doi:10.20944/preprints201609.0107.v1

Article

Comprehensive Safety Evaluation of the Chemicals with Human and Environmental Relevance

Matteo Floris [1,2,*,§], Federica Albanese [3,§], Ricardo Medda [1] and Emilio Benfenati [3]

1 CRS4 - Center for advanced studies, research and development in Sardinia, Loc. Piscina Manna, Building 1, 09010 Pula (CA), Italy; ricardo.medda@crs4.it
2 Department of Biomedical Sciences, University of Sassari, Viale San Pietro, 07100 Sassari, Italy
3 IRCCS – Istituto di Ricerche Farmacologiche "Mario Negri", Department of Environmental Health Sciences, Laboratory of Environmental Chemistry and Toxicology, Via La Masa 19, 20159 Milan, Italy;
federica.albanese@marionegri.it (F.A.); emilio.benfenati@marionegri.it (E.B.)

*** Correspondence: matteo.floris@gmail.com**
§ These authors contributed equally to this work

Abstract.

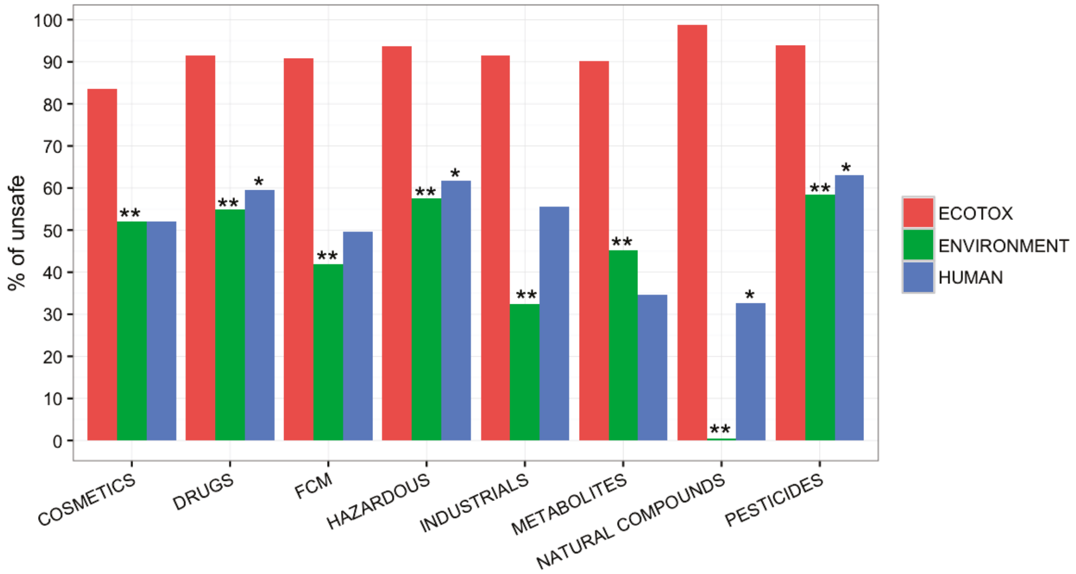
Reducing the levels of toxic chemicals that cause adverse health and environmental impacts is a challenge for the international community. There is a need of a global strategy. Indeed, too often the problems associated to the exposure of chemical substances is recognized a posteriori, in the presence of consequences already observed. The prediction of the likely effects of chemical exposure on human health is based on classical tests with animals, which are time and money consuming, may deviate from an accurate prediction towards humans, and arises ethical concerns. Regulations are now considering the adoption of *in silico* (or computational) methods, which can be used for prioritizing substances according to the probability to be toxic for the biosphere. Several initiatives have prioritized chemicals, typically based according to some criteria, such as chemicals which may be endocrine disruptors, or persistent, bioaccumulative, or toxic (PBT). However, these initiatives focused on a certain range of adverse properties, and covered a certain number of substances only. We applied a set of largely validated and widely used predictive methods to large collections of chemicals: (i) to about 340,000 with a defined function, and (ii) to about 6 millions, which have been synthetized. The aim of this study is to quantify the putative impact of existing and future chemicals on towards human health, ecological and environment properties. The impact on the environment is the cause of major concern. This is the case of pesticides and hazardous, which is quite expected; however, also pharmacologically active candidate compounds of natural origin may have a high level of ecotoxicity. Pesticides and hazardous are also the categories of higher concern for humans, followed by pharmaceuticals. The pesticides and the hazardous are the categories of higher concern also on the environmental point of view. The results of our analysis could be the basis for the identification of new safety rules.

Keywords: QSAR; Toxicology

Introduction.

Chemical substances are the basic components of our organisms, and anthropogenic chemicals are continuously produced. In the past it has been argued if natural compounds are safer than anthropogenic ones (**Ames1973**). Besides a debate on the risks of synthetic substances, which

Figure 1. Figure 1: Percentage of putatively unsafe compounds in each chemical category. Significance is calculated by comparing the % of putatively safe and unsafe compounds for a given category of chemical to the % of putatively safe and unsafe compounds in the SIGMA catalog. Legend: * = chi square p-value < 0.05, ** = p-value < 0.001.



may recall a new edition of the beau savauge dispute on the risks of the progress (**Rousseau1762**), it is important to analyse on a neutral point of view the possible impact of chemical substances in general, because there is quite a large ignorance of the effects that substances have on the human health and on the environment. The obtained information may drive initiatives focussed on the identification of prioritises, and substitutions or avoidance of risky substances.

First, we compiled and analysed a collection of about 340,000 substances with different functions (see Methods for a detailed list of sources). The attention was given to properties commonly used for the risk assessment of chemical: in details, endpoints of concern for the human health, such as carcinogenicity, mutagenicity and reproductive toxicity (CMR), for ecotoxocological properties, such as fish and Daphnia Magna acute toxicity, and environmental properties, such as bioconcentration factor (BCF), ready biodegradability, and persistence in water, soil and sediments. These endpoints have been estimated using the software VEGA (**Benfenati2013**). As a general remark, we notice that the predictive models are conservative. We may expect that this will affect the overall results, and thus the main lesson is the relative impact of the different chemical categories, and not the absolute percentage of chemicals with a given property value.

Results.

Impact towards human health.

On average, the categories of substances with the highest relative risk for human health are pesticides and hazardous. In particular, when a substance belongs to one of the above categories, the probability of being harmful for humans is higher than 60%. The third category of higher concern is approved drugs, followed by industrial chemicals.

Impact towards ecotoxicological endpoints.

Also in this case the highest environmental impact is predicted for pesticides and hazardous, together with substances of natural origin. Thus, the last category has an opposite behaviour towards human and ecotoxicological endpoints.

Impact towards environmental properties.

Pesticides and hazardous are categories of higher concern also towards environmental properties, together with pharmaceuticals. On the opposite side, natural compounds have almost a nearly absent impact.

Overall, these results indicate that certain categories of chemicals, like pesticides and hazardous, represent a general risk for the different targets. Metabolites and food contact materials are among the less critical substances in general.

Conversely, other categories, like natural compounds, have different behaviour, since they are the safest substances towards humans and the environment (considering BCF and persistence) while they are quite toxic towards aquatic animals. Industrial chemicals are also with different behaviour. In order to assess a significance of the above described observations, we compared the percentage of putatively unsafe compounds in each category to the percentage of compounds from a huge database of commercial compounds.

Evaluation of the commercial chemical space.

The results we presented refer to the predictions obtained with higher reliability, as defined by the VEGA software. This is related to the data and the chemicals used to build up the models, which are about few hundreds or few thousands in the most profitable case (mutagenicity). The extrapolation of these results to the much higher number of compounds, as we did, is an exercise which is useful to get rough indications, and to identify priorities and situations of concern. With these limits we also addressed an even larger number of compounds, about 6 millions, from a commercial chemical catalogue. These substances are in the domain of research, and in most cases are not physically available. Nevertheless, they may represent the future substances. Thus, it becomes important to join the information on the possible reasons of concern which may arise in case of their applications and use. Overall, considering only highly reliable predictions, more than 99% of compounds are into the applicability domain of VEGA in at least 1 of the 3 human endpoints, and of these 46% are labelled as toxicant; only less than 1% of the database is into the applicability domain for at least 1 of the ecotoxicological endpoints, and in this case about 94% of the chemicals are labelled as toxicants; finally, for about 5% of the database compounds we have reliable predictions for at least 1 of the environmental endpoints, showing that only 10% of the compounds would be active (See Fig. 2).

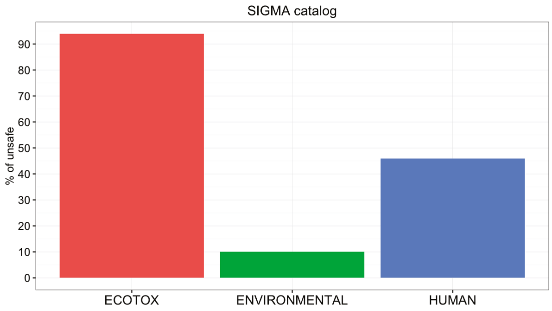


Figure 2. Figure 2: percentage of putatively unsafe compounds in the "chemical universe" (here intended as the SIGMA Aldrich database, considering only highly reliable predictions).

Discussion

Within this study we evaluated a large universe of existing, and maybe next generation substances. We identified the relative adverse impact of the different chemicals, together with the reasons of concerns. The other side of the evaluation is also useful: the safety of the substances, and in particular the identification of reasons associated to this behaviour. This is an important feature of the in silico methodology we used: the results on the adverse effect is in many cases related to rules which can be used to avoid the adverse effect, or, on the same premises, the complementary rules related to safety can be exploited to drive the preparation of a new generation of substances with reduced impact.

The strategy we adopted was balanced, covering both effects towards human health and the environment. At the same time, our approach was transparent, allowing spotting the main causes of concern.

Material and Methods.

Dataset.

We collected different collection of chemicals belonging to different categories as described below:

1. **COSMETICS – compounds used as cosmetic ingredients.**

   We used the COSMOS database v1.0, freely available on `http://www.cosmostox.eu`, that contains more than 80,000 chemical records with more than 40,000 unique structures; it is the result of a project funded by the European Commission and Cosmetic Europe (European trade association for the cosmetic, toiletry and perfumery industry), bringing together expertise from Europe and USA industry, academia and regulatory agencies.

2. **FOOD AND CONTACT MATERIALS (FCM)**.

   This category comprises materials either intended to be brought into contact with food, are already in contact with food, or can reasonably be brought into contact with food or transfer their constituents to the food under normal or foreseeable use (definition on Regulation (EC) No 1935/2004). This includes direct or indirect contact.
   We processed with our approach a U.S. Food and Drug Administration (FDA) database of indirect food additives (over 3000 substances) [Title 21 of the U.S. Code of Federal Regulations(21CFR) Parts 175, 176, 177, and 178]. Indirect food additives are used in food contact articles, including adhesives and components of coatings, paper and paperboard components, polymers, adjuvants and production aids. The database is free available on `http://www.fda.gov/Food/IngredientsPackagingLabeling/PackagingFCS/IndirectAdditives/default.htm`.
   We processed also other lists obtained in collaborations with European Authorities: (1) the FCMs database available from the European Joint Research Center (JRC) that includes 4610 entries. (2) Inventory list of 4948 chemicals about paper, board and printing inks substances **Boriani2016** Printing inks data is from Federal Office of Public Health Switzerland and paper and board from the inventory list produced by Council of Europe. (3) FCMs lists of 2484 chemicals from the Scientific Institute of Public Health of Belgium (WIV-ISP).
   Furthermore, we processed the 24394 entries of the FooDB v.1 (free available: `http://foodb.ca/` by HMDB of TMIC- Canadian Metabolomic Information Center), that is the largest resource on food constituents in the world, containing information on both macronutrients and micronutrients, including many of the constituents that give foods their flavor, color, taste, texture and aroma.

3. **METABOLITES.** Here we considered small compounds that are intermediates and products of biological metabolism.

The Human Metabolome Database (HMDB) v3.6 of Canadian Metabolomic Information Center (TMIC) is a freely available electronic database, containing detailed information on 41993 small molecule metabolites found in the human body (`http://www.hmdb.ca/`) (**Wishart2007**) (**Wishart2013**) (**Wishart2009**).

4. **DRUGS**

   Pharmaceutical drugs, commonly abbreviated as drugs, are subministered as a medicine and care.
   They can be synthetic or have natural origin. The database we gathered is the DrugBank v.5 (freely available at `http://www.drugbank.ca/`) that combines detailed drug data with comprehensive drug target information. The database contains 8206 drug entries including 1991 FDA-approved small molecule drugs, 207 FDA-approved biotech (protein/peptide) drugs, 93 nutraceuticals and over 6000 experimental drugs. For our purposes we used approved drugs only (**Wishart2006**).

5. **NATURAL COMPOUNDS** This category includes compounds and mixtures produced by living organisms.
   Natural products include large groups of substances from a variety of sources such as plants, animals, fungi and bacteria. The definition include also some minerals, vitamins and probiotics. They can be extracted from nature or synthetized as natural-like chemicals. Natural products are considered a rich source for new drugs.
   To assess this category we chosen the ASINEX BioDesign (v. November 2015), a free available library (download: `http://www.asinex.com/natprod`) of 129359 natural product-like compounds. The ASINEX's BioDesign approach incorporates key structural features of known pharmacologically relevant natural products (e.g. alkaloids and other secondary metabolites) into synthetically feasible medicinal chemistry scaffolds.

6. **PESTICIDES.** Formulations of chemicals released intentionally in the environment to kill weeds (herbicides), insects (insecticides), fungi (fungicides), rodents (rodenticides), and others. Pesticides have been linked to a wide range of human health and ecological hazards so monitoring the presence of their residues in environment and in foods commits the authorities worldwide.
   We applied our methods on the list of EPA Registered Pesticides under FIFRA (Federal Insecticide, Fungicide, and Rodenticide Act), corresponding to a dataset of 1377 chemicals. The list is available on `http://scorecard.goodguide.com/chemical-groups/one-list.tcl?short_list_name=pest`.

7. **INDUSTRIALS.** All substances and their derivatives produced and/or processed on an industrial scale. We applied our classification algorithm to the full list of the 8771 registered substances under REACH Regulation (CE) n. 1907/2006 (release February 2016), and to the 270 SVHC (substance of very high concern) of the priority list (release February 2016), both exported from ECHA website (`https://echa.europa.eu`).
   We assessed also chemicals included into the Distributed Structure-Searchable Toxicity (DSSTox) Database collected by EPA U.S (v. October 2015) and into the EPA Toxic Substances Control Act (TSCA) Chemical Substance Inventory (v. January 2015). DSSTox provides a high quality public chemistry resource for supporting improved predictive toxicology. It is freely available at `ftp://ftp.epa.gov/dsstoxftp`, and contains 154925 entries. TSCA Inventory (freely available `https://www.epa.gov/tsca-inventory/how-access-tsca-inventory`) contains non-confidential portion of the chemical substance listings on the TSCA Inventory, identified by Chemical Abstract Service (CAS) Registry Number and Chemical Abstracts (CA) Index Name for amount of 67635 entries.

8. **HAZARDOUS.** Substances or energy forms introduced into environment able to cause adverse effects. They may be endogenous for the environment considered, or may be

exogenous and/or xenobiotic. Many known substances and mixtures are identified as hazardous chemicals and are cause of concern for human and environmental health. We used a database of 401 chemicals (free download http://www.nite.go.jp/en/chem/chrip/chrip_search/intSrhSpcLst?_e_trans=&slScNm=TD_01_002) available by NITE (Japan) Chemical Risk Information Platform (NITE-CHRIP) as part of the Safety Inspections for Existing Chemical Substances carried out by the Ministry of Health, Labour and Welfare of Japan to promote chemical substance safety inspections. Japan cooperates on the OECD: HPV (High Production Volume) Chemicals Program.

In addition to all the databases described by category, we also processed a database that contains molecules referred to all the categories described above and that can be cause of concern because they should be toxic or because they are produced in high volumes. For the first case the database is the Toxin and Toxin Target Database (T3DB) that combines detailed toxin data with comprehensive toxin target information. The free version (downloaded: http://www.t3db.ca/downloads) dated in February 2016 contains 3526 toxins, including hazardous, pesticides, drugs, and food toxins. T3DB project is supported by the Canadian Institutes of Health Research. For the second case, we investigated on the U.S. High Production Volume (USHPV) database (freely available https://iaspub.epa.gov/sor_internet/registry/substreg/list/details.do?listId=74), that includes 3357chemicals manufactured in or imported into the United States in amounts equal to or greater than 1 million pounds per year.

SIGMA Aldrich catalogue.

We applied our classification algorithm to the full Sigma Aldrich catalogue (version: March 2016). The catalogue is a collection of 7,282,931 chemical structures. We normalized the structures (Methods, Schroedinger's normalizer tool), then we processed only the non-redundant set of 6,644,492 structures.

When the data were available as CAS number, we used the CIRpy, a Python interface for the Chemical Identifier Resolver (CIR) by the CADD Group at the NCI/NIH (https://github.com/mcs07/CIRpy), to obtain the corresponding SMILES strings (http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html). All the SMILES string were then converted to SDF structural files with the Openbabel tools (**OBoyle2011**).

Models.

For each dataset, we then calculated the following models available within VEGA platform v.1.1.1. (free download: http://www.vega-qsar.eu/):

1. Models with human health interest: We used models able to assess the toxicity for cardinal endpoints to evaluate the impact on human health such as Mutagenicity, Carcinogenicity and Developmental Toxicity. For each endpoint, more than one model was used: four for Mutagenicity; four for Carcinogenicity and two for Developmental Toxicity. Each model is described below.

   (a) Mutagenicity

      i. Mutagenicity (Ames Test) model (CAESAR) 2.1.13 provides a qualitative prediction of mutagenicity on Salmonella typhimurium (Ames test). The model extends the original CAESAR model, freely available at: http://www.caesarproject.eu/software/. Structural Alerts have been taken from the Benigni/Bossa rulebase for mutagenicity and carcinogenicity (implemented as a module of Toxtree software).

      ii. Mutagenicity (Ames Test) model (SARpy/IRFMN) 1.0.7 provides a qualitative prediction of mutagenicity on Salmonella typhimurium (Ames test). The model has been built as a set of rules, extracted with SARpy software from the training set from the Mutagenicity CAESAR model. The original work has been extended, resulting in two sets of rules for mutagenicity (112 rules) and non-mutagenicity (93 rules).

iii. Mutagenicity (Ames Test) model (ISS) 1.0.2 provides a qualitative prediction of mutagenicity on Salmonella typhimurium (Ames test). The model has been built as a set of rules, taken from the work of Benigni and Bossa (ISS) as implemented in the software ToxTree version 2.6 (`http://toxtree.sourceforge.net`). The training set for the model has been extracted from ToxTree, and consists of 670 compounds.

iv. Mutagenicity (Ames Test) model (KNN/Read-Across) 1.0.0 performs a read-across and provides a qualitative prediction of mutagenicity on Salmonella typhimurium (Ames test). The model performs a read-across on a dataset of 5770 chemicals.

(b) Carcinogenicity

i. The carcinogenicity CAESAR model (v. 2.1.9) is a CP-ANN neural network developed using data for carcinogenicity in rat extracted from the CPDB database.

ii. The Carcinogenicity (ISS) model (v. 1.0.2) was built implementing the same alerts (55SA in total, of which 22 for non-genotoxic carcinogenesis) used for Carcinogenicity (genotoxic and non-genotoxic) and mutagenicity ToxTree 2.6.13 ISS rule base (see below the ToxTree description for more details).

iii. The Carcinogenicity IRFMN/ANTARES model (v. 1.0.0) was built as a set of rules (127 structural alerts), extracted with the SARpy software from a dataset of 1543 chemicals obtained from the carcinogenicity database of EU-funded project ANTARES. This database is a collection of chemical rat carcinogenesis data (presence of carcinogenic effects in male or female rats) obtained from the CAESAR project database and the "FDA 2009 SAR Carcinogenicity - SAR Structures" database.

iv. The Carcinogenicity IRFMN/ISSCAN-CGX model (v. 1.0.0) is based on a set of rules (43 structural alerts), extracted with the SARpy software from a dataset of 986 compounds. This dataset was obtained from the combination of two data sets: i) The long-term carcinogenicity bioassay on rodents (rat and mouse) ISSCAN data set; ii) The carcinogenicity (different species) data set provided by Kirkland (**Kirkland2005**).

(c) Developmental Toxicity

i. The Developmental Toxicity CAESAR model (v. 2.1.7) is a QSAR binary classification model based on a Random Forest method, implemented using 13 descriptors developed using a dataset of compounds classified according to FDA categories for pregnancy (A and B categories are considered as non-toxicant, categories C, D and X are considered toxicant) .

ii. The Developmental/Reproductive Toxicity library (PG) model (v. 1.0.0) implements a virtual library of toxicant compounds as described in the study from P&G (Wu et al., 2013). The model identifies the category in which the given compound falls and generates a list of virtual compounds for each category. The model implements these categories, and tries to find an exact match between the given compound and any of the virtual compounds in the library.

2. Models with ecological interest: We used models able to assess the acute toxicity (LC50) on two of the indicator organisms chosen into the ecotoxicological regulation framework such as Daphnia Magna and Fish. For each endpoint two models are available. Each model is described below.

(a) Daphnia Magna Acute Toxicity (LC50)

i. Daphnia Magna LC50 48h (EPA) 1.0.7 provides a quantitative prediction for Daphnia Magna LC50 (48 hour), given in -log(mol/l) and its conversion in mg/L.

The model is a re-implementation of one of the individual models available inside T.E.S.T. software developed by Todd Martin for US EPA (single model). The model is a linear regression made on 17 molecular descriptors.

ii. Daphnia Magna LC50 48h (Demetra) 1.0.4 provides a quantitative prediction for Daphnia Magna LC50 (48 hour), given in -log(mol/l) and its conversion in mg/L. The model is a re-implementation of the original model developed inside the DEMETRA EU project (**Benfenati2011**). The model is a hybrid model base on multiple linear regressions, using on 16 molecular descriptors.

(b) Fish Acute Toxicity (LD50)

i. Fish Acute Toxicity Read-Across version 1.0.0 performs a read-across on a dataset of 972 chemicals and provides a quantitative prediction of acute toxicity in fish, given in -log(mg/L). This dataset has been made by Istituto di Ricerche Farmacologiche Mario Negri, merging experimental data from several reliable sources: the database compiled by the MED-Duluth group, the OECD Toolbox, the DEMETRA Project (Rainbow Trout toxicity model) and the work of Su et al. (**SuLMLiuX2014**). The read-across model has been built with the istKNN application (developed by Kode srl, `http://chm.kode-solutions.net`) and it is based on the similarity index developed inside the VEGA platform.

ii. 96 h Fathead Minnow (LC50) Model version 1.0.7 provides a quantitative prediction for Pimephales promelas LC50 (96hour), given in -log(mol/l) and its conversion in mg/L. It is a linear regression made on 21 molecular descriptors and is re-implementation of the original model developed by Todd Martin inside TEST software for US EPA. The TEST software is freely available at: `http://www.epa.gov/nrmrl/std/cppb/qsar/`.

3. Models with environmental interest: We used models able to evaluate the substance environmental fate such as Biodegradability and Persistence. The properties predicted by models were Ready Biodegradability by one model, Bioconcentration Factor (BCF) by three models and Persistence in all the three compartments, one model for each water, sediment and soil, considered by regulations. Each model is described below.

(a) Ready Biodegradability model (IRFMN) 1.0.9 is based on the OECD TG 301C - modified MITI -I test data and provides a qualitative evaluation (binary classification) of ready biodegradability properties. It has been developed using Sarpy software, by Istituto di Ricerche Farmacologiche Mario Negri and Politecnico di Milano. The model has been built as a set of rules, extracted from the training set with Sarpy software and manually by human experts. The final set of fragments obtained come from a work that involved both a statistical and an expert-based part.

(b) BCF CAESAR Model version 2.1.14 provides a quantitative prediction of bioconcentration factor (BCF) in fish, given in log(L/kg). The model extends the original CAESAR model, freely available at: `http://www.caesarproject.eu/software/`.

(c) BCF Meylan Model version 1.0.3 provides a quantitative prediction of bioconcentration factor (BCF) in fish, given in log(L/kg). The model implements the Meylan model, as described in EPI Suite BCFBAF module: `http://www.epa.gov/oppt/exposure/pubs/episuite.htm`.

(d) BCF Read-Across version 1.1.0 performs a read-across on a dataset of 860 chemicals and provides a quantitative prediction of bioconcentration factor(BCF) in fish, given in log(L/kg). This dataset has been made by Istitutodi Ricerche Farmacologiche Mario Negri, merging experimental data from several reliable sources, including the original dataset of the CAESAR BCF model (note that experimental values may differ from the ones in the CAESAR BCF dataset, as this new dataset has been built including more sources). The read-across model has been built with the istKNN application (developed by Kode srl, `http://chm.kode-solutions.net`) and it is based on the similarity index developed inside the VEGA platform.

(e) Persistence in water Model version 1.0.0 is based on the half-lives test data and provides a qualitative evaluation (four classes) of persistence property in the water compartment. It has been developed using an ensemble of k-NN modelling and a set of alerts extracted with Sarpy software, by Istituto di Ricerche Farmacologiche Mario Negri by human experts with the support of the istChemFeat application (developedby Kode srl, `http://chm.kode-solutions.net`), both developed on a dataset of 351 compounds (**Manga2015**).

(f) Persistence in soil Model version 1.0.0 is based on the half-lives test data and provides a qualitative evaluation (four classes) of persistence property in the water compartment. It has been developed using an ensemble of k-NN modelling and a set of alerts extracted with Sarpy software, by Istituto di Ricerche Farmacologiche Mario Negri by human experts with the support of the istChemFeat application (developedby Kode srl, `http://chm.kode-solutions.net`), both developed on a dataset of 297 compounds. [A. Manganaro, F. Pizzo, A. Lombardo, A. Pogliaghi, E. Benfenati, "Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN)algorithm", Chemosphere (2015)]

(g) Persistence in sediment Model version 1.0.0 is based on the half-lives test data and provides a qualitative evaluation (four classes) of persistence property in the water compartment. It has been developed using an ensemble of k-NN modelling and a set of alerts extracted with SARpy software, by Istituto di Ricerche Farmacologiche Mario Negri by human experts with the support of the istChemFeat application (developedby Kode srl, `http://chm.kode-solutions.net`), both developed on a dataset of 568 compounds (**Manga2015**).

Prioritization algorithm.

We then developed an algorithm for the prioritization of each single substance in the three biosphere levels:

1. Human health toxicity:

   (a) For carcinogenicity, predictions of two models were considered: Carcinogenicity CAESAR model (v. 2.1.9 and Carcino IRFMN/ANTARES model (v. 1.0.0). Only highly reliable predictions (defined by ADI index higher than 0.75) were considered. A consensus approach combines linearly the predictions from the two models (**Cassano2014**). A single substance is considered carcinogenic if the result of the consensus formula is higher than 0.

   (b) For developmental toxicity, we used a similar approach by combining the results of two models (DevTox CAESAR model v. 2.1.7 and Dev/ReproTox (PG) library model v. 1.0.0) into a single consensus model (**Cassano2014**).

   (c) For mutagenicity we developed a consensus formula by combining the predictions from 4 models (**Cassano2014**).

2. Ecological toxicity: We applied four different models for Ecotoxicity: Daphnia Magna LC50 48h (Demetra) 1.0.4, Daphnia Magna LC50 48h (EPA) 1.0.7, Fish Acute Toxicity Read-Across version 1.0.0, 96 h Fathead Minnow (LC50) Model version 1.0.7. Only substances into the applicability domain (defined by ADI index higher than 0.75) were

considered for prioritization; in a separate run, all the substances partially into the AD (defined by ADI index higher than 0.5) were also included.
If a substance shows an alert in at least 1 of the 4 domains, it is considered toxic in the ecological sphere.

3. Environmental toxicity: We applied five different models for Environmental toxicity: Ready Biodegradability model (IRFMN) 1.0.9; BCF CAESAR Model version 2.1.14; BCF Read-Across version 1.1.0 ; Persistence in water Model version 1.0.0; Persistence in soil Model version 1.0.0 and Persistence in sediment Model version 1.0.0. Only substances into the applicability domain (defined by ADI index higher than 0.75) were considered for prioritization; in a separate run, all the substances partially into the AD (defined by ADI index higher than 0.5) were also included.
If a substance shows an alert in at least 1 of the 5 domains, it is considered toxic in the environmental sphere. Only if a substance is toxic for one of the persistence compartments (water, soil and sediment), it is considered as persistent.

Acknowledgments

This study was supported by EC project PROSIL LIFE12/ENV/IT/000154. Special thanks to Carlo Podda (CRS4) for his continuous technical assistance and to Alberto Manganaro for providing custom VEGA code.

References

A. Manganaro F. Pizzo, A Lombardo A Pogliaghi E Benfenati (2015). “Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN)algorithm”. In: *Chemosphere*.

Ames BN Lee FD, Durston W E et al. (1973). “An improved bacterial test system for the detection and classification of mutagens and carcinogens”. In: *PNAS* 70.3, pp. 782–6. DOI: 10.1073/pnas.70.3.782. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=433358%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Benfenati E Boriani E, Craciun M Malazizi L Neagu D Roncagioni A (2011). *Databases for pesticide ecotoxicity in: Benfenati E (Elsevier) Quantitative structure-activity relationships (QSAR) for pesticide regulatory purposes*.

VEGA-QSAR: AI inside a platform for predictive toxicology (2013). CEUR Workshop Proceedings Vol-1107. URL: http://www.vega-qsar.eu/.

Boriani E Ernstoff A, Benfenati E Nicolaas E (2016). “Classification and prioritization of chemicals present in food contact material”. In: *ILSI conference, Barcelona*.

Kirkland D Aardema M, Henderson L Müller L et al. (2005). “Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity”. In: *Mutat Res* 608.1, pp. 29–42. ISSN: 0027-5107 (Print). DOI: 10.1016/j.mrgentox.2006.04.017.

OBoyle NM Banck M, James C A Morley C Vandermeersch T Hutchison G R (2011). “Open Babel: An open chemical toolbox”. In: *J Cheminform.* DOI: 10.1186/1758-2946-3-33.

Rousseau, Jean-Jacques (1762). *Emile ou De l’education*. Garnier.

Su LM Liu X, Wang Y Li J J Wang X H Sheng L X Zhao Y H (2014). “The discrimination of excess toxicity from baseline effect: effect of bioconcentration”. In: *Sci Total Environ.* DOI: 10.1016/j.scitotenv.2014.03.040.

Wishart DS Jewison T, Guo A C Wilson M Knox C et al. (2013). “HMDB 3.0 — The Human Metabolome Database in 2013”. In: *Nucleic Acids Research*.

Wishart DS Knox C, Guo A C Shrivastava S Hassanali M Stothard P Chang Z Woolsey J (2014). "Evaluation of QSAR models for the prediction of ames genotoxicity: a retrospective exercise on the chemical substances registered under the EU REACH regulation". In: *J Environ Sci Health C Environ Carcinog Ecotoxicol*. DOI: 10.1080/10590501.2014.938955.

Wishart DS Knox C, Guo A C Shrivastava S Hassanali M Stothard P Chang Z Woolsey J et al. (2006). "DrugBank: a comprehensive resource for in silico drug discovery and exploration". In: *Nucleic Acids Research* 34.Database issue, pp. D668–72. URL: http://www.ncbi.nlm.nih.gov/pubmed/16381955$%5Cbackslash$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1347430.

Wishart DS Knox C, Guo A C et al. (2009). "HMDB: a knowledgebase for the human metabolome". In: *Nucleic Acids Research*.

Wishart DS Tzur D, Knox C Eisner R Guo A C Young N Cheng D Jewell K Arndt D Sawhney S Fung C Nikolai L Lewis M Coutouly M A Forsythe I Tang P Shrivastava S Jeroncic K Stothard P Amegbey G Block D Hau D D Wagner J Miniaci J Clements M Gebremedhin M Guo N Zhang Y Duggan G E M (2007). "HMDB: the Human Metabolome Database". In: *Nucleic Acids Research*.