EXTERNAL SCIENTIFIC REPORT

APPROVED: 31 March 2017 AMENDED 14 July 2017

doi:10.2903/sp.efsa.2017.EN-1206

Developing innovative *in silico* models with EFSA's OpenFoodTox database

Emilio Benfenati, Francesca Como, Marco Marzo, Domenico Gadaleta,

Andrey Toropov and Alla Toropova

Instituto di Ricerche Farmacologiche Mario Negri

Abstract

The present scientific report summarises the development of innovative *in silico* quantitative structure activity relationship (QSAR) models as tools to predict toxicity values or classify thresholds for human and environmental risk assessment (HRA and ERA). These QSAR models have been developed using EFSA's Chemical hazards database: openfoodtox, other relevant databases (e.g. US-EPA terrestrial database, Fraunhofer RepDose) and the open source VEGA platform. Two continuous QSAR models relevant to HRA were developed using data for sub-chronic toxicity in rats and a global model was developed to predict toxicity values in OpenFoodTox using the VEGA platform. For ERA, two QSAR models were developed for predicting acute toxicity in rainbow trout as a continuous model and to predict acute contact toxicity data in bees as a classification model.

These innovative QSAR models open novel avenues for chemical risk assessment and will be published in the near future in the VEGA platform as open source tools. Further, these approaches demonstrate the usefulness of large open source toxicological databases, providing historical data to boost bioinformatics analysis and *in silico* modelling particularly for compounds with scarce toxicological data. Future research is proposed particularly to develop systematic and harmonised approaches for the use of QSAR and read across for a number of endpoints and species relevant to HRA, ERA as well as for the refinement of the Threshold of Toxicological Concern (TTC) approach.

Key words: In silico models; QSAR, toxicity, risk assessment, NOAEL, LD_{50}, rat, bee, trout, VEGA.

Question number: EFSA-Q-2015-00170

Correspondence: sc.secretariat@efsa.europa.eu

Acknowledgements: EFSA wishes to thank the following for the support provided to this scientific output: Emilio Benfenati and Marco Marzo from the Mario Negri Institute, Arianna Bassan, Lidia Ceriani and Manuela Pavan from Soluzione Informatiche and EFSA staff for their support: Jean Lou Dorne and Nikolaos Georgiadis.

Amendment: An editorial correction for one of the author's names was carried out and it does not materially affect the contents or outcome of this scientific output. To avoid confusion the older version has been removed from the EFSA Journal, but is available on request, as is the version showing all the changes made.

Suggested citation: Benfenati E, Como F, Marzo M, Gadaleta D, Toropov A and Toropova A, 2017. Developing innovative in silico models with EFSA's OpenFoodTox database. EFSA supporting publication 2017:EN-1206. 19 pp. doi:10.2903/sp.efsa.2017.EN-1206

ISSN: 2397-8325

Table of contents

1. Introduction

1.1. Background as provided by EFSA

Modern methodologies for toxicological testing and chemical risk assessment are currently a topic of great interest amongst researchers and the regulatory community, because of their potential for predicting chemical toxicity and reducing animal testing. In 2014, EFSA Scientific Committee and Emerging Risks Unit published a scientific report on "modern methodologies and tools for human hazard assessment of chemicals" and identified this topic a priority for EFSA in the technical report on EFSA Horizon 2020 priority topics (EFSA, 2014). The report provided a review of a number of methodologies and tools and their application to investigate toxicokinetic (TK) and toxicodynamic (TD) processes of chemicals, i.e. mode of Action (MoA)/Adverse Outcome pathway (AOP) at different levels of biological organization (organism, organ, cellular and molecular level). These included in vitro systems, physiologically-based (PB) models (such as PB-TK and PB-TK-TD models), in silico and decision making tools ((Quantitative) Structure Activity Relationship (Q)SAR) systems, Threshold of Toxicological Concern (TTC) and read across methods) and OMICs technologies (transcriptomics, proteomics, and metabolomics). Recommendations for the potential applications of these modern methodologies and tools were also proposed including the development of TK models and in silico tools such as QSAR and read across models as tools for EFSA's future work. A working example of their applicability is the current development of a guidance document by EFSA's Scientific Committee to integrate results from these modern tools into weight of evidence approaches for chemical risk assessment.

Finally, these modern methods have also been identified as a high priority for the EFSA Strategy 2020 implementation plan and the Scientific Committee particularly regarding objective 4 "Prepare for future risk assessment challenges" 4.2 "support for the development and use of harmonised methodologies for risk assessment across the EU and internationally" and 4.3 "Become a hub in methodologies and tools for risk assessment".

1.2. *In Silico* tools and models: (Q)SAR models and read-across

In the broad sense of the term, *in silico* models refer to computer tools and models available to scientists to simulate biological processes for a range of applications, species and level of biological organisation (cellular, molecular, species, population, ecosystem, landscape etc). In the area of toxicology and risk assessment, *in silico* tools often aim to predict toxicity of chemicals and cover a wide range of methodologies that would also comprise molecular modelling approaches and general computational toxicology tools, including theoretical models based on the intrinsic structural and physicochemical properties of chemicals and rule-based expert systems.

Structure-Activity Relationships (SAR) and **Quantitative Structure Activity Relationships (QSAR)** models are often collectively referred to as (Q)SAR as mathematical models that relate the structure of chemicals to their biological activities. The term 'quantitative' refers to the fact that the molecular descriptors are quantifiable on a continuous scale and thus provide a quantitative relationship with toxicity (which may itself be expressed in quantitative or categorical terms). Molecular descriptors of chemicals include their inherent physicochemical properties (i.e. atomic composition, structure, sub-structures, hydrophobicity, surface area charge, and molecular volume). QSARs can be classified in relation to their dimensionality 1.1D-QSAR as a system for which the effect is correlated with a single (e.g. physicochemical) property. 2. 2D-QSAR associated with atomic connectivity or two-dimensional (e.g. pharmacophore) patterns. 3.3D-QSAR associated with three-dimensional structure of a compound. 4. Multi-dimensional QSAR (dimensions $n > 3$) or short 'mQSAR' include multiple representation of ligands such as 4D-QSAR Dimensionalities (Vedani et al., 2000; Tseng et al., 2012) and the protein 5D/6D (Vedani et al., 2006).

SAR and QSAR models can provide a fast method for the toxicity screening of untested substances, for identifying emerging chemicals in the food chain that have not yet been tested for their safety to human health or the environment. They are typically used in combination with other non-testing (e.g. read-across) and testing (e.g. *in vitro*) methods in the context of integrated testing strategies (ITS) and Weight-of-Evidence assessments (EFSA, 2014).

Read-across represents 'a technique for predicting endpoint information for one substance (target substance), by using data from the same endpoint from (an)other substance(s), (source substance(s))' as defined by the European Chemical Agency (ECHA) (ECHA, 2008). ECHA used two key approaches for read across: 1. *Analogue approach* for which read-across is applied within a group of a very limited number of substances e.g. simplest read-across from a single source substance to a target substance 2.*Category approach* for which compounds can be grouped in the case of a high number of substances and comprehensive guidance on grouping and read-across has been published by the OECD (OECD, 2007) and ECHA (ECHA, 2008). Examples of criteria to group chemicals include physico-chemical properties, functional/mechanistic/structural alert groups, chemical similarity, e.g. based on the Tanimoto coefficient and similarity in breakdown or metabolic products (Schilter et al, 2014; EFSA, 2014).

The development of QSAR and read-across approaches for predicting toxicity of chemicals ideally involves quantitative understanding and data relating both toxicokinetics and toxicodynamic processes and some of the underlying parameters as predictor variables take into account TK (e.g. partitioning coefficients) or TD (e.g. electronic properties) in the QSARs modelling. Key databases for QSAR and read-across have been described elsewhere (EFSA, 2014) and include:

-The OECD QSAR Toolbox (http://www.qsartoolbox.org/) as a hazard identification tool which contains QSAR relationship methodologies to group chemicals into categories sharing the same structural characteristics and/or MoA.

-The eChemPortal hosted by the OECD allows simultaneous searching of reports/datasets by chemical name and number and by chemical property providing direct links to collections of use, exposure, hazard and risk information from government chemical review programs at national, regional and international levels.

-The VEGA platform (www.vega-qsar.eu) includes a large number of models including physico-chemical properties, ecotoxicological and toxicological properties. VEGA provides a quantitative measurement of the applicability domain of the target chemical specific for each endpoint, based on a set of parameters considering the chemical and/or toxicological parameters, and those related to the algorithm. VEGA platform also includes tools for read across (ToxRead), prioritisation (JANUS) and results are are integrated within the ToxWeight tool using weight-of-Evidence approches.

-Other databases include Chembase, ChemIDplus, ChemSpider, Pubchem, Carcinogenic Potency Database, DSSTox, European chemical Substances Information System, NTP Database: (, IPCS ToxNet.

Key QSAR software and models used by international and national organizations including the Toxicity Estimation Software Tool (TEST), the OECD QSAR toolbox models and High-throughput Virtual Molecular Docking (HTVMD), MetaCore, DEMETRA, CAESAR, DEREK, METEOR, Multicase, PASS, OASIS Times (EFSA,. 2014). Recently, the ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) SAR database, abbreviated admetSAR, has been published as open source, text and structure searchable, and is continually updated (Cheng et al., 2013). AdmetSAR provides ADMET-associated properties data from the published literature with over 210 000 ADMET annotated data points for over 96 000 compounds with 45 kinds of ADMET-associated properties, proteins, species, or organisms and allows queries for specific chemical profiles using the CAS registry number, the common name, or structure similarity. ADMET-SAR includes 22 qualitative classification and 5

quantitative regression models allowing the estimation of ecological/mammalian ADMET properties for "new" or emerging chemicals (Cheng et al., 2013).

In this context, EFSA has consulted its scientific panels and formulated a number of recommendations regarding the future of *in silico* tools (EFSA, 2014):

1. Further explore the application of *in silico* tools in chemical risk assessment using of publicly) available databases containing large sets of physicochemical and toxicological data and validate available predictive models to reduce animal use. In addition, it will provide the opportunity to explore the applicability domain of the predictive methods and their degree of specificity. It can be foreseen that the domain of applicability of such tools will be encompassing human health (HRA), animal health (ARA), and ecological risk assessment (ERA).

2. Further develop a framework for systematic and harmonised approach use of *in silico* tools (SAR, QSAR, read-across) as potential tools to: 1. support the hazard identification of genotoxic compounds by building batteries of models based on structural alerts, toxicity data and existing databases, 2. design physiologically-based models, 3. elucidate the mode of action (including toxicity pathways) for the prioritisation of chemicals. A key aspect of these applications is the need to compare the currently available (Q)SAR tools in a transparent way to allow optimisation and calibration of the models.

3. Further develop read-across methodologies in hazard assessment of chemicals, particularly to integrate (Q)SAR and physicochemical properties with TK and TD data (potency estimates, AOP etc) using specific chemicals as case studies for 'proof of concept'. This can also be useful to explore category-approaches for prioritisation of chemicals, especially for data-poor substances (e.g. flavourings, emerging contaminants…) using for example the OECD QSAR toolbox or the ADMET-SAR tool.

1.3. EFSA's chemical Hazards Database: OpenFoodTox

Since its creation in 2002, EFSA has been performing risk assessment for chemicals in food and feed which has defined as 'a scientifically based process consisting of four steps: hazard identification, hazard characterisation, exposure assessment and risk characterisation' (EC, 2002; WHO, 2009). Overall, more than 4400 substances have been assessed in over 1650 opinions and have been structured into EFSA's chemical hazards database: OpenFoodTox. OpenFoodTox is an open source database which contains for each individual substance: the substance characterisation, the links to EFSA's related output(s), the background European legislation, and a summary of the critical toxicological endpoints and reference values (HRA, ARA and ERA depending on the relevant legislation and intended uses)(Dorne et al., 2017).

For regulated compounds, individual risk assessments have been performed by five scientific panels and four supporting units:

1. HRA for a) food additives and nutrient sources added to food (ANS panel and FIP unit); b) food contact materials, enzymes, flavourings and processing aids (CEF panel and FIP unit), c) vitamins, minerals and novel foods Dietetic products, nutrition and allergies (NDA panel and unit).

2) HRA, ARA and ERA for pesticides performing the peer review of plant protection product opinions (PPPs) and publishing conclusions on single pesticides (pesticide unit) and for feed additives (Additives, products and substances used in animal feed (FEEDAP panel and unit).

For contaminants, the Scientific Panel and Unit on contaminants in the food chain (CONTAM panel and unit) have been dealing with HRA and ARA for a) contaminants of anthropogenic origin (e.g. brominated flame retardants, dioxins), environmental contaminants (e.g. heavy metals), b) compounds resulting from food and/or feed processing (e.g. acrylamide), c) natural toxins produced as undesirable substances in food and feed by plants, fungi and other micro-organisms (e.g. alkaloids,

mycotoxins, marine biotoxins). For compounds falling under the remit of more than one panel (e.g. carvone), the risk assessment has been performed by the Scientific Committee of EFSA[1] (EFSA Scientific Committee, 2014a).

With reference to the use of toxicological data for hazard identification and characterisation in the food safety, reference values are derived to set safe levels of exposure for substances with regards to HRA, ARA and ERA from pivotal toxicology studies that provide the basis for a reference point. The reference points are then divided by uncertainty factors to derive reference values. Examples of reference points for human health and animal health effects include Lowest or No-Observed–Adverse-Effect-Level (LOAEL/NOAEL), Benchmark dose limit (BMDL, e.g. $BMDL_{10}$), LD_{50} or No Observed Effect Concentration (NOECs) for eco-toxicological effects (daphnia, fish, bees etc). Examples of reference values include health-based guidance values for setting safe levels of chronic exposure in humans such as acceptable daily intake (ADI) for food and feed additives, pesticides and food contact materials, Upper Limits (UL) for vitamins and minerals and tolerable daily intake (TDI) for contaminants.

OpenFoodTox has been designed and developed using a data model taking into account OECD Harmonised templates[2] to collect and structure the data in a harmonised manner. Detailed description of the data and the structure of OpenFoodTox have been published elsewhere (S-IN, 2013, 2014, 2015).

For the user, OpenFoodTox provides options to search data for each substance using chemical descriptors and generate a summary datasheet (pdf/Excel) using an online micro-strategy tool:

- Chemical Characterisation (e.g. name, formula, CAS and EU number, IUPAC, smile etc…)
- EFSA outputs (Scientific opinions, Statement or Conclusions) and background regulations where the corresponding bibliographic details, Digital Object Identifier and link are provided.
- Critical toxicological study including study design (length of study, species, type), reference point, or toxicity value for HRA, ARA or ERA.
- Conclusions on the mutagenicity and the genotoxicity/of the substance
- Reference Values and uncertainty factors applied for the derivation of health based guidance values for humans (e.g. ADI, TDI) and environmental standards (e.g. NOECs or predicted NOECs).

Openfoodtox is available in EFSA's data warehouse under:

[https://dwh.efsa.europa.eu/bi/asp/Main.aspx?rwtrep=400].

Overall, OpenFoodTox contributes actively to objective 2 of EFSA's 2020 Strategy[3] which aims at "widening EFSA's evidence base and optimise access to its data" as a valuable open source database that can be shared with all scientific advisory bodies and stakeholders with an interest in chemical risk

[1] The Scientific Committee develops harmonized risk assessment methodologies on scientific matters of a horizontal nature in the fields within EFSA's remit where EU-wide approaches are not already defined. It provides general co-ordination to ensure consistency in the scientific opinions prepared by EFSA's scientific panels. It also provides strategic scientific advice to EFSA's management.

[2] http://www.oecd.org/ehs/templates/

[3] http://www.efsa.europa.eu/en/corporate/pub/strategy2020

assessment. In addition, OpenFoodTox has been submitted to the OECD's Global Portal to Information on Chemical Substances (eChem Portal) so that individual substances can be searched as part of national and international databases. By making this summary data available in a readily accessible format, it is anticipated that this further analysis of the data by the wider scientific community will be stimulated and generate new knowledge and tools in the area of chemical risk assessment.

In the context of this report, OpenFoodTox has been explored as a data source to support the development of QSAR models for HRA and ERA.

1.4. Objectives

The further development of *in silico* models such as QSAR tools and models for HA and ERA of chemicals has been identified as a key priority for EFSA (EFSA, 2014), sister agencies (ECHA and EMA) as well as for the general scientific community, national and international scientific advisory bodies (OECD, US-EPA, WHO etc..). These predictive models allow the use of historical toxicological data, provide a means to reduce animal testing. In the food safety area, these models can be particularly useful tools for weight of evidence approaches for compounds for which toxicity data are not available. In this context, the integration of such *in silico* tools including QSAR models in chemical risk assessment has been recently explored in the development of the guidance document of the Scientific Committee on "*the use of the weight of evidence approach in scientific assessments*" and exemplified for data poor compounds such as emerging contaminants of anthropogenic or natural origin (EFSA, 2017).

The purpose of this scientific report performed under the Negotiated procedure reference: NP/EFSA/AFSCO/2016/01 is to illustrate the development of innovative *in silico* QSAR models using EFSA's OpenFoodTox and other databases for HRA and ERA.

2. Data and Methodologies

2.1. Toxicological Databases

Reference Toxicological databases have been used to develop the 4 QSAR models and these include the EFSA's OpenFoodTox database, the Fraunhofer RepDose Database, the Terrestrial US-EPA ECOTOX database and the VEGA platform.

1. For the NO(A)EL QSAR models in rats, sub-chronic toxicity data were collected from OpenFoodTox and the Fraunhofer RepDose Database;

2. For the bee toxicity model, acute contact toxicity data (LC_{50}) in bees were collected from the DEMETRA database (Benfenati et al., 2011), the terrestrial US-EPA ECOTOX database present in the OECD QSAR Toolbox, vers. 3.3. (www.qsartoolbox.org) and from OpenfoodTox.

3. For the rainbow trout model, acute contact toxicity data (LC_{50}) were collected was extracted from OpenFoodTox.

4. For the model predicting the toxicity values from Openfootox, the available models from the VEGA platform were used.

2.2. Data Curation and Datasets

Once the data sources have been identified, as described in the previous section for the relevant endpoints, data curation was performed.

For OpenFoodTox, each individual toxicological study were retrieved from EFSA's Data collection framework and move to Excel with an individual ID number assigned to facilitate easy finding of data when needed. For the development of (Q)SAR model, a correct identifier for each chemical is needed,

including "SMILES", and all data were checked for correctness for the substance identification "SUB_NAME", "SUB_CASNUMBER", "COM_CASNUMBER" and "SMILESNOTATION". Consequently, each substance was checked for having a unique SMILES which was then associated with a unique CAS and where possible. For this purpose, tools such as ChemID Plus (ChemIdPlus, 2016), Chemspider (ChemSpider, 2015) and Instant JChem (ChemAxon, 1998) were used. Circumstances under which 1.SMILES and CAS would identify different structures, the CAS associated to the SMILES were identified considering the most frequent association, 2. SMILES or CAS number were missing, the data gap were filled in when possible. At the end of this data curation step, each SMILES set was associated to a unique CAS and a "FINAL CAS" column named was added. Inorganic compounds were not used for the exercise since QSAR models are typically developed for organic compounds.

In second step, all structures were normalised for their neutral form with IstMolBase (IstMolBase, 2013) and a column named "SMILES neutralised" was added. In this process, the salts (such as sodium salt, or salt with hydrochloric acid) were transformed into the corresponding neutral form, since the *in silico* models most commonly use the neutral form. It was also noticed that the chemical form used in the protocol for the experimental assay applies a controlled pH, obtained with a buffer solution. Under these conditions, the original salt and the neutral substance were both transformed into the salt of the corresponding buffer and further checks were performed for completeness (i.e. presence of redundant compounds after transformation of the salt into the neutral form).

The final data curation step was the transformation of all SMILES in the VEGA format to have a unique harmonised codification and avoid multiple SMILES formats so that the column "SMILES VEGA" was added to the dataset.

At the end of the data curation process, each SMILES structure in the dataset was associated to a unique CAS so that each substance could be identified easily in the dataset.

After the data curation, a unique set of 1218 substances were pruned following specific criteria as follows: substances without CAS or SMILES or both, substances with complex salt, inorganic compounds or metal and multiple compounds. After the end of the data curation process, the OpenFoodTox dataset was ready to be analysed to extract subset of data for different endpoints and species. Further pruning was necessary, depending on the endpoint, considering parameters specific to the experimental protocol, as described below.

2.3. Methodology for the development of the QSAR models

In the context of toxicological sciences, QSAR models aim to predict toxicity values while relating a training set of data as a set of variables to predict the potency or toxicity of a test set of data either as continuous models or classification models. In this report, all models used a training set and a test set, and in some instances, virtual sets and validation sets. Most models described in this report were continuous models aiming to predict NOAEL (rat model), LC_{50} (rainbow trout) or different toxicity values (VEGA model) with the exception of the bee model which is a classifier aiming to predict threshold values predefined in the training set.

2.3.1. QSAR models for the prediction of NOAEL in rats

Sub-chronic toxicity data (90 day studies) in rats for HRA were extracted from OpenFoodTox using the following columns and filters were: "TESTTYPE" filtering "subchronic", "SPECIES" filtering "Rat", "ROUTE" filtering "oral:" (all oral route was included), "EXP_DURATION_DAYS" filtering greater that 76 days, "ENDPOINT" filtering "NOAEL" and "NOEL", "QUALIFIER" filtering "=" and "COM_STRUCTURESHOWN" filtering "compound". Furthermore, substances that were pruned, for

reasons indicated in 2.1, were removed from the final dataset. Further checks of the database were performed particularly for duplicates and for substances with more than one experimental value the lowest value was used. The NOEL/NOAEL rat dataset was composed of 166 compounds amongst which a cluster of aliphatic chain of aldehydes, carboxylic acids and alcohols was found including 39 compounds with the same value as a read across based on the experimental value of one congener. In this situation, only the experimental value for the single chemical was used. The final database from OpenFoodTox provided a set of 128 compounds.

The final database was then complemented with the Fraunhofer @RepDose Database dataset composed of a dataset of 362 sub-chronic NOELs in rat with a final value set, after analysis of duplicates, composed of 357 substances (Bitsch et al., 2006). The two data sets, from OpenFoodTox and Fraunhofer @RepDose Database, were merged using the lowest value for duplicate and producing final database containing 487 compounds.

The CORAL software (http://www.insilico.eu/coral) was used as a tool to calculate optimal molecular descriptors using the Monte Carlo technique, which involves the maximisation of correlation coefficients between the descriptor and endpoint. The CORAL descriptor finds the presence (or absence) of eight chemical elements (nitrogen, oxygen, sulfur, phosphorus, fluorine, chlorine, bromine, and iodine) and different kinds of chemical bonds (double bond, triple bond, and stereo chemical bond). Hybrid descriptors were calculated for the NOEL endpoint. The hybrid descriptor takes into account molecular features extracted from simplified molecular input-line entry system (SMILES) and from hydrogen-suppressed graph (HSG). In fact, SMILES and HSG are different (similar but non-identical) representations of the molecular structure. Figure 1 shows the interconnection between SMILES and HSG. The model developed with CORAL (Figure 1) is a model for continuous values.

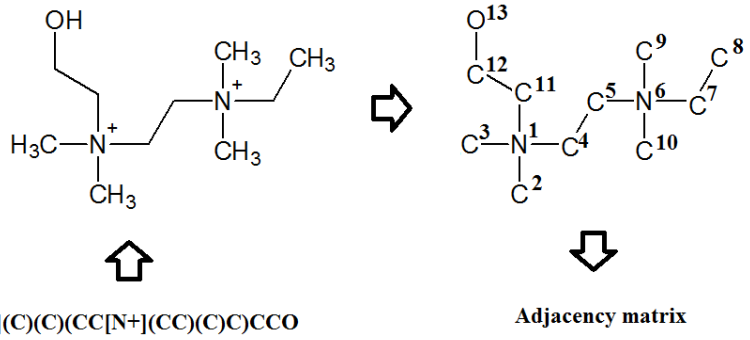

Figure 1.Interconnection between SMILES and HSG represented by the adjacency matrix

2.3.2. QSAR models for the prediction of acute toxicity in rainbow trout

The datasets reporting acute toxicity in rainbow trout (LC_{50}) for pesticides were extracted from the OpenFoodTox database (EFSA, 2017). To extract the dataset, the datasets were extracted using the following filters and columns: "STUDY_CATEGORY" filtering "Ecotox (water compartment)",

"TESTTYPE" filtering "acute toxicity", "SPECIES" filtering "Rainbow trout", "EXP_DURATION_DAYS" filtering "4", "ENDPOINT" filtering "LC50", "QUALIFIER" filtering "=" and "SUB_OP_CLASS" filtering Pesticides. Furthermore, substances pruned for reasons indicated in 2.1. were removed from the final dataset. 13 compounds had one or more values, and the lowest was selected for a total of 116 compounds in the final dataset.

2.3.3. QSAR model for the prediction of acute contact toxicity in bees

The Bee data set for acute contact toxicity in bees was extracted from the following sources:

- DEMETRA database (Benfenati et al., 2011).
- Terrestrial US-EPA ECOTOX database present in the OECD QSAR Toolbox, vers. 3.3 (www.qsartoolbox.org).
- OpenFoodTox from EFSA

Toxicity data for pesticides expressed as LD_{50} were selected from each database described above following criteria from the OECD guideline (OECD, 1998). LD_{50} values were relative to contact exposure in honey bees (*Apis mellifera*) and measured after 48h of exposure (LD_{50}48h). Compounds in the datasets were classified as low toxicity for LD_{50} greater than 100 µg/bee, moderately toxic for LD_{50} between 1 and 100 µg/bee, and highly toxic for LD50 lower than 1 µg/bee according to the Pesticide Properties Database (PPDB) (http://sitem.herts.ac.uk/aeru/iupac/docs/Background_and_Support.pdf). For compounds with more than one experimental value, the associated variability was evaluated using the strategy employed by Benfenati and collaborators (Benfenati et al., 2011) as the ratio between duplicated experimental data (x/y) as follows:

a. If x/y was <= 5 the average of the experimental data was kept;

b. If x/y was >5 the compound was removed from the dataset.

The datasets contained overall 256 pesticides with acute contact toxicity LD_{50} from OpenFoodTox, the DEMETRA database and the OECD QSAR toolbox version 3.3. The datasets were split in training sets (205 compounds as 80% of original data set) and validation sets (51 compounds as 20% of the original datasets).

The Bee model was constructed using the k-nearest neighbors (k-NN) methodology, applying the in-house software istkNN (Manganaro et al., 2016). This model was run on acute toxicity data after contact exposure (LD50 48 H) in bees extracted from OpenFoodTox, the terrestrial US- EPA ECOTOX database from the OECD QSAR Toolbox and the DEMETRA database. Conflicting values were rejected. Overall, the dataset covered 256 chemicals, which were the basis to develop the *in silico* model. Three toxicity classes based on LD50 values were defined with chemicals with toxicity values >100 µg/bee, between 1 and 100 µg/bee and <1 µg/bee. Two models were developed using two threshold to classify as toxic the compounds; the first model uses threshold under 1 µg/bee to classify toxic a compounds and the second model uses under 100 µg/bee to classify compounds as toxic. This k-NN predictive model has been published in the literature (Como et al., 2016).

2.3.4. QSAR models for toxicity prediction in OpenFoodTox

The VEGA platform is available as open-source software (http://vega-qsar.eu) and contains a large number of models to predict different endpoints using SMILES as input. Models present in VEGA are organised in four categories:

- Human toxicity
    - 4 models to predict mutagenicity and one mutagenicity consensus model
    - 4 models to predict carcinogenicity
    - 2 models to predict developmental toxicity
    - 2 models to predict effects on estrogen receptors
    - 1 model to predict skin sensitization
    - 1 model to predict hepatotoxicity
- Ecotoxicological toxicity
    - 4 models to predict LC50 on fish
    - 2 models to predict LC50 on Daphnia Magna
    - 1 models to predict LD50 on bee
- Environmental fate and toxicity
    - 3 models to calculate BCF
    - 1 model to calculate ready biodegradability
    - 3 models to calculate persistence
- Physical/chemical properties
    - 3 models to calculate LogP

The description of the models is present in VEGA clicking on "?" button next to all models in the model selection section. VEGA models were applied to predict toxicity of organic compounds from the OpenFoodTox database. In this context, inorganic, organo-metallic compounds, mixtures or chemicals without unique chemical structure were excluded from the modelling. After curation of openfoodtox, a total of 933 compounds were suitable for the exercise and all predictive QSAR from the VEGA platform described above models were run.

2.4. QSAR predictions of NOAEL in rats

The best QSAR model developed with the CORAL software for the prediction of NOAEL values is illustrated in figure 2. The continuous model developed gave good results with correlation coefficient R^2 values ranging between 0.58 and 0.70 in the training sets and between 0.50 and 0.70 in the virtual set. The QSAR models cover all compounds of the dataset and summary statistics are given in figure 3. The coverage of 100% demonstrates that the model is applicable to all the chemicals included in the modelling.

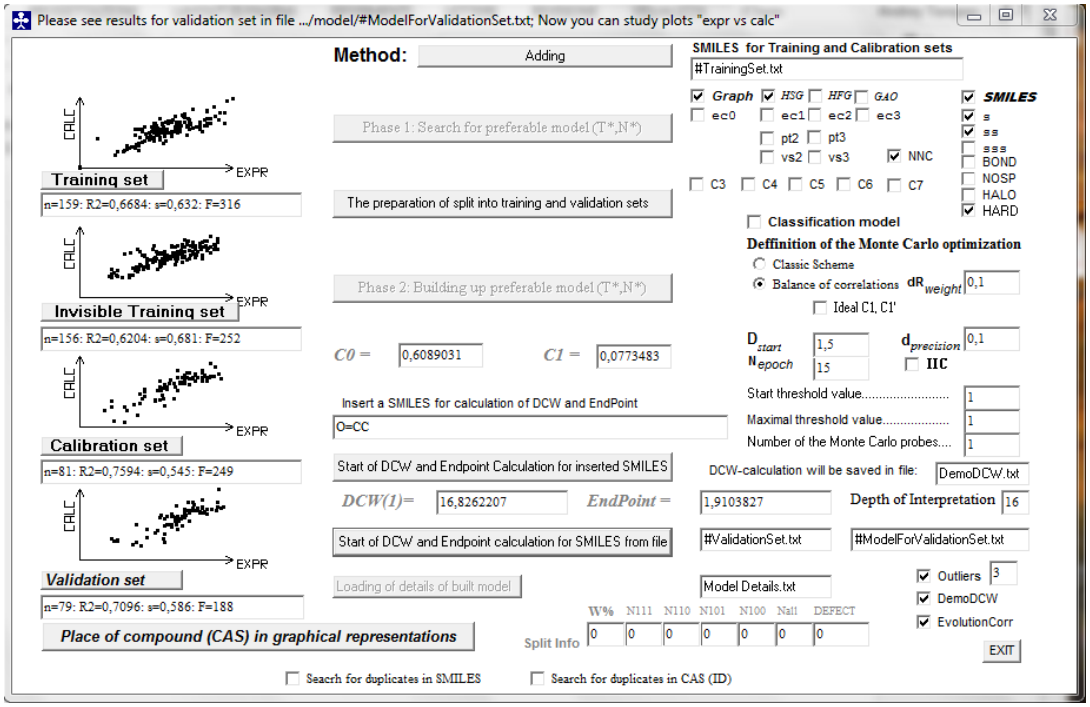

Figure 2. Best QSAR model for the prediction of NOAEL values in rats

R2 = squared correlation coefficient; s is Root squared mean error; F is Fischer F-ratio

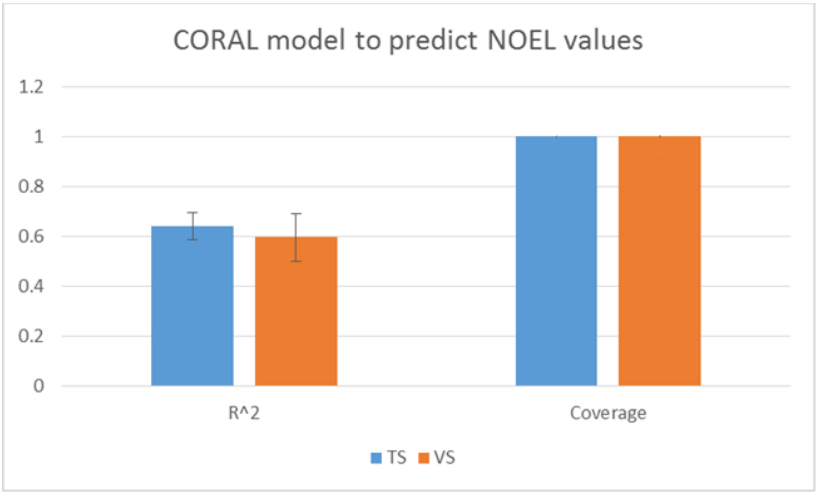

Figure 3. Mean values for the QSAR prediction of NOAEL values in rats

2.5. QSAR predictions of LC_{50} in rainbow trout

The CORAL software was used to predict LC_{50} values in rainbow trout using pesticide toxicity data extracted from OpenFoodTox. The datasets were randomly split into training (≈40%), invisible training (≈40%), calibration (≈10%) and validation (≈10%) sets on three separate sampling. The best model is illustrated below in figure 4.

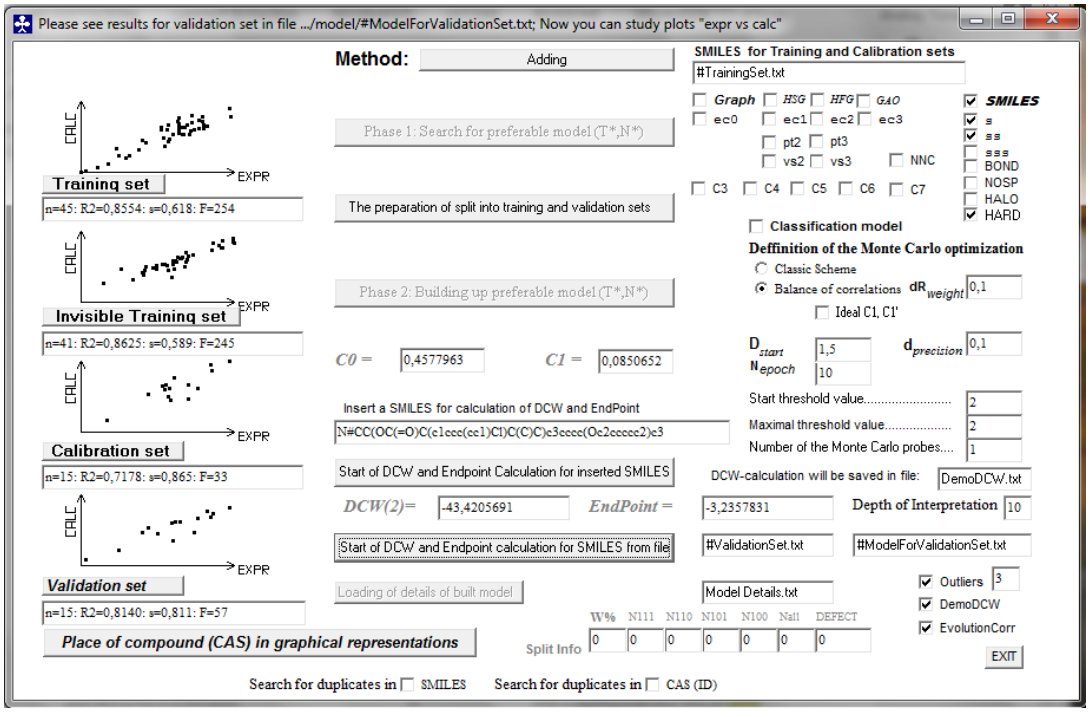

Figure 4. Best predictions from the CORAL model for prediction of LC_{50} values in rainbow trout

R2 = squared correlation coefficient; s is Root squared mean error; F is Fischer F-ratio

The QSAR model developed with CORAL for the prediction of LC_{50} values in rainbow trout resulted in correlation statistics of R^2 between 0.85 and 0.88 in TS and 0.63 and 0.89 in VS; all models cover all compounds of the dataset and summary statistics are given. Results are shown in Figure 5.

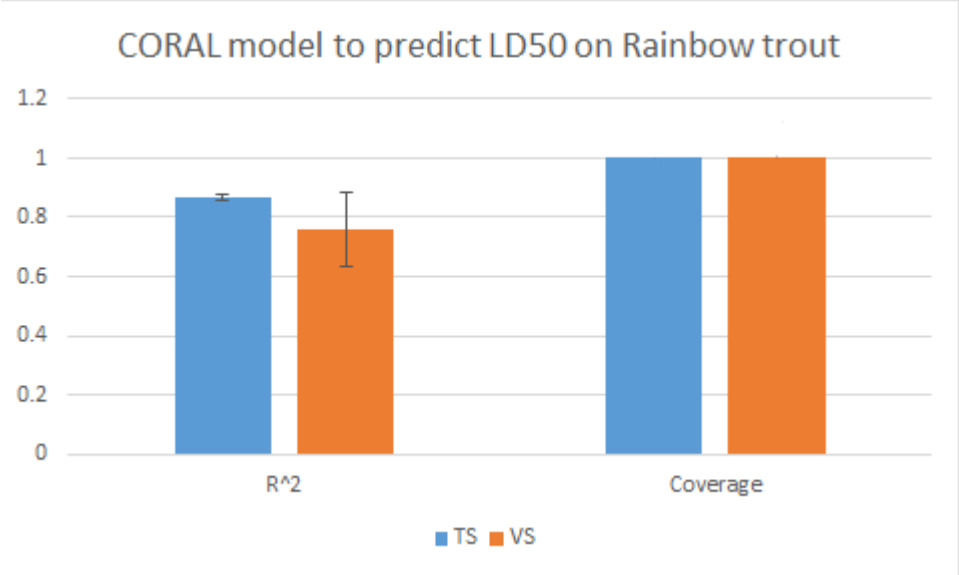

Figure 5. Mean values for the QSAR prediction of LC_{50} values in rainbow trout

2.6. QSAR prediction of acute contact toxicity in bees

The classification model developed by k-NN gave very good statistics for specificity at demonstrated with the Matthew Correlation Coefficient. Results are illustrated in Figure 6.The model is described in details elsewhere (Como et al., 2016).

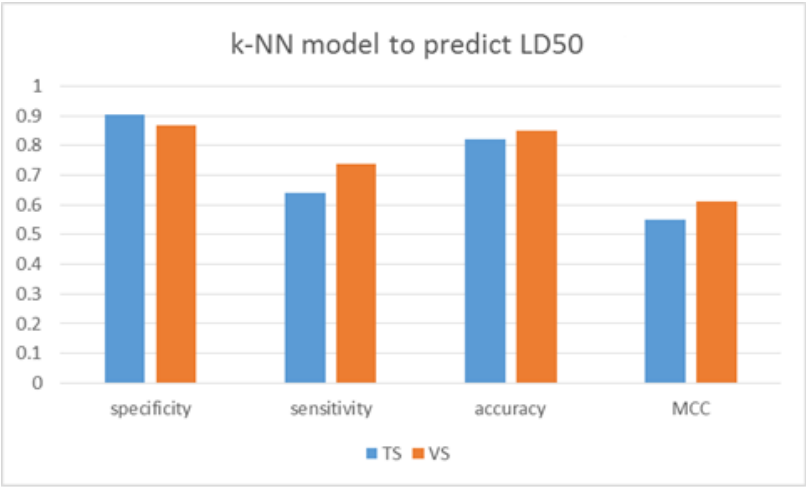

Figure 6. Mean values for the QSAR predictions of contact LD_{50} in Bees

2.7. QSAR predictions of toxicity in OpenFoodTox using VEGA

QSAR Predictions with 33 models present in VEGA were made on 933 substances of the OpenFoodTox database. These predictions provide a large amount of information for all substances and provide a very useful preliminary analysis for the hazard assessment of these substances. In addition to predictions, VEGA provides other useful information such as the presence or absence of experimental values in the training set of the models, the presence of known structural alerts for toxicity or for non-toxicity (if models are based on structural alerts) and the applicability domain index, that permits to know the model reliability for a specific chemical. The applicability domain index is calculated through taking in account a number of parameters:

- Similarity: This index takes into account how similar are the first three or two (depend on different models) most similar compounds found in the model TS.
- Accuracy: This index takes into account the classification accuracy in prediction for the three or two (depend on different models) most similar compounds found in the model TS.
- Concordance: This index takes into account the difference between the predicted value and the experimental values of the three or two (depend on different models) most similar compounds in the model TS.
- Atom-centred fragments: This index takes into account the presence of one or more fragments that aren't found in the model TS, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur less than three times in the TS), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND.
- Model descriptors range. This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. This index is present only if model is based on descriptors.

An applicability domain index value between 0.8 or 0.9 (model dependent) is associated with reliable prediction whereas applicability domain indexes value between 0.8 and 0.75 or lower are associated with medium or low reliable prediction. Such an applicability domain index provides a powerful quantitative instrument to measure the reliability of the prediction power of the model and an intrinsic parameter of all models present in the VEGA platform.

3. Conclusions and recommendations

This scientific report aimed to illustrate the development of *in silico* quantitative structure activity relationship (QSAR) models. These models provide innovative tools to predict toxicity values or classify thresholds for HRA and ERA using EFSA's Chemical hazards database: openfoodtox and other relevant databases including the US-EPA terrestrial database and the Fraunhofer RepDose and the open source VEGA platform. Two continuous QSAR models were developed using data for sub-chronic toxicity in rats For ERA, two QSAR models were developed for predicting acute toxicity in rainbow trout as a continuous model and to predict acute contact toxicity data in bees as a classification model. Finally, a global model was developed to predict toxicity values from OpenFoodTox using the VEGA platform.

This report shows that these QSAR models provide support to scientists in the evaluation of human and eco-toxicological properties and will be published in the near future in the VEGA platform as open source tools. The use of these *in silico* models has a strategical impact, related to the exploitation of this novel database, offering the possibility in the future to navigate through the data and extend the information present in the database towards chemicals not present in the data using *in silico* tools. Good prediction results for toxicological endpoints have been achieved. Very promising results were achieved in the modelling of NOEL endpoint. In addition, the work performed during the development of the model will be useful for future modelling, since a "factory" of models using different approaches has been tested within KNIME (Berthold et al, 2007). This suite of models allows the generation of many predictive models simultaneously and the selection of the most predictive model. In this manner, it is foreseen in the future that scientists will be able to develop many models based on consistent data in a high throughput manner.

Further, these approaches further demonstrate the usefulness of large open source toxicological databases, providing historical data to boost bioinformatics analysis and *in silico* modelling particularly for compounds with scarce toxicological data.

The use for the OpenFoodTox database is potentiated thanks to the availability of values on tens of endpoints obtained by VEGA. It is acknowledged that these are predicted values, not to be confused with experimental values. However, this large set of predicted values represents a valuable starting point to further, deeper assessment, and for prioritization purposes.

Further work is recommended to explore the application of *in silico* tools in chemical risk assessment. Of particular interest is the development of systematic and harmonised approaches for the use of QSAR, read across using physico-chemical properties, toxicological and toxicokinetic data for a number of endpoints and species relevant to HRa and ERA as well as for the refinement of the TTC approach.

References

Benfenati E, Boriani E, Craciun M, Malazizi L, Neagu D and Roncagioni A, 2011. Databases for pesticide ecotoxicity. In: Benfenati, E. (Ed.), Quantitative Structure-activity Relationships (QSAR) for Pesticide Regulatory Purposes. Elsevier, pp. 59e81.

Benfenati E, Manganaro A and Gini G, 2013. VEGA-QSAR: AI inside a platform for predictive toxicology. Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy Published on CEUR Workshop Proceedings Vol-1107

Bitsch A, Jacobi S, Melber C, Wahnschaffe U, Simetska N and Mangelsdorf I, 2006. RepDose: A database on repeated dose toxicity studies of commercial chemicals—A multifunctional tool. Regul. Toxicol. Pharmacol. 46, 202–210.

ChemAxon, 1998. Instant JChem was used for structure database management, search and prediction, Instant JChem 16.5.30.0, 2016, ChemAxon (http://www.chemaxon.com)"

ChemIdPlus, 2016. National Institutes of Health 1887. https://chem.nlm.nih.gov/chemidplus/

ChemSpider, 2015. Royal Society of Chemistry 2015. http://www.chemspider.com/

EFSA (European Food Safety Authority), 2014. Report on 'Further development and update of EFSA's Chemical Hazards Database NP/EFSA/EMRISK/2012/01', S-IN Soluzioni-Informatiche; Supporting Publications 2014: EN-654. [103 pp.]. Available online: http://www.efsa.europa.eu/it/supporting/pub/654e.htm

EFSA (European Food Safety Authority), 2017. OpenFoodTox: EFSA's open source toxicological database on chemical hazards in food and feed. EFSA Journal 2017;15(1):e15011 [3 pp.]. DOI: 10.2903/j.efsa.2017.e15011

IstMolBase, 2013. IstMolBase v.1.0.2 https://www.kode-solutions.net/

Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B, 2007. KNIME: the konstanz information miner. In Stud. Classif. Data Anal. Knowl. Organ: Springer;

Manganaro A, Pizzo F, Lombardo A, Pogliaghi A and Benfenati E, 2016. Predicting persistence in the sediment compartment with a new automatic software based on the K-Nearest Neighbour (k-NN) algorithm. Chemosphere 144, 1624e1630.

OECD (Organisation for Economic Cooperation and Development), 1998. Test No. 214: Honeybees, Acute Contact Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2. OECD Publishing, Paris

Pizzo F, Lombardo A, Manganaro A and Benfenati E, 2016. A New Structure-Activity Relationship (SAR) Model for Predicting Drug-Induced Liver Injury, Based on Statistical and Expert-Based Structural Alerts. *Front. Pharmacol*. 7:442. Doi: 10.3389/fphar.2016.00442

Abbreviations

EFSA	European Food Safety Authority
QSAR	Quantitative Structure activity relationship
US-EPA	Environmental Protection Agency of the United States