QSARINS (QSAR-INSubria)

Software for QSAR MLR model development and validation

From ideas of Prof. Paola Gramatica, applied in "QSAR Research Unit in Environmental Chemistry and Ecotoxicology" at University of Insubria, Italy

The approach to validated development and validation of QSAR MLR models in QSARINS is presented in the following papers:

- Gramatica, P., Principles of QSAR modeling. Comments and suggestions from personal experience, *International Journal of QSPR*, **2020**, 5(3) ,61-97 **Open Access** DOI: 10.4018/IJQSPR.20200701.oa1
- Gramatica, P., Chirico, N., Papa, E., Kovarich, S., Cassani, S. QSARINS: A New Software for the Development, Analysis, and Validation of QSAR MLR Models. *Journal of Computational Chemistry*, *Software news and updates*, **2013**, 34, 2121-2132, DOI: 10.1002/jcc.23361.

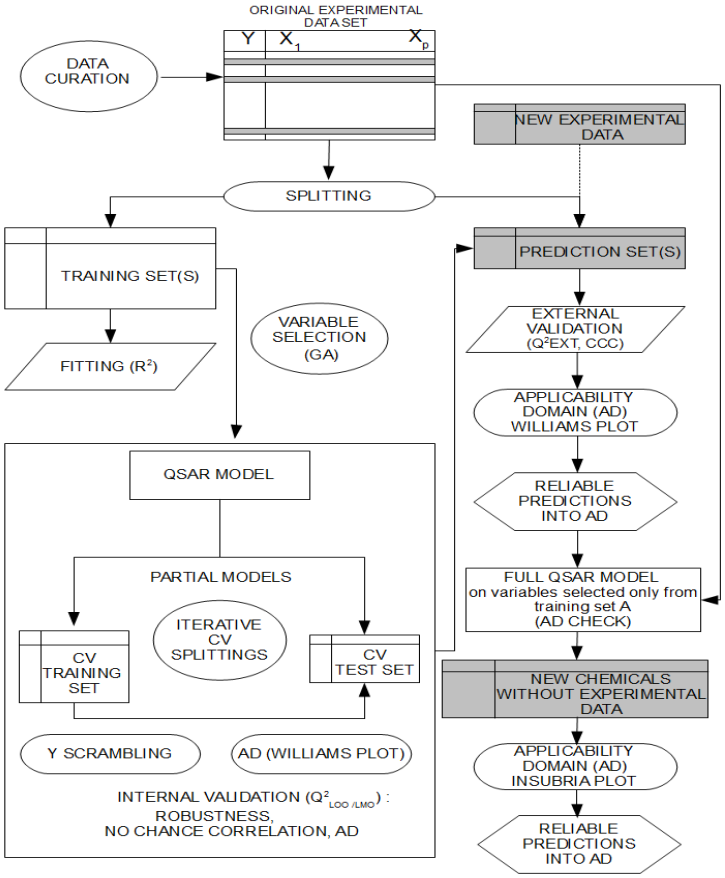
- Gramatica, P., Cassani, S., Chirico, N. QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS. *Journal of Computational Chemistry*, *Software news and updates*, **2014,** 35, 1036–1044. DOI: 10.1002/jcc.23576

The researchers interested in QSARINS (whether for QSAR models' development/validation or for application of QSARINS-Chem models or for use of the stored Insubria datasets or for analysis and validation of personal models) should read these papers to better know QSARINS.
These papers should also be kindly cited in papers where QSARINS is used.

QSARINS can be obtained by sending a request to Prof. Paola Gramatica (paola.gramatica@uninsubria.it), illustrating the personal experience in QSAR modeling.

General information about QSARINS

QSARINS (QSAR-INSubria) is a software for the development and validation of Multiple Linear Regression (MLR) models by Ordinary Least Squares (OLS) and Genetic Algorithm (GA) for variable selection, based on QSAR experience of Prof. Paola Gramatica since 1995 and developed by Nicola Chirico (2008-2012). It is implemented according to the statistical/chemometric approach for QSAR models' predictivity (Tropsha et al. 2003; Gramatica 2007, 2009, 2012, 2013, 2014, 2016, 2020; Chirico and Gramatica 2011, 2012; Gramatica et al. 2012; Gramatica and Sangion 2016), which is applied by the QSAR group of the University of Insubria in all the modeling researches.
In the module **QSARINS-Chem** the 3D- chemical structures of many Insubria datasets are stored, in addition to several QSAR models, based on PaDEL Descriptors (Yap 2011), developed by the Insubria QSAR group and applicable also to new chemicals, directly in QSARINS, where the PaDEL Descriptors can be calculated.
The following scheme (from Gramatica et al. 2012) summarizes the procedure for a validated QSAR modeling, applicable in QSARINS:



QSARINS provides a user-friendly interface that includes tools for:

- Data normalization;

- Dataset analysis (by Principal Component Analysis (PCA), etc.);

- Splitting of the data sets (by Random, Ordered response or Structure by PCA of molecular descriptors);

- OLS Model development by All Subset Selection;

- Genetic Algorithm (GA) for Variable Selection (with tracing of the models while being developed);

- Analysis of all individual variables, as well as those more frequently selected in the models;

- Internal validation, including the following criteria: Q^2_{LOO}, Q^2_{LMO}, Concordance Correlation Coefficient (CCC), RMSE, MAE, Y-Scrambling, etc…;

- External validation, including the following criteria: Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, Golbraikh and Tropsha parameters, r^2_m metrics, CCC_{EXT}, RMSE, MAE;

- Structural applicability domain by the leverage from the diagonal values of the Hat matrix

(including Williams plot for chemicals with experimental data, y axis: standardized residuals and Insubria graph for chemicals without data, y axis: predicted values);

- Principal Components (PCs) Regression;

- Multi-Criteria Decision Making (MCDM) for selection of the best models;

- Consensus modeling (weighted and not, different tools for selection of models to be average);

- Check and validation of imported single models;

-Calculation of molecular descriptors and fingerprints with the open source software PaDEL-Descriptor (Yap 2011) (version currently included: 2.21).

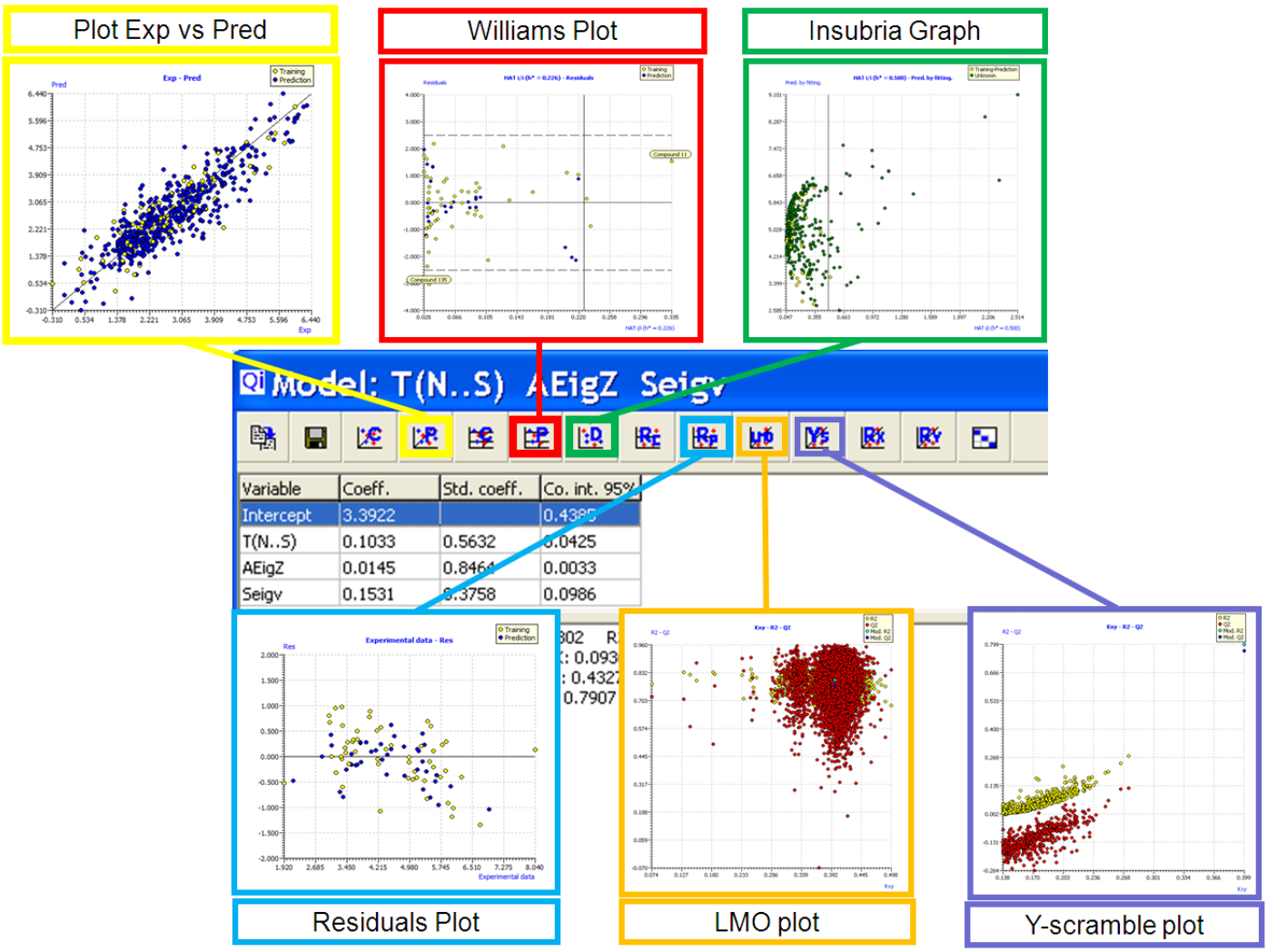
- Application of several QSAR/QSPR models for environmental pollutants, developed by PaDEL-Descriptor software and stored in the QSARINS-Chem module, with the corresponding QMRF (QSAR Model Reporting Format). In particular: the PBT Index model (Papa and Gramatica 2010), re-developed (Gramatica et al. 2013, 2015) using PaDEL-Descriptor (Yap 2011);

- Database of compounds used to develop Insubria QSAR models: chemicals can be explored in different ways (CAS, SMILES, names) and visualized in 3D;

- Ranking of chemicals, based on PCA and MCDM

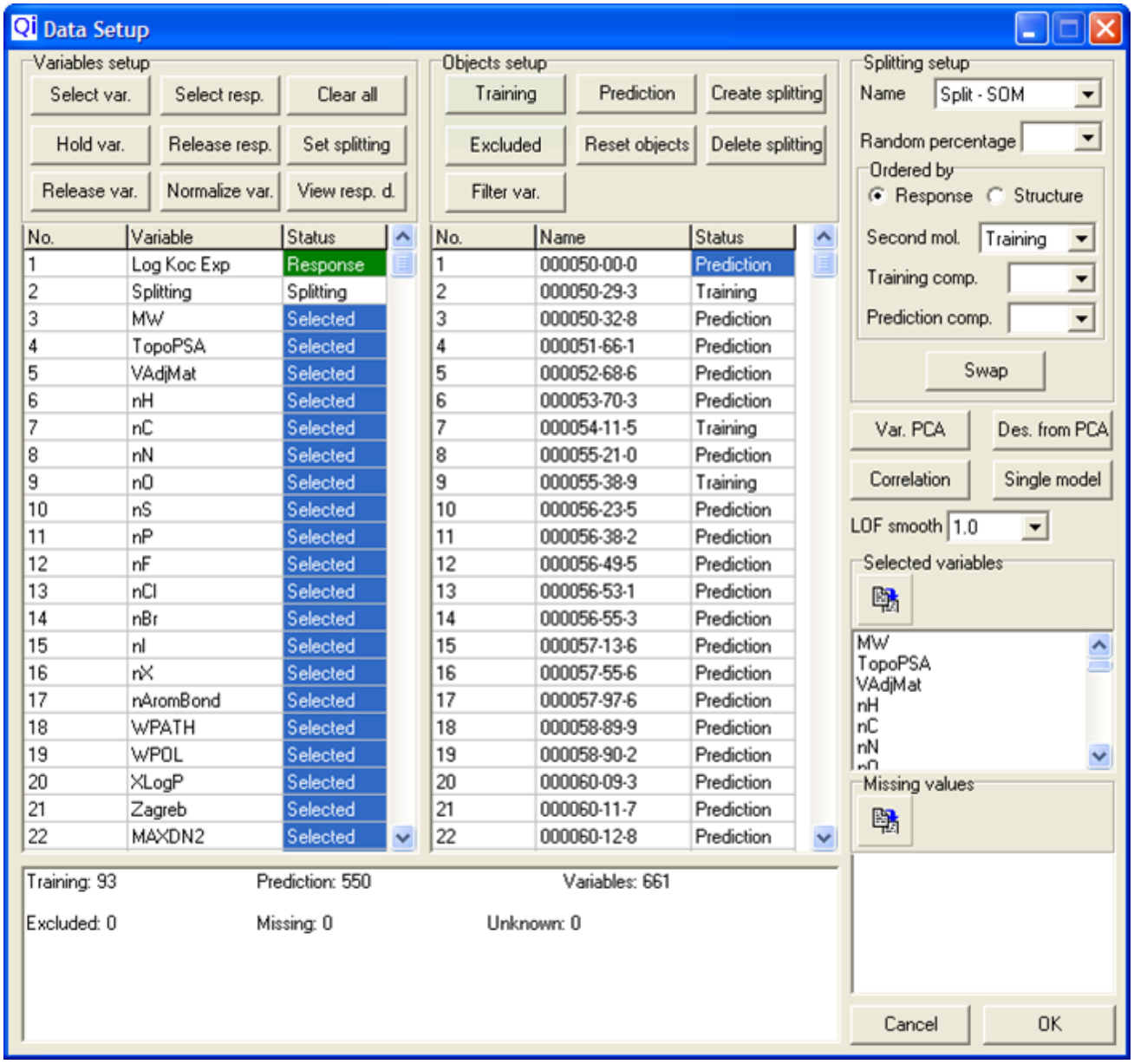
All outcomes of QSARINS are reported in the software with corresponding **plots** for quick graphical interpretation.

Example of some available **plots:**



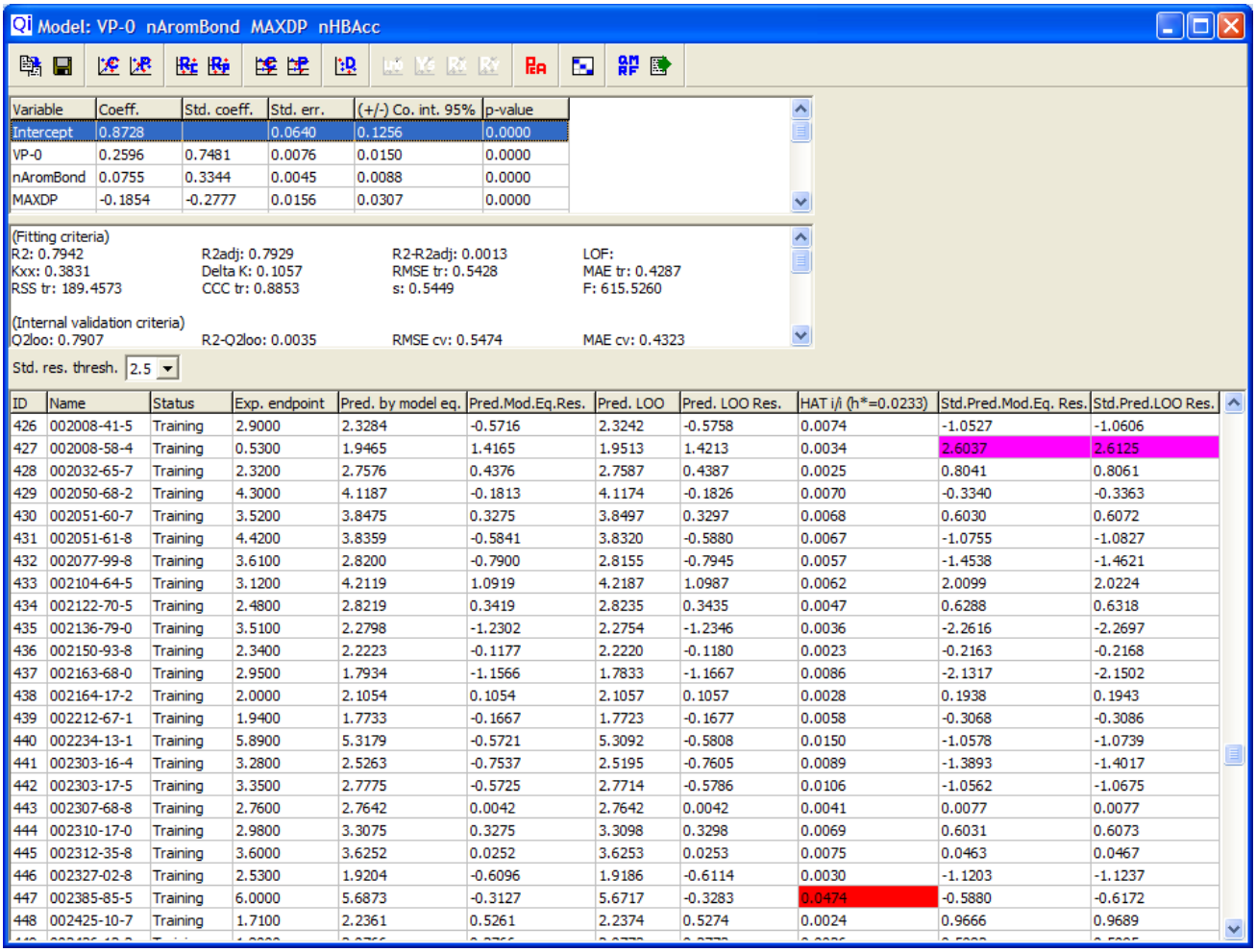
SCREENSHOTS

DATA SETUP - Selection of variables and response. Three different splitting (Random, by ordered response, by structure based on PCA of molecular descriptors). PCA of the selected descriptors and generation of PC scores as variables for Principal Components Regression.



SINGLE MODEL – Model's parameters related to internal and external validation, predicted values, HAT values, and standardized residuals are calculated.
In addition, it is possible to view the PCA and the correlation matrix of the modeling descriptors. If available, the QMRF can be exported. It is present for the large majority of the models available in QSARINS-Chem.

IMPORTANT INFORMATION: Any user can upload personal MLR models and use QSARINS to manage them for storing, visualization, validation, ranking etc..



QSARINS-Chem Module in QSARINS:

QSARINS-Chem is the module where 3012 chemicals, collected from the literature, curated and modeled by the Insubria group, are available with their 3D structure and experimental responses. In addition, 45 QSAR models of environmental end-points for organic pollutants, based on free software for molecular descriptors (Yap 2011), are available. These models, supported by their QMRF, can be applied for any new chemical without experimental data or also not yet synthesized, verifying the structural applicability domain by the Insubria graph.

List of QSAR models in QSARINS-Chem Module:

1. Brominated Flame Retardants (BFR) Log Koa (Papa et al. 2009; Gramatica et al. 2014)[$]
2. BFR MP (Papa et al. 2009; Gramatica et al. 2014)[$]
3. BFR VP (Papa et al. 2009; Gramatica et al. 2014)[$]
4. Benzo-Triazole (BTAZ) *D. magna* tox (Cassani et al. 2013; Gramatica et al. 2014)[$]
5. BTAZ Log Kow (Bhhatarai and Gramatica 2011a; Gramatica et al. 2014)[$]
6. BTAZ MP (Bhhatarai and Gramatica 2011a; Gramatica et al. 2014)[$]
7. BTAZ *O. mykiss* tox (Cassani et al. 2013; Gramatica et al. 2014)[$]
8. BTAZ *P. subcapitata* tox (Cassani et al. 2013; Gramatica et al. 2014)[$]
9. BTAZ Sw (Bhhatarai and Gramatica 2011a; Gramatica et al. 2014)[$]
10. BTAZ VP (Bhhatarai and Gramatica 2011a; Gramatica et al. 2014)[$]
11. Endocrine Disruptor Chemicals (EDC) Estrogen Receptor Binding (Li and Gramatica 2010; Gramatica et al. 2014)[$]
12. Esters *D. magna* tox (Papa et al. 2005a; Gramatica et al. 2014)[$]
13. *Esters P. promelas* tox (Papa et al. 2005a; Gramatica et al. 2014)[$]
14. Fish Biotrans. logHLn M1_day (Papa et al. 2014)*
15. Fish Biotrans. logHLn M2_day (Papa et al. 2014)*
16. Fish Biotrans. logHLn M3_day (Papa et al. 2014)*
17. Global Half-Life Index (GHLI) (Gramatica and Papa 2007; Gramatica et al. 2014)[$]*
18. Human Biotrans. logHLB1_h (Papa et al. 2018)*
19. Human Biotrans. logHLB2_h (Papa et al. 2018)*
20. Human Biotrans. logHLB3_h (Papa et al. 2018)*
21. Human Biotrans. logHLB4_h (Papa et al. 2018)*
22. Human Biotrans. logHLT_h (Papa et al. 2018)*
23. Log Koc of Pesticides (Gramatica et al. 2007a, 2014)[$]*
24. Nitro PAH Mutagenicity (Gramatica et al. 2007b, 2014)[$]
25. *Pimephales promelas* tox (Papa et al. 2005b; Gramatica et al. 2014)[$]*
26. Persistence Bioaccumulation Toxicity (PBT) Index (Papa and Gramatica 2010; Gramatica et al. 2013)[$]*
27. Personal Care Products (PCP) Aquatic Toxicity Index-ATI (Gramatica et al. 2016)*
28. PCP *D.magna* acute tox (Gramatica et al. 2016)*
29. PCP *P.promelas* acute tox LogP-based (Gramatica et al. 2016)*
30. PCP *P.promelas* acute tox (Gramatica et al. 2016)*

31. PCP *P.subcapitata* acute tox (Gramatica et al. 2016)*
32. PerFluorinated Chemicals (PFC) Mouse Inhalation tox (Bhhatarai and Gramatica 2010; Gramatica et al. 2014)[$]
33. PFC Rat Inhalation tox (Bhhatarai and Gramatica 2010; Gramatica et al. 2014)[$]
34. PFC Rat Oral tox (Bhhatarai and Gramatica 2011b; Gramatica et al. 2014)[$]
35. PFC Sw (Bhhatarai and Gramatica 2011c; Gramatica et al. 2014)[$]
36. PFC VP (Bhhatarai and Gramatica 2011c; Gramatica et al. 2014)[$]
37. Pharm. Aquatic Toxicity Index-ATI (Sangion and Gramatica 2016a)*
38. Pharm. *D.magna* acute tox (Sangion and Gramatica 2016a)*
39. Pharm. O.mykiss acute tox (Sangion and Gramatica 2016a)*
40. Pharm. *P.promelas* acute tox (Sangion and Gramatica 2016a)*
41. Pharm. *P.subcapitata* acute tox (Sangion and Gramatica 2016a)*
42. PPCP intertox *D.magna-O.mykiss* (Sangion and Gramatica 2016b)
43. PPCP intertox *D.magna-P.promelas* (Sangion and Gramatica 2016b)
44. PPCP intertox *O.mykiss-P.promelas* (Sangion and Gramatica 2016b)
45. PPCP intertox P.promelas-O.mykiss (Sangion and Gramatica 2016b)

[$]*Models originally developed using proprietary software (DRAGON v 5.5 or former versions) and updated using PaDEL Descriptor software (models are described in Gramatica et al. 2014).*

**Models reported also in QSARINS-Chem Standalone version (free downloadable from www.qsar.it)*

Additional information

QSARINS can be used for every modeling work involving Multiple Linear Regression (MLR) calculations, based on Genetic Algorithm for variable selection and Ordinary Least Squares (OLS) as modeling method. However, any personal MLR model, even if developed by other software, can be analyzed by the plots available in QSARINS as well as validated by several statistical validation parameters (both for internal and external validation: see above list).

Other chemometric tools (Principal Component Analysis (PCA), Multicriteria Decision Macking (MCDM)) for explorative analysis and ranking are also implemented, **therefore QSARINS is not limited to QSAR studies.**

It is also important to note that not only chemicals, but any kind of objects can be analyzed or modeled in a multivariate way.

Bibliography

Bhhatarai B, Gramatica P (2011a) Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. Water Res 45:1463–1471. doi: 10.1016/j.watres.2010.11.006

Bhhatarai B, Gramatica P (2010) Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat and Mouse Using QSAR Modeling. Chem Res Toxicol 23:528–539. doi: 10.1021/tx900252h

Bhhatarai B, Gramatica P (2011b) Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse. Mol Divers 15:467–476. doi: 10.1007/s11030-010-9268-z

Bhhatarai B, Gramatica P (2011c) Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. Environ Sci Technol 45:8120–8128. doi: 10.1021/es101181g

Cassani S, Kovarich S, Papa E, et al (2013) Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity–activity modelling. J Hazard Mater 258–259:50–60. doi: 10.1016/j.jhazmat.2013.04.025

Chirico N, Gramatica P (2011) Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. J Chem Inf Model 51:2320–2335. doi: 10.1021/ci200211n

Chirico N, Gramatica P (2012) Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection. J Chem Inf Model 52:2044–2058. doi: 10.1021/ci300084j

Gramatica P (2007) Principles of QSAR models validation: internal and external. Qsar Comb Sci 26:694–701. doi: 10.1002/qsar.200610151

Gramatica P (2009) Chemometric Methods and Theoretical Molecular Descriptors in Predictive QSAR Modeling of the Environmental Behaviour of Organic Pollutants, Chapter 12 in Recent Advances in QSAR Studies. In: Puzyn T, Leszczynski J, Cronin MT (eds) Recent Advances in QSAR Studies. Springer Netherlands, pp 327–366

Gramatica P (2012) Modelling Chemicals in the Environment. In: Livingstone D, Davis A (eds) Drug Design Strategies Quantitative Approaches. RSC Publishing

Gramatica P (2013) On the Development and Validation of QSAR Models. In: Reisfeld B, Mayeno AN (eds) Computational Toxicology. Humana Press, pp 499–526

Gramatica P (2014) External Evaluation of QSAR Models, in Addition to Cross-Validation: Verification of Predictive Capability on Totally New Chemicals. Mol Inform 33:311–314

Gramatica P (2016) Prioritization of Chemicals Based on Chemoinformatic Analysis. In: Leszczynski J (ed) Handbook of Computational Chemistry. Springer Netherlands, pp 1–33.

Gramatica P, (2020) Principles of QSAR modeling. Comments and suggestions from personal experience, International Journal of QSPR, 5(3) ,61-97 Open Access DOI:10.4018/IJQSPR.20200701.oa1

Gramatica P, Cassani S, Chirico N (2014) QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS. J Comput Chem 35:1036–1044. doi: 10.1002/jcc.23576

Gramatica P, Cassani S, Roy PP, et al (2012) QSAR modeling is not "push a button and find a correlation": a case study of toxicity of (benzo-)triazoles on algae. Mol Inf 31:817–835. doi: 10.1002/minf.201200075

Gramatica P, Cassani S, Sangion A (2015) PBT assessment and prioritization by PBT Index and consensus modeling: Comparison of screening results from structural models. Environ Int 77:25–34. doi: 10.1016/j.envint.2014.12.012

Gramatica P, Cassani S, Sangion A (2016) Aquatic ecotoxicity of personal care products: QSAR models and ranking for prioritization and safer alternatives' design. Green Chem 18:4393 – 4406. doi: 10.1039/C5GC02818C

Gramatica P, Chirico N, Papa E, et al (2013) QSARINS: A new software for the development, analysis and validation of QSAR MLR models. J Comput Chem 34:2121–2132. doi: 10.1002/jcc.23361

Gramatica P, Giani E, Papa E (2007a) Statistical external validation and consensus modeling: A QSPR case study for K-oc prediction. J Mol Graph Model 25:755–766. doi: 10.1016/j.jmgm.2006.06.005

Gramatica P, Papa E (2007) Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure. Environ Sci Technol 41:2833–2839. doi: 10.1021/es061773b

Gramatica P, Pilutti P, Papa E (2007b) Approaches for externally validated QSAR modelling of Nitrated Polycyclic Aromatic Hydrocarbon mutagenicity. SAR QSAR Environ Res 18:169–178. doi: 10.1080/10629360601054388

Gramatica P, Sangion A (2016) A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. J Chem Inf Model 56:1127–1131. doi: 10.1021/acs.jcim.6b00088

Li J, Gramatica P (2010) The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders. Mol Divers 14:687–696. doi: 10.1007/s11030-009-9212-2

Papa E, Battaini F, Gramatica P (2005a) Ranking of aquatic toxicity of esters modelled by QSAR. Chemosphere 58:559–570. doi: 10.1016/j.chemosphere.2004.08.003

Papa E, Gramatica P (2010) QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure. Green Chem 12:836–843. doi: 10.1039/B923843C

Papa E, Kovarich S, Gramatica P (2009) Development, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers. QSAR Comb Sci 28:790–796. doi: 10.1002/qsar.200860183

Papa E, Sangion A, Arnot JA, Gramatica P (2018) Development of human biotransformation QSARs and application for PBT assessment refinement. Food Chem Toxicol 112:535–543. doi: screening level

Papa E, van der Wal L, Arnot JA, Gramatica P (2014) Metabolic biotransformation half-lives in fish: QSAR modeling and consensus analysis. Sci Total Env 470–471:1040–1046

Papa E, Villa F, Gramatica P (2005b) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow). J Chem Inf Model 45:1256–1266. doi: 10.1021/ci050212l

Sangion A, Gramatica P (2016a) Ecotoxicity interspecies QAAR models from Daphnia toxicity of pharmaceuticals and personal care products. SAR QSAR Environ Res 0:1–18. doi: 10.1080/1062936X.2016.1233139

Sangion A, Gramatica P (2016b) Hazard of pharmaceuticals for aquatic environment: Prioritization by structural approaches and prediction of ecotoxicity. Environ Int 95:131–143. doi: 10.1016/j.envint.2016.08.008

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. Qsar Comb Sci 22:69–77. doi: 10.1002/qsar.200390007

Yap CW (2011) PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. JComput Chem 32:1466–1474. doi: 10.1002/jcc.21707

QSARINS is currently used by the Insubria group and several international QSAR groups: more than 800 free licenses active into 2021.