QSAR models and their practical use, the experience of Kode

Alberto Manganaro (a.manganaro@kode.srl)

Genova, June 20th 2019

Kode Chemoinformatics

We are a team of experts in the field of chemoinformatics and QSAR modelling, with over a decade of experience and involvement in research activities and projects.

We provide high-value integrated consultancy and novel IT tools to support activities for industry (pharmaceutical and chemical) and research institutions.

We offer:
- Chemical handling solutions: integrated databases and tool (structure similarity search, gathering of specific datasets, etc)
- QSAR modelling: molecular descriptors calculation, robust modelling tools
- Property predictions: application and customization of existing models for eco-toxicological predictions

Development of the open-source tool **VEGA**: several QSAR models for eco-toxicological endpoints (financed by the European Union).

Development of the open-source tool **ToxRead**: platform for read-across assessment of chemicals (financed by the European Union).

Development of the freely available tools **Prometheus** and **Janus**, for the German environmental agency (UBA): prioritization of PBT substances with QSAR models batteries. Currently working on a new project for the UBA, **toDivine**: platform for assessment of substances based on the integration of QSAR and read-across.

Partner in european research projects:
- **LIFE Vermeer**
- **LIFE Concert REACH**

QSAR and in-silico tools

REACH legal framework led to an increasing request of QSARs and in-silico tools both from industries and regulatory bodies.

The need of data to perform the REACH registration of chemicals pushes (small and medium) industries towards the use of **in-silico tools** (for economic reasons). Due to this, several EU funded projects have the goal to make available QSARs and other non-testing methods for eco-toxicological assessment.

REACH contains a number of specific measures and general provisions designed to establish and enforce the principle that **animal testing should be performed only as a last resort**. On the side of industries, animal testing is expensive and in particular SMEs can not just afford it.

The objective is to deliver flexible and user-friendly software tools, called SPHERA and ToxEraser, for the **substitution of harmful chemicals**. SPHERA will be a diagnostic tool for identifying the adverse effects related to chemicals, and ToxEraser will provide the remedy by suggesting suitable alternative chemicals.

These tools will enable multiple assessments for human health or other species, and the persistence or bioaccumulation of chemicals taking into account different exposure scenarios. The tools will be based on **MERLIN-Expo** for exposure, on **VEGA** for assess multi-endpoint hazard, on **ERICA** for the risk aggregation and on**ToxRead** for the identification of structural alerts in the substitution strategy. The tools will have broad application, which will be demonstrated through a series of case studies (food contact materials, biocides, petroleum and oil fraction, greener solvents, dispersants and cosmetics).

PARTNERS:
Istituto di Ricerche Farmacologiche Mario Negri
Electricité de France SA
Kode s.r.l.
Institut Scientifique de Santé Publique
SC Sviluppo chimica S.p.A
Federal Institute for Risk Assessment
Angel Consulting SAS
Institut National de l'Environnement industriel et des Risques
Facilia AB, Sweden

LIFE CONCERT REACH (LIFE17 GIE/IT/000461)

Aim of the project is to establish a **network of in silico tools** aimed to reshape the strategy of evaluating the chemical substances, becoming, in the long term, the first step in the evaluation of the chemical substances, to be followed by the classical methods.

This goal will be pursued by the integration of currently well-known and established solutions such as **VEGA**, **Ambit** and **Danish QSAR Database**. New (Q)SAR and read-across tools, potentially covering all possible properties relevant for the substance REACH registration, will be also developed. The integrated models will be developed and updated to be compliant with regulatory requirements, paying also particular attention on the generated prediction report, with the aim of facilitating their evaluation and, in the end, improving their usability.

PARTNERS:
Istituto di Ricerche Farmacologiche Mario Negri
Technical University of Denmark
IdeaConsult
Dr. Knoell Consult GmbH
Kode s.r.l.
SC Sviluppo chimica S.p.A

VEGA is an open-source freely available tool, that makes available a **wide range of QSAR models**.

All implemented QSAR models come from other scientific projects or published papers, and newly developed models mainly by Istituto di Ricerche Farmacologiche "Mario Negri" and Kode.

- Easy-to-use interface, designed for users (toxicologists, chemists) who are not QSAR experts
- Rich documentation on the produced predictions, to support the use of the results for the REACH registration of chemicals
- Recognized as a valid tool by ECHA

VEGA is freely available at: https://www.vegahub.eu/

List of all available QSAR models in the current version of VEGA (1.1.5):

Mutagenicity (Ames test) CONSENSUS model
Mutagenicity (Ames test) model (CAESAR)
Mutagenicity (Ames test) model (SarPy/IRFMN)
Mutagenicity (Ames test) model (ISS)
Mutagenicity (Ames test) model (KNN/Read-Across)
Carcinogenicity model (CAESAR)
Carcinogenicity model (ISS)
Carcinogenicity model (IRFMN/Antares)
Carcinogenicity model (IRFMN/ISSCAN-CGX)
Carcinogenicity oral classification model (IRFMN)
Carcinogenicity oral Slope Factor model (IRFMN)
Carcinogenicity inhalation classification model (IRFMN)
Carcinogenicity inhalation Slope Factor model (IRFMN)
Developmental Toxicity model (CAESAR)
Developmental/Reproductive Toxicity library (PG)
Zebrafish embryo AC50 (IRFMN/CORAL)
Estrogen Receptor Relative Binding Affinity model (IRFMN)
Estrogen Receptor-mediated effect (IRFMN/CERAPP)
Androgen Receptor-mediated effect (IRFMN/COMPARA)
Skin Sensitization model (CAESAR)
Skin Sensitization model (IRFMN/JRC)
Hepatotoxicity model (IRFMN)
Fish Acute (LC50) Toxicity classification (SarPy/IRFMN)
Fish Acute (LC50) Toxicity model (KNN/Read-Across)
Fish Acute (LC50) Toxicity model (NIC)
Fish Acute (LC50) Toxicity model (IRFMN)
Fish Acute (LC50) Toxicity model (IRFMN/Combase)
Fish Chronic (NOEC) Toxicity model (IRFMN)
Fathead Minnow LC50 96h (EPA)
Fathead Minnow LC50 model (KNN/IRFMN)

Daphnia Magna LC50 48h (EPA)
Daphnia Magna LC50 48h (DEMETRA)
Daphnia Magna Acute (EC50) Toxicity model (IRFMN)
Daphnia Magna Acute (EC50) Toxicity model (IRFMN/Combase)
Daphnia Magna Chronic (NOEC) Toxicity model (IRFMN)
Guppy LC50 model (KNN/IRFMN)
Algae Acute (EC50) Toxicity model (IRFMN)
Algae Acute (EC50) Toxicity model (ProtoQSAR/Combase)
Algae Chronic (NOEC) Toxicity model (IRFMN)
Algae Classification Toxicity model (ProtoQSAR/Combase)
Bee acute toxicity model (KNN/IRFMN)
Sludge (EC50) Toxicity model (ProtoQSAR/Combase)
Sludge Classification Toxicity model (ProtoQSAR/Combase)
BCF model (CAESAR)
BCF model (Meylan)
BCF model (KNN/Read-Across)
BCF model (Arnot-Gobas)
kM/Half-Life model (Arnot/EpiSuite)
Ready Biodegradability model (IRFMN)
Persistence (sediment) model (IRFMN)
Persistence (sediment) quantitative model (IRFMN)
Persistence (soil) model (IRFMN)
Persistence (soil) quantitative model (IRFMN)
Persistence (water) model (IRFMN)
Persistence (water) quantitative model (IRFMN)
LogP model (Meylan/Kowwin)
LogP model (MLogP)
LogP model (ALogP)
Water solubility model (IRFMN)

VEGA

VEGA is distributed as a stand-alone Java application.

It does not rely on internet (server) connection, indeed processed data are not shared with any third party.
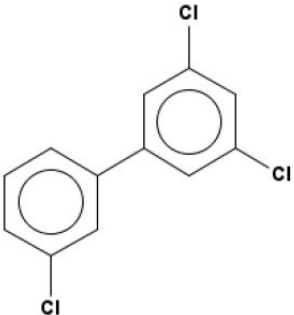
VEGA
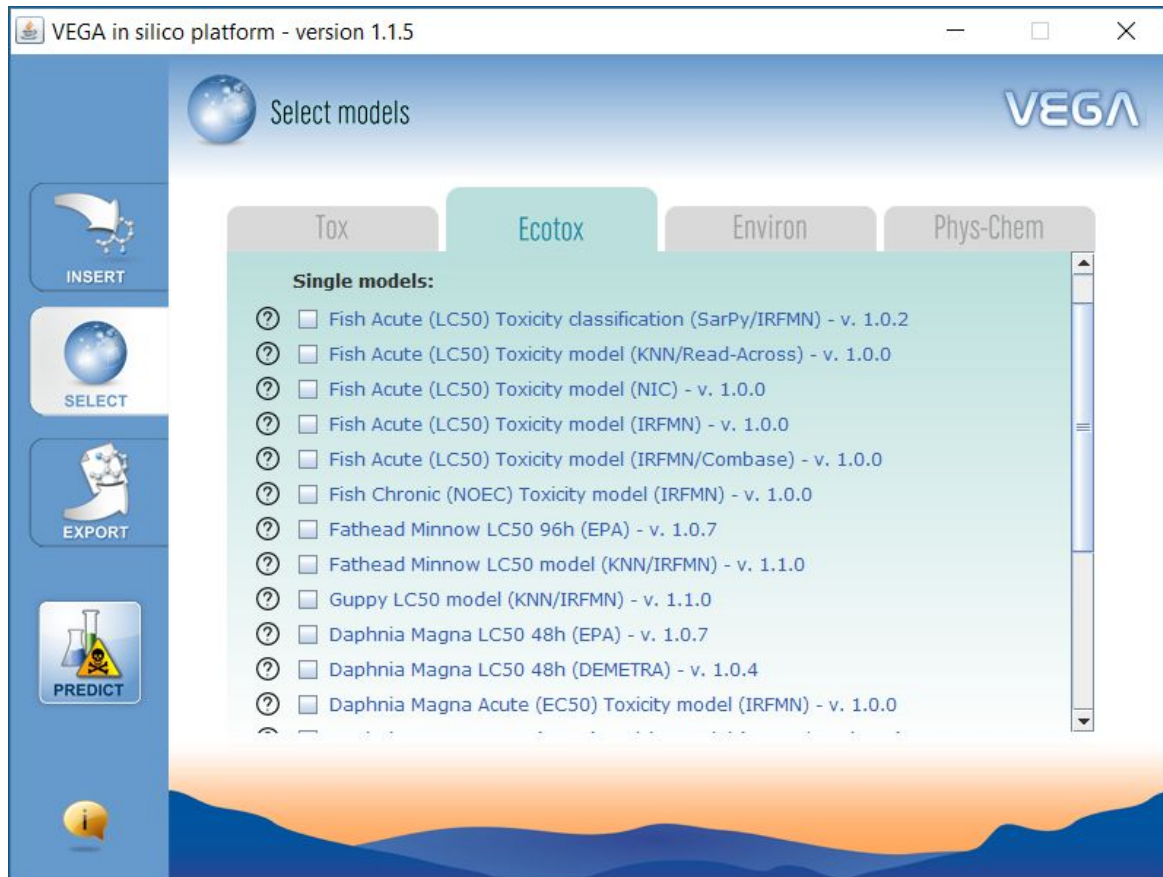
The PDF report provides full information about the processed molecule and the details (reliability, applicability domain) of the prediction.

1. Prediction Summary

Prediction for compound Molecule 1

Prediction: 🔴   Reliability: ⭐⭐⭐

Prediction is 0.3853 mg/L, the result appears reliable. Anyhow, you should check it through the evaluation of the information given in the following sections.

Compound: Molecule 1
Compound SMILES: c1cc(cc(c1)Cl)c2cc(cc(c2)Cl)Cl
Experimental value [log(1/(mmol/L))]: -
Predicted fish toxicity [log(1/(mmol/L))]: 2.82
Predicted fish toxicity [mg/l]: 0.3853
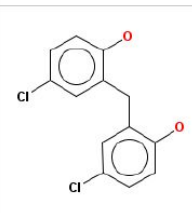Molecular Weight: 257.16
Experimental value [mg/l]: -
Reliability: the predicted compound is into the Applicability Domain of the model

VEGA

The PDF report includes also the list of most similar compounds with known experimental value.

Indeed, this helps the user to assess the provided prediction (with a read-across-like approach) so to accept or interpret the given result.

3.1 Applicability Domain:
Similar Compounds, with Predicted and Experimental Values

Compound #1

CAS: 000091-94-1
Dataset id: 868 (Training set)
SMILES: Nc(c(cc(c(ccc(N)c1Cl)c1)c2)Cl)c2
Similarity: 0.884

Experimental value [log(1/(mmol/L))]: 2.696
Predicted value [log(1/(mmol/L))]: 3.008

Compound #2

CAS: 006639-30-1
Dataset id: 426 (Training set)
SMILES: c(c(cc(c1Cl)Cl)Cl)(c1)C
Similarity: 0.828

Experimental value [log(1/(mmol/L))]: 2.06
Predicted value [log(1/(mmol/L))]: 1.852

Compound #3

CAS: 000097-23-4
Dataset id: 824 (Test set)
SMILES: Oc(c(cc(c1)Cl)Cc(c(O)ccc2Cl)c2)c1
Similarity: 0.805

Experimental value [log(1/(mmol/L))]: 2.94
Predicted value [log(1/(mmol/L))]: 2.316

VEGA

Details about the applicability domain of the given prediction (indeed, about its reliability) are provided.

3.2 Applicability Domain:
Measured Applicability Domain Scores

Global AD Index
AD index = 1
Explanation: the predicted compound is into the Applicability Domain of the model.

Similar molecules with known experimental value
Similarity index = 0.853
Explanation: strongly similar compounds with known experimental value in the training set have been found.

Accuracy of prediction for similar molecules
Accuracy index = 0.26
Explanation: accuracy of prediction for similar molecules found in the training set is good.

Concordance for similar molecules
Concordance index = 0.446
Explanation: similar molecules found in the training set have experimental values that agree with the predicted value.

Maximum error of prediction among similar molecules
Max error index = 0.312
Explanation: the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability.

Model's descriptors range check
Descriptors range check = True
Explanation: descriptors for this compound have values inside the descriptor range of the compounds of the training set.

Atom Centered Fragments similarity check
ACF index = 1
Explanation: all atom centered fragment of the compound have been found in the compounds of the training set.

VEGA

ECHA (the European CHemical Agency) mentioned VEGA in one of its practical guide, as a valid tool to be used in REACH registrations.

4.2 Ready biodegradability (VEGA)

a) Introduction

Ready biodegradability is a REACH requirement for all substances produced or imported above one tonne/year (REACH Annex VII). The main outcome of a ready biodegradability test is the classification of the chemical either as "readily biodegradable" or as "not readily biodegradable".

The VEGA platform contains several QSAR models for various endpoints. One of these models predicts ready biodegradability (model developed by the Istituto di Ricerche Farmacologiche Mario Negri). This model is based on structural alerts.

Four sets of substructures (i.e. fragments) are included in this model and these sets are classified as "non-re... biodegradable" and ... non-biodegradable i...

At the time of writin... used to prepare this...

Link to the (Q)SAR p...

ECHA
EUROPEAN CHEMICALS AGENCY

Practical guide

How to use and report (Q)SARs

Version 3.1 – July 2016

ToxRead

ToxRead is an open-source freely available tool, that assists the user in making reproducible read across evaluations of chemicals.

Given a target chemical, the software searches for similar compounds with known experimental value, and arrange the found compounds using a set of rules and structural alerts that provide useful information to estimate the target compound.

ToxRead is freely available at: https://www.vegahub.eu/

ToxRead

ToxRead currently provides six end-points, and it will be soon updated with new end-points developed in different ongoing projects.

PROMETHEUS / JANUS
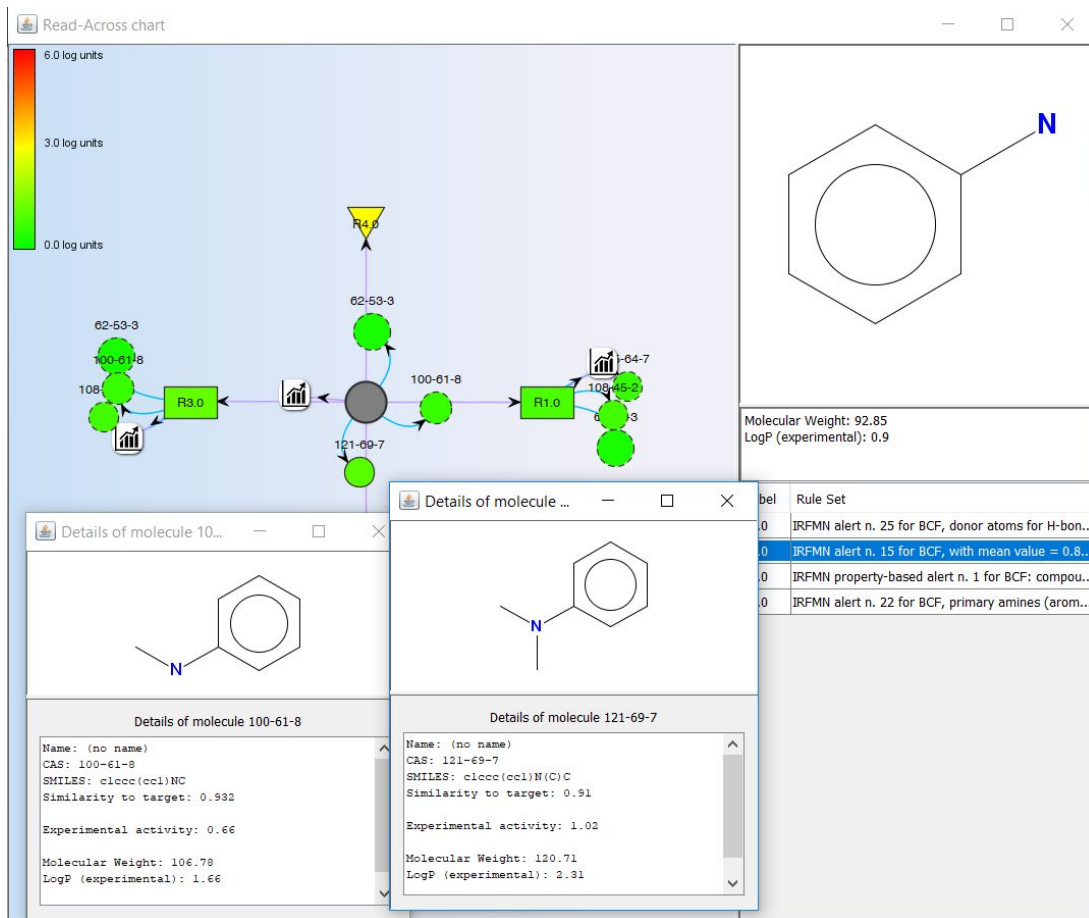
We worked on two projects for the **UBA (German environmental agency)** to develop a platform for the PBT assessment of chemicals. This is needed by the UBA to screen large sets of chemicals and prioritize them, reducing the number of needed tests.

The software is based on a battery of QSAR models, integrated with a specific workflow for each end-point. The final predictions are combined in a unique score, that allows to rank and prioritize the list of target compounds.

The initial project **Prometheus** led to a first version of the software, enhanced by the following project **Janus** (the final application will be publicly available in a few months)

PROMETHEUS / JANUS

Janus Result

NUMBER OF COMPOUNDS **3**

☑ PBT ☐ CMR ☐ ED ☑ PARTIAL SCORES ☑ FINAL SCORES

	No. ∨	Metabolite	Id	SMILES	Label	P	rel.	score	B [log(L/kg)]	rel.	score	T [mg/l]	rel.	score	Score(vPvB)	Score(SVHC)	Score(PBT)
🔍	1		Molecule 1	OC(OC)C	**PBT-CMRE**	nP	0.45	0.211	**0.16**	0.9	0.078	**2.18**	0.57	0.28	0.129	0.853	0.242
🔍	2	Metabolite	Molecule 1 [M-01]	O=CC	**PBT-CMRE**	nP	0.5	0.243	**0.31**	0.5	0.191	**1.08**	0.62	0.313	0.215	0.523	0.291
🔍	3	Metabolite	Molecule 1 [M-02]	OC	**PBT-CMRE**	nP	0.8	0.174	**0.15**	0.75	0.114	**0.78**	0.62	0.336	0.141	0.875	0.26

Export as plain text

PROMETHEUS / JANUS

Molecule

Info

Molecule number: 1
Molecule id: Molecule 1
Molecule SMILES: OC(OC)C
Metabolites:
    Molecule number 2: go to molecule ❯
    Molecule number 3: go to molecule ❯

Note

SUMMARY

Persistence	Bioaccumulation	Toxicity
nP (reliability: 0.45)	**0.16** (reliability: 0.9)	**2.18** (reliability: 0.57)

Carcinogenicity	Mutagenicity	Reproductive toxicity
NON Carcinogenic (reliability: 0.8)	**NON Mutagenic** (reliability: 0.8)	**TOXICANT** (reliability: 0.9)

Endocrine Disruptor
Inactive (reliability: 0.5)

Conclusions

- There is an emerging need and request for **in-silico tools**

- Several publicly funded projects support their **development** and **use** in regulatory context

- QSARs and other in-silico tools are meeting the favor of **industry** and **regulatory** bodies

- In our experience, we aimed to develop tools **designed for the final user** (and not for the QSAR expert), so to support the **real use and acceptance** of non-testing methods.

Thanks for your attention and thanks to