Throughput Screen.* **2016**, *19*, 260-261.
[90] D. Fourches, E. Muratov, A. Tropsha, *Nat. Chem. Biol.* **2015**, *11*, 535.
[91] I. V. Tetko, *J. Comput. Aided. Mol. Des.* **2012**, *26*, 135-136.

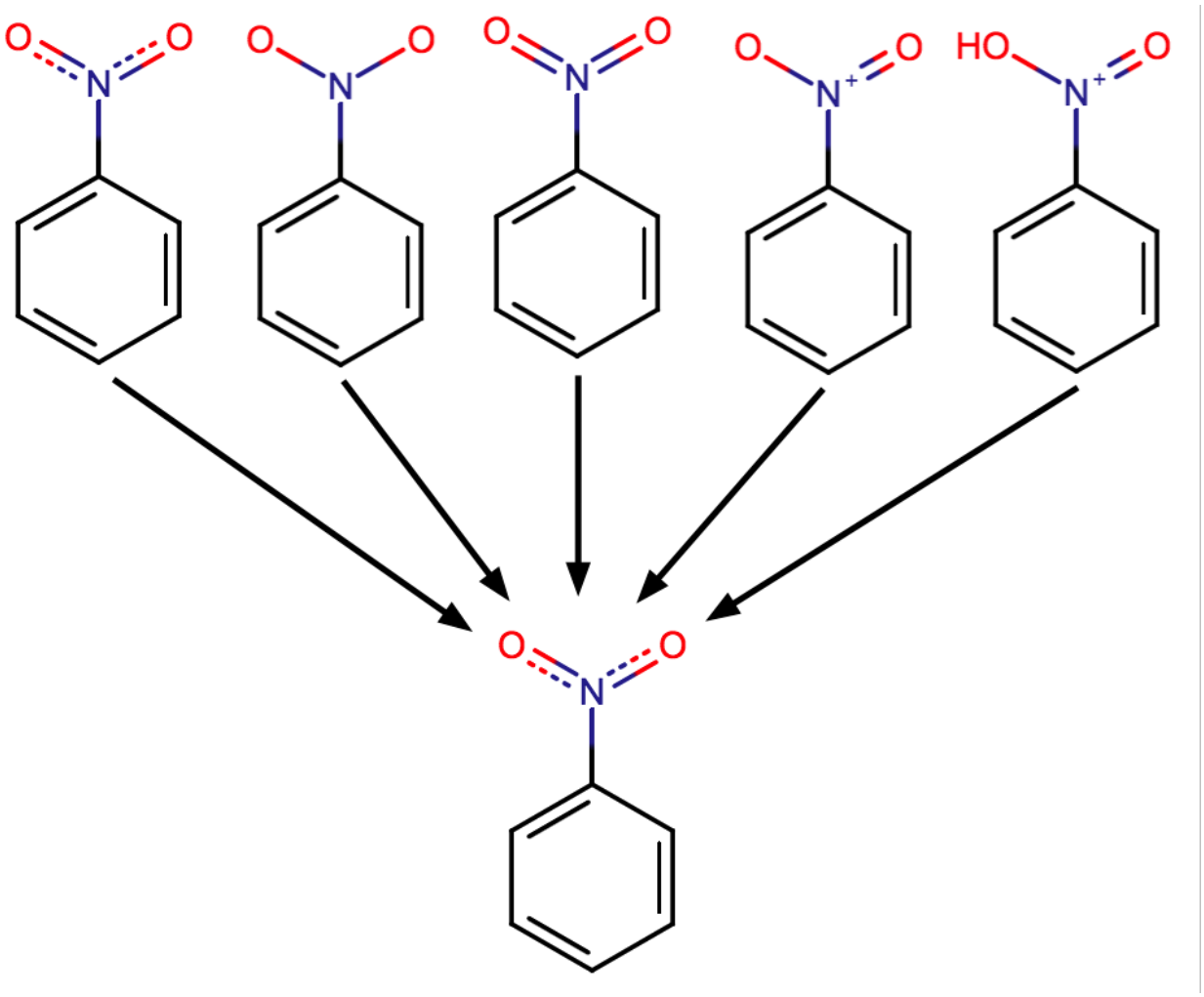

Figure 1. Structure normalization: five types of nitro group representations retrieved in the nitroaromatics toxicity dataset.[13]

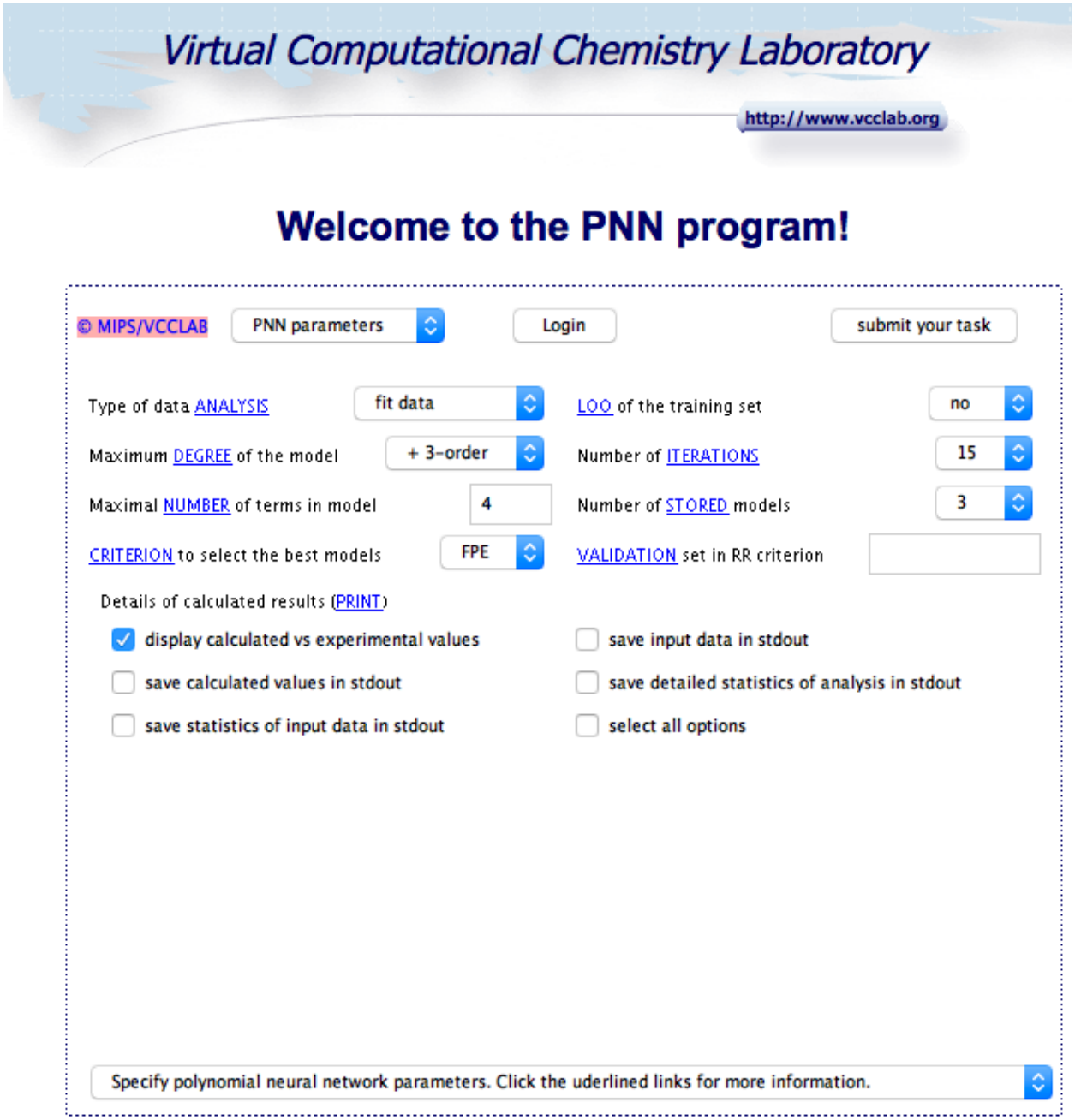

Figure 2. The applet interface to introduce parameters for Polynomial Neural Networks (PNN) used in the first published distributed QSAR calculation application on the web[37] as currently available on the VCCLAB (http://www.vcclab.org). The interface was slightly updated in 2003 following publication of an improved version of the PNN.[39]

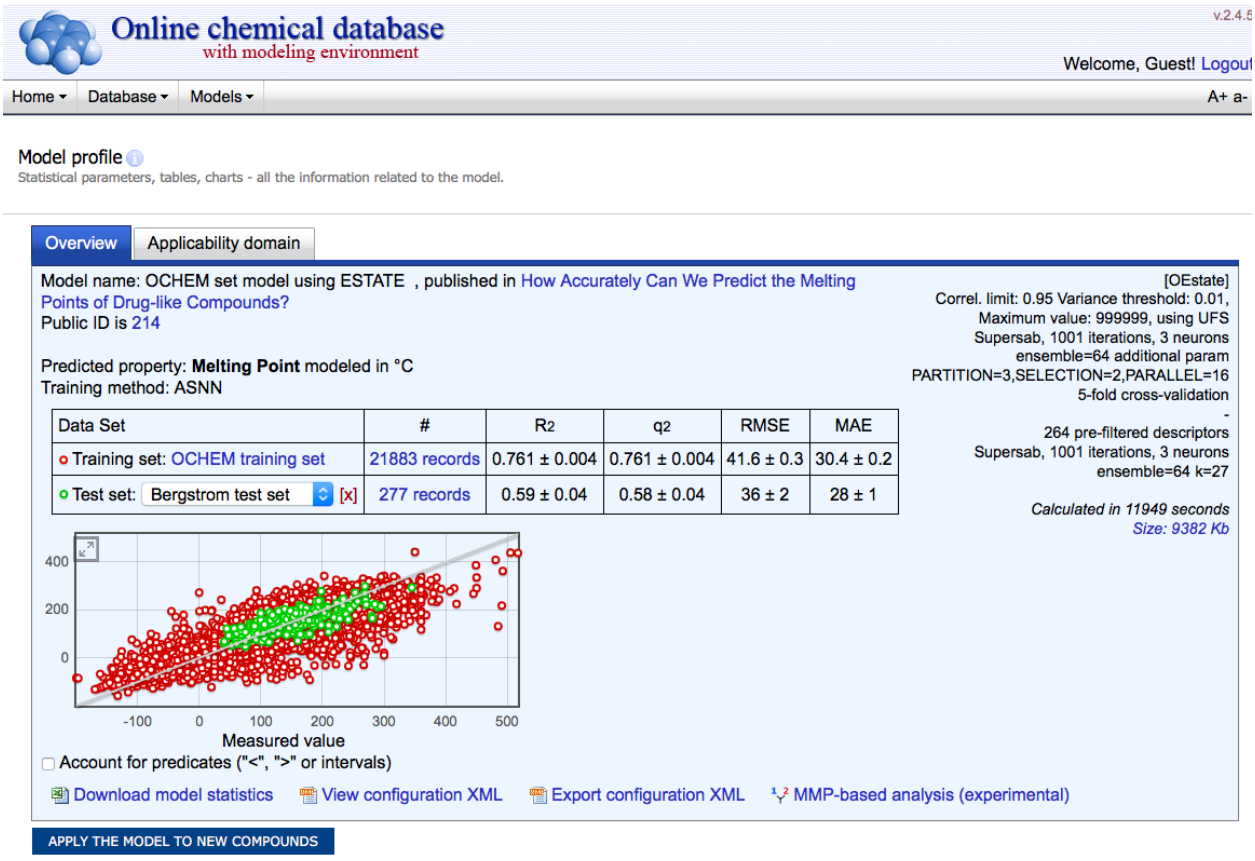

Figure 3. Profile of a Melting Point http://ochem.eu/model/214 at the OCHEM web site which is part of the larger consensus model from ref [49]. The calculated statistical parameters allow estimation of a quality of the model. A short summary of the model parameters is provided in the right corner. The "Export configuration XML" link downloads configuration, which can be uploaded to exactly reproduce all modeling steps and redevelop the model. Download model statistics link exports data and predicted values. The model can be applied to predict new compounds with one click.

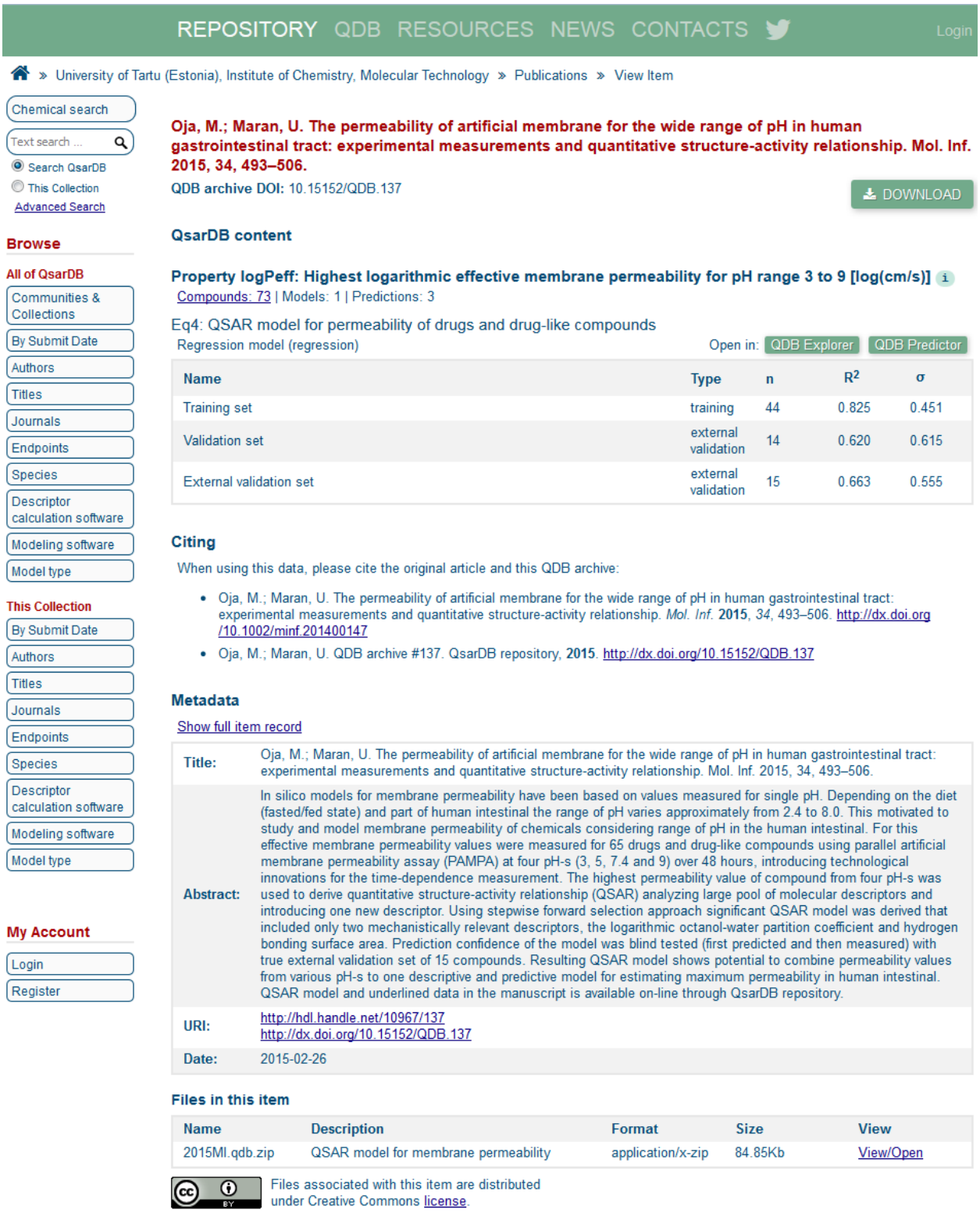

Figure 4. QDB archive Item View that shows an example of the information delivered to user together with links to 'QDB Explorer' and 'QDB Predictor' modules that allow further exploring of model content and to download the original archive file.

Biographical Sketches

Igor V. Tetko is head of Chemoinformatics group at the Helmholtz Zentrum München, Germany. He is also CEO of BigChem GmbH as well as coordinator of Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorates, "Big Data in Chemistry". He received his PhD from the Ukrainian Academy of Sciences in 1994 and habilitation in chemoinformatics from the University of Strasbourg (France) in 2011. His research interests include predictive modeling in cheminformatics, big data analysis and development of on-line web tools.

Uko Maran is Senior Researcher at the Institute of chemistry, University of Tartu, Estonia. He received his doctoral degree (PhD) in theoretical and computational chemistry from University of Tartu in 1997. Prior to joining University of Tartu research staff he has been studying and working at the University of Joensuu (now University of Eastern Finland), the University of Uppsala (Sweden) and the University of Florida (USA). His main areas of research are descriptive and predictive modelling, data analysis and computer-aided molecular design for chemical, toxicological, biochemical and biotechnological, etc., processes.

Alexander Tropsha is K.H. Lee Distinguished Professor and Associate Dean for Pharmacoinformatics and Data Science at the UNC Eshelman School of Pharmacy, UNC-Chapel Hill. Prof. Tropsha obtained PhD in Chemical Enzymology in 1986 from Moscow State University, Russia. His research interests are in the areas of Computer-Assisted Drug Design, Cheminformatics, Structural Bioinformatics and Computational Toxicology. His has authored or co-authored more than 200 peer-reviewed research papers, reviews and book chapters and co-edited two monographs. His research has been supported by multiple grants from the NIH, NSF, EPA, DOD, and private companies.