International Journal of Quantitative Structure-Property Relationships
Volume 5 • Issue 3 • July-September 2020

Principles of QSAR Modeling:
Comments and Suggestions From Personal Experience

Paola Gramatica, University of Insubria, Varese, Italy

ABSTRACT

At the end of her academic career, the author summarizes the main aspects of QSAR modeling, giving comments and suggestions according to her 23 years' experience in QSAR research on environmental topics. The focus is mainly on Multiple Linear Regression, particularly Ordinary Least Squares, using a Genetic Algorithm for variable selection from various theoretical molecular descriptors, but the comments can be useful also for other QSAR methods. The need for rigorous validation, also external, and for applicability domain check to guarantee predictivity and reliability of QSAR models is particularly highlighted. The commented approach is the "predictive" one, based on chemometrics, and is usefully applied to the prioritization of environmental pollutants. All the discussed points and the author's ideas are implemented in the software QSARINS, as a legacy to the QSAR community.

KEYWORDS

Applicability Domain, Consensus, Cumulative Environmental End-Points, External Validation, Multiple Linear Regression, Predictive Modeling, Splitting, Theoretical Molecular Descriptors

INTRODUCTION

At the end of my academic career, I wish to share with QSAR model developers my thoughts on QSAR modeling, as well as some personal anecdotes, based on my relatively short experience in this field: a field that was developed "only" in the course of the last 23 years of an academic career that lasted almost 46 years. Over these years many authoritative QSAR modelers have published papers on the development of QSARs, and have expressed their opinion on the best way to develop good validated QSAR models (e.g. some selected examples: Dearden, 2016; Dearden et al. 2009; Eriksson et al., 2003; Golbraikh & Tropsha, 2002; Gramatica, 2007, 2014; Kubini, 2002; Livingstone, 2000; Roy, 2007; Scior et al., 2009; Tropsha, 2010, Tropsha & Golbraikh, 2007; Tropsha et al. 2003). There are also interesting chapters in several books on the subject (too many to be listed here, so to quote only a few examples: Gramatica, 2009, 2013; Roy et al. 2015b, 2015c). Thus, this article, which is only "my small drop in the ocean" of QSAR research, is essentially focused on my ideas, it is most certainly not a review of the many papers that have been published on the various topics addressed here. For this reason, I apologize beforehand to the many authoritative authors who will not find

DOI: 10.4018/IJQSPR.20200701.oa1

citations of their relevant articles. This is a paper solely of words, which translate my thoughts, thus it does not include the formulas and equations reported in the cited publications.

BACKGROUND

The fact that I started research in computational chemistry, particularly in QSAR modeling, after long experience at the University of Milano in a completely different field of chemistry, organic chemistry, carrying out experimental works on the synthesis and structural identification of natural products, certainly had relevant influence on how I dealt with the new research field in QSAR. Another fundamental element in my researches was to learn QSAR modeling from a leader in chemometrics and theoretical molecular descriptors, Prof. Roberto Todeschini of the University of Milano-Bicocca.

QSAR Dualism

My 23 years of experience in QSAR modeling, my "second life" of academic research, has convinced me that the approach to QSAR modeling is highly influenced by the personal background of the QSAR modellers.

If a biologist is particularly interested in exploiting QSAR in order to understand how molecular structure influences the mode of action, or the biological mechanism, of studied chemicals, the "mechanistic" or "descriptive approach" will be the method of choice. In these "descriptive QSARs" the main attention is focused on modelling the existing data, fitting them the best as possible by using a few well known molecular descriptors that are personally selected by the modellers, because they are considered interpretable for "understanding" the correlation/causality in terms of mechanism. Such QSAR models are particularly useful for biological end-points (toxicity, mutagenicity, carcinogenicity, etc. and for drug design) for an attempted mechanism interpretation, mainly in local models, developed on homogeneous data sets of congeneric compounds. They can help towards a better understanding of the influence of some particular structural feature of the modelled chemical behaviour.

A chemist, instead, could be more interested in the possibility of representing a chemical structure in various ways, and would therefore use, *a priori*, numerous molecular descriptors to find, by methods of variable selection, the few descriptors really related to the end-point of interest. Moreover, if he/she has a chemometric background, particular attention will be to the statistical validation of QSAR models to verify their performance in robustness and predictivity also for new chemicals. A rigorous check to exclude a chance correlation in the model will also be applied. In this second case, the "predictive" approach will be the method of choice. The "predictive QSAR", also named "statistical QSAR", developed on more heterogeneous compounds and focused mainly on validation for new applications, can be usefully applied mainly to screen and rank more diversified compounds. This approach is very useful for big data sets, in prioritizing, for experimental tests, those compounds that are *in silico* highlighted as potentially more dangerous.

Thus, both approaches are specifically useful in different application fields. In several of my papers and meeting presentations I have stressed this point of the dualism that can be found in QSAR modelling, a dualism that could explain some of the lack of comprehension between researchers with different approaches. Indeed, Zefirov and Palyulin have already distinguished between "descriptive QSARs" and "predictive QSARs" (Zefirov & Palyulin, 2001). Also Prof. Fujita, a founder of modern QSAR, commented on this dualism in one of his last papers (Fujita & Winkler, 2016), expressing the need for the reconciliation of the "two QSARs".

"Predictive QSAR" has always been my personal approach: as a chemist with interest in environmental topics. My aim was to exploit the limited information available on compounds dangerous for the environment, taking advantage of the "fantastic" ability of good QSAR modeling to screen and highlight chemicals that could be of environmental concern, despite there being no experimental data available. This prioritization can also be applied to new, not yet synthesized, compounds, allowing a Green Chemistry approach to the synthesis of new potentially safer chemicals. Thus, the core of

my research was the validation of QSAR models to rigorously check the possibility of predicting reliable data also for new compounds.

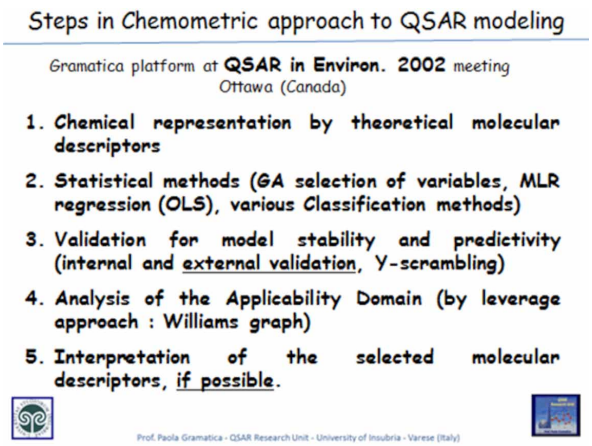
Principles of QSAR Models in Regulation

During preparation years of the new European REACH regulation (https://ec.europa.eu/environment/chemicals/reach/legislation_en.htm) (2001-2007), the role of QSAR was highlighted as a useful tool for filling the data gap: frequent meetings of researchers working on QSAR modeling were organized. I was selected to be a member of the QSAR Expert Group for JRC and for OECD, and collaborated in the proposal of the Setubal Principles (2002), and then their modification in the OECD principles for the validation, for regulatory purposes, of (Q)SAR models (https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm) (2004). In all those meetings I was in the group of researchers which stressed the need for validation, also external, and for the applicability domain check. In particular, I highlighted my personal idea that mechanistic interpretation is not always possible ("if possible", in this 5° principle, was my comment at the final meeting in Paris, 2004). I supported also the need for a glossary and I collaborated with various international experts to prepare the Guidance document on the validation of QSAR models (ENV/JM/MONO(2007)2) that includes useful information on good practices in QSAR modeling.

In the present commentary paper (probably my last paper) I will address the most important aspects in QSAR modeling according to my experience on Multiple Linear Regression (MLR) (particularly the Ordinary Least Squares (OLS) method), but most of my comments are also valid for other modeling methods (regression and classification). The line of the famous OECD Principles of QSAR validation is here followed: (1 - a defined end point, 2 - an unambiguous algorithm,3 - a defined domain of applicability, 4 - appropriate measures of goodness-of-fit, robustness and predictivity and 5 - a mechanistic interpretation, if possible). In fact, these principles, defined after much discussion in QSAR' and regulatory communities, are, in my opinion, an optimum summary of the most important points that need to be addressed to obtain reliable QSAR models, and the best line for the steps that must be followed for good QSAR modeling, with only one switch between Principles 3 – on Applicability Domain and 4 –on Validation.

It is interesting to note that the steps in my chemometric approach to QSAR modeling, steps that I always applied in my research and that I presented at the QSAR meeting in Ottawa (2002) (Figure 1), are very similar to the OECD principles that were later established in 2004 (Paris).

Figure 1. Slide of Gramatica at the 2002 QSAR in Environmental Sciences meeting in Ottawa (Canada) to illustrate her approach



All my ideas, illustrated here and applied in my research on organic compounds of environmental concern, are now implemented in the software QSARINS (QSAR-INSubria) (Gramatica et al., 2013), for MLR model development and validation, freely available by request on the web: http://www.qsar.it. I decided to put my experience in QSAR modeling in QSARINS and also to make available many curated data sets and some models, developed by my Insubria research group in the period 2003-2018, in the module QSARINS-Chem (Gramatica et al., 2014), as my personal legacy to QSAR community. A selection of our QSAR models is now more easily available in the recent Standalone version of QSARINS-Chem, free downloadable from our website.

STEPS IN QSAR MODELING

Data Curation and Data Set Preparation - Partly in the OECD Principle 1: "A Defined End-Point"

The first point, and sometimes the most challenging in QSAR, is that QSAR modellers need experimental data as input for their models. A popular statement among QSAR model developers is "garbage in, garbage out"; in fact it is well known that the output of any model cannot be better in quality than the input. Thus, specific work must be devoted *a priori* to the data curation step to avoid the proposal of "garbage"-models, simply because "garbage"-data were used. In this context the careful selection of good input data is crucial (Livingstone, 1995; Tropsha, 2010).

Experimental Data

My previous experience in a chemical laboratory showed me that experimental data are neither perfect nor always reproducible, which is contrary to the conviction of some people who place their confidence only in experimental data and not in QSAR predictions. However, as experimental data are the unique source of input information available to modelers, QSAR models must exploit, as much as possible, all the experimental information available, even if this is limited and often characterized by high uncertainty in its measurement.

Data variability can arise from chemical purity, variability in experimental protocol and in biological response. Much of the variability resulting from chemical purity and protocol could be avoided, while it must be remembered that the reproducibility of biological tests is much lower than that for other properties (for instance physico-chemical properties). Therefore, it is obvious that QSAR models for biological activities have generally lower performances than those for physico-chemical properties. Within a specific test, intra-laboratory and inter-laboratory variability in the measured endpoint add to the difficulties of selecting a training set. A "golden" training set should be one where data are obtained by the same method, in the same lab and possibly by the same researcher, but this is obviously seldom possible.

Due to the above reasons, unfortunately good experimental data with quality useful for QSAR modeling are often scarce, especially for end-points of regulatory interest and for environmental compounds belonging to a wide variety of chemical classes. Often experimental data are available only for few and the most studied chemicals, while data are not available for the majority of chemicals. Moreover, the data arise frequently from different sources, by experiments that are not comparable in terms of applied experimental test, time, conditions, etc. This causes dangerous consequences in the derived modeling (Zhao et al., 2017).

Even though it is not always possible (even almost impossible) for a QSAR modeler to select homogeneously determined data, however, some serious mistakes could be certainly avoided: for instance, the mixing of toxicity end-points on different species of fishes with the purpose of proposing a supposed general model for any fish (small or large…).

In conclusion, as is evident, the preliminary work on selection of experimental data is often the most difficult step for a QSAR developer, especially because experience in selecting and evaluating

the meaning and quality of some experimental data is not always possible (it is not possible to have experience in every field of the experiments used as modeling input). This problem of understanding the meaning and reliability of input data has become even more crucial in recent years, as use is being made of very big sets of experimental data of various origins, often of very complex and/or different biological meaning. The quality of each data could be good when they derive from certified laboratories, but often some experimental data cannot be fully understood, due to the lack of a unique ontology of the experimental protocols and results. Main problem is that these kinds of mixed data, as said above, could be not useful as homogeneous input data for reliable QSAR modeling.

Chemical Structures

Obviously, data curation is relevant not only for experimental response but also for the chemical structures of the studied compounds (Young et al, 2008). A careful check of the chemical structure of the training molecules must be done to avoid the introduction of incorrect structural information.

The huge problem of data curation (for both experimental data and chemical structures) was the subject of many useful papers in recent years (Fourches et al., 2010; Li & Gramatica, 2010; Gramatica et al, 2012; Tropsha, 2010; Waldman et al., 2015) and some automated workflows are also proposed (Gadaleta et al. 2018, Mansouri et al. 2016, Ruusmann & Maran, 2013).

It is evident that the higher the number of available data for modelling, the bigger the information and the more reliable could be the data predicted by the models. But attention to quantity must not overcome that to quality. This comment is particularly relevant for more recent works on modeling big data sets. In these cases, in addition to my previous comments on mixed experimental data of various end-points of different biological mechanism, in my opinion, the simple automatic use of SMILES notation as input for chemical structure, without personally verifying the hidden molecular structure, could lead to unreliable QSAR prediction due to the possibility of some wrong input structures.

Other important aspects to be taken into account are: the response distribution and the structural domain of the modeled chemicals.

Response Distribution

A requirement for regression modeling, mainly by Ordinary Least Squares (OLS), which is the simplest method that I mainly applied in my research, is that the response distribution should be approximately normal. Least squares estimates for regression models are highly sensitive to response outliers, observations which break the assumption of normally distributed data and do not follow the pattern of the other data. Therefore, particular attention must be given to the presence of these outliers at the beginning of the modelling, and cleaning of the data set is required.

Some published models where one or two chemicals with the highest (or lowest) response value are isolated, and very far from the rest of the studied set (high leverage compounds for their response, therefore strongly influential), are not reliable. Their plot of experimental vs predicted values is similar to a regression line passing between 2 points. Obviously the R^2 value is very high! But the model is not robust, being completely dependent on the unique influential chemical. For this reason, I had recommended (Chirico & Gramatica, 2013) that QSAR models are presented in scientific publications not only with their statistical performances, but also with the plot of predicted versus experimental values.

Concluding, the analysis of the response distribution is a necessary preliminary step that must be always done.

Structural Domain

Also, the structural variability of the studied dataset must be carefully analyzed. In fact, both kinds of high leverage chemicals, for response or for structure, could have too strong influence in selecting the descriptors in the model just to make these molecules fit into the model domain: the result would be a good fitting model where the high leverage chemicals are well calculated, but with limited or

even null performances in prediction for new chemicals. Other possibility: a highly isolated structural outlier could be not sufficiently influent in the variable selection, so it will be not well calculated, resulting finally in a response outlier. All these outliers must be identified in *a priori* careful analysis of the studied data set: it could be better to delete them before the modeling to obtain more stable and predictive models. If these kinds of compounds are conserved in the training set there is the possibility to have a larger domain of the model for future applicability, but the model itself would be more instable. This possible pitfall must be taken into consideration.

Regarding the *a priori* analysis of the structural domain, this can be done, as always in my researches, by applying Principal Component Analysis (PCA) to the studied set of compounds represented by different molecular descriptors. This PCA can highlight some structural outliers (high leverage chemicals) too far from the majority of the compounds to be well modeled: in these events, as said before, their deletion is suggested for a more robust modeling. However, in some cases it could be interesting to enlarge the chemical domain of the model and if a high leverage compound is not too far from the majority of the data set, near the cut-off Hat value, it could be kept in the model, but careful analysis of the relative results must be done.

PCA, applicable to various molecular descriptors, is implemented in our software QSARINS, where it can be applied on the complete data set and also on the split data in order to verify the distribution of the compounds in the training and prediction sets to guarantee the structural representativity in both sets. This check after the splitting is important, not because a splitting of the full data set, done for verifying external predictivity, could hide a cheat to do well QSAR models, but simply because any QSAR model can be reliably applied only in its Applicability Domain (see below).

A distinction between "local" and "global" models is here useful.

Local and Global Models

"Local" models are generally developed on smaller sets of similar chemicals or structurally-related or, for instance, with the same hypothesized Mode of Action (MOA), consequently they have a limited and specific applicability. These models could have the advantage of an easier interpretability of the few modeling descriptors and for this reason are preferred in the "mechanistic" approach.

"Global" models are built on bigger data sets, using a rich enough structural diversity of compounds in training sets, to ensure as much as possible generalizable model. These models are based on a greater number of descriptors, taking into account the various structural features that are related to the studied end-point in a heterogeneous data set. For this reason, they could be often of difficult interpretation for a potential common mechanism, mechanism that, on the contrary, could be different in a large data set.

"Global" models are more appropriate for virtual screening of big heterogeneous data sets, where the main concern is to prioritize the most dangerous chemicals, in order to focus attention on them and to reduce the number of compounds that need to be experimentally tested.

In my work on *Pimephales* toxicity (Papa et al. 2005) a comparison was done on local models, developed on small sub-sets of compounds with the same predefined MOA, and global models (DTP-models: Direct Toxicity Prediction-Models), based on the complete set of studied chemicals with mixed MOAs. The predictive performances are similar, but the applicability is different: local models can be reliably applied only to compound with a predefined MOA (but this is not always possible), while global models can be applied without any *a priori* knowledge of a toxicity mechanism. This second opportunity is certainly highly useful for environmental chemicals, mainly for screening purposes with prioritization aim.

Also, Puzyn et al. (Puzyn et al., 2011) recommended that:

Whenever global models fulfil all quality criteria proposed by the OECD principles, they should be applied in practice without necessity of developing a series of local QSPRs. Such a recommendation is reasonable, because of three reasons. First, the global models allow for simultaneous predictions

of physicochemical properties for even many hundreds of compounds. This feature is very important from the economic point of view, regarding that the number of new chemicals synthesized and/or identified in the environmental compartments is growing exponentially.

Second, the global modeling approach may be the only possibility of modeling, when the number of chemicals from one specific class demonstrated, the performance (predictive ability) of global models is not always worse than these of local ones of the chemically related compounds is insufficient to calibrate and appropriately validate a local QSPR model.

Third, as both kinds of models can be useful and predictions obtained from both local and global models in consensus can be the best. In fact, I completely agree also with the comment: it is probable that a combination of global and local QSPR models within framework of consensus models may be the optimal approach for construction of stable predictive QSPR models with mechanistic interpretation. (Raevsky, et al., 2015)

Development of the QSAR Model - The OECD Principle 2: "An Unambiguous Algorithm"

The development of a QSAR model must produce a transparent model algorithm that generates predictions of an endpoint from information on chemical structure. The algorithms used in QSAR modelling (in terms of methods and molecular descriptors) should be described thoroughly, so that a user can understand exactly how the estimated value was produced, and be able to reproduce the calculations, if desired. In fact, without information on how QSAR estimates are derived, the performance of a model cannot be independently verified. The algorithm of a QSAR model must be not only always reproducible, but also preferably with easy applicability. For this reason, and in line with the Occam's razor requiring simplicity, there is a general preference for simpler and more widely understandable models, mainly for applicability also by the not experts in regulation contest. In fact, some methods have simple and explicit equations (for instance MLR), whereas other methods (i.e. machine learning, such as Neural Networks (NN), Support Vector Machine (SVM), etc.) result in more complex expressions and, for this reason, are erroneously perceived by not experts as "black-boxes".

I believe that this position is determined by the request of not experts in QSAR of understanding in details the method they wish to apply. However, in their "push a button" approach the users should be confident on the models that are hopefully well developed by experts in QSAR modeling.

Non-linear methods are obviously those that are more able to better model any end-point, simply because they are able to capture also the non-linear aspects of the response. Thus, it is trivial to write, as reported in several published papers, that NN or SVM models have better performances of MLR. The aim of MLR is to model in the simplest way, capturing only the most relevant aspects of the chemical structure related to the end point.

As said, in my researches I developed MLR models, in particular by OLS method: some of these models are available in the module QSARINS-Chem (Gramatica et al. 2014) of QSARINS software (Gramatica et al., 2013).

Molecular Descriptors

The basis of QSAR modeling (where S means Structure) is that any property/biological activity/reactivity of chemicals is related to their molecular structure. Thus, the crucial point is how to represent a molecular structure translating it in a number to be inserted in an algorithm. Molecular descriptors have this role and are obviously the core of any QSAR models.

Descriptor of Hydrophobicity: Log P (log Kow)

QSAR models for various environmental end-points, and mainly QSARs for toxicities, are highly dominated by the historical background of QSAR. Therefore, a lot of models for ecotoxicity are based on a descriptor of a molecular property more than of the molecular structure: log Kow (or older notation logP), the partition coefficient between octanol and water. It is supposed that this hydrophobicity term has a clear "mechanistic" meaning as it would reproduce the ability of a substance to enter cells through the lipid membranes and would indicate both the toxicant uptake and baseline toxicity. For this reason, it is widely used as a modeller of toxicity or other partition coefficients (such as Bioconcentration Factor (BCF)) or Soil sorption on Organic carbon (Koc).

The reason of this wide use of logKow-based models for environmental chemicals (implemented also in EPI-Suite (https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface) and ECOSAR (https://www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model) is mainly related to the diffused supposed "interpretability" of this parameter. However, there are a lot of papers that highlight some problems related to this descriptor.

The problem of logKow reliability was dealt by various authors (for instance: Kaiser, 2003; Renner, 2002). In fact, the experimental determination of logKow can be a complex matter, and experimental values can differ greatly even when referred to the same compound. Thus, several approaches have been developed for the theoretical calculation of logKow; they are other QSAR models that interpret molecular structure information in their own peculiar way. But also, in these calculations, strongly dependent on the applied QSAR, it is not uncommon to have differences of several orders of magnitude (Benfenati et al., 2003). Therefore, logKow cannot be considered a univocal descriptor since it could provide, for each studied compound, different end-point predicted values, strictly dependent on the logKow used for the QSAR development, compromising model reproducibility, especially when no indication is reported on which kind of logKow was used. It is also important to highlight that a calculated logKow looks like a single descriptor, but it is, really, the condensation of all the structural information represented by fragments and correction factors, applied for its calculation. For this reason, the dimensionality of a logKow-based model appears lower than a corresponding model based on other theoretical descriptors, but it is simply a guise.

Additionally, if an experimental logKow is used for modeling an activity, it is a property (P) not a structural descriptor (S), thus QSAR becomes QPAR.

The arbitrary choice of one specific logKow, as is usual in QSA(P)R studies does not necessarily lead to the highest quality model for the analysed data set. This parameter is not always selected as modelling descriptor, by variable selection methods, when other theoretical descriptors are available. Therefore, in my opinion, it should be possible to omit logKow as a variable when, included in the list of input variables, there is the same type of information on which logKow computation depends, i.e. theoretical structural descriptors able to furnish analogous information.

This is exactly the approach I decided to apply in my group, in line with analogous studies of other authors (Kaiser, 2003; Ren, 2003) in models for various toxicities and properties: i) general Direct Toxicity Prediction (DTP) in *Pimephales* (Papa et al. 2005), ii) various aquatic toxicities of Personal Care Products (PCPs) (Gramatica et al., 2016a) and of pharmaceuticals (Sangion & Gramatica, 2016a), iii) Bioconcentration (Gramatica & Papa, 2003, 2005), iv) soil sorption coefficient on organic carbon, Koc, of pesticides (Gramatica et al. 2007), among others. In these researches, we had demonstrated that the logKow-free models had the highest predictive power, in comparison with models based on different kinds of logKow, were reproducible, and additionally gave more detailed information on the structural aspects responsible of the modelled end-points. In addition, these global models, verified for their predictive on external sets, are applicable even in the absence of "*a priori*" information about the supposed mechanism of a chemical.

Fragments

Frequently used molecular descriptors are counters of some groups, substructure descriptors or fragments. They are the simplest descriptors, collectively named as mono-dimensional or 1D-descriptors. In many QSAR models of the "mechanistic" approach, some fragments are identified as structural alerts for some biological end points, i.e. this kind of descriptors can highlight which parts of the molecule confer the activity. They are obviously more easily interpretable and can be useful because biological activities are often predicted by the presence or absence of a particular group of atoms in a molecule.

However, the use of substructure descriptors has some troubles.

First, the dimensionality is usually much higher, i.e., thousands of substructures could be present in a typical training set.

Second, a new molecule may have substructures that are not present in the training set, i.e., it can be outside the structural space of the training set. Thus, although the fragment-based approaches appear to give reliable predictions for some end-points, they cannot reliably estimate the response for structures containing a "new" fragment, and the number of correction factors can continually grow as compound diversity increases.

Third, these are substructure descriptors that cut the molecule in many fragments, losing, in this way, any possible relevance of the contemporaneous presence and interactions of various functional groups in each chemical structure.

Last but not least, the final consequence is that the fragment- models are frequently based on a high number of descriptors, often against an acceptable ratio between chemicals and modeling descriptors (see below in Ratio between Chemicals and Descriptors).

Holistic Descriptors

It is important not to forget that a molecule should be considered in its physico-chemical properties and chemical or biological behaviour/reactivity as a whole, not simply as a sum of parts, thus whole-molecular or holistic descriptors are highly useful in QSAR modeling. Holistic descriptors can take into consideration not only the presence (yes or not), as in fragments-models or the count of some groups, but also the relative position of different functional groups or fragments in a molecule. There are a plenty of holistic descriptors that can consider the atom connectivity and the topology of molecular graph on a plane (2D-descriptors) and also the three-dimensional conformation of a compound (3D-descriptors).

A comprehensive analysis of thousands of molecular descriptors was done by Todeschini & Consonni in their Handbooks (Todeschini & Consonni, 2000, 2008).

The existence of a huge number of different molecular descriptors, experimental or theoretical, to describe chemical structure is a great resource as it allows to have different X-variables available that take into account each structural feature under various aspects.

In some cases, to avoid any bias introduced by the selection of a specific 3D-conformation (often unknown for the modelled response), it is better to avoid the use of 3D-descriptors, even if for some complex end-points, such as biological behaviours, the best models are often based on three-dimensional information.

A lot of developers of QSAR models, mainly in the "predictive" or "statistical" approach (including my group), prefer to have *a priori* available as many molecular descriptors as possible, including holistic descriptors. The use of large number of descriptors, some of difficult interpretability, was largely criticized in these years, by several researchers, in particular those that apply the "descriptive" or "mechanistic" approach. However, it is evident that the number and diversity of descriptors guarantee the coverage of the molecular structure space more efficiently than if limited to only a few descriptors.

Variable Selection for Correlation

When dealing with a lot of molecular descriptors, rigorous variable selection must be done to find a simple and predictive QSAR model, which must be based on the least number of descriptors (following the Occam's razor philosophy), descriptors that must be the least inter-correlated as possible.

The "statistical" approach, as already mentioned and is well known, is based on the fundamental conviction that the QSAR modeller should not influence *a priori*, or personally, the descriptor selection through mechanistic assumptions, but should apply unbiased mathematical tools to select, from a wide pool of input descriptors, those descriptors most correlated to the studied response, verifying their ability in predicting new chemicals (through validation and Applicability Domain check).

The variable selection should be done in two following reduction steps: a preliminary objective pre-reduction, using only the molecular descriptors, the independent variables (X), and a subsequent modeling variables' selection, which additionally use the response values, the dependent variable (Y).

In the first objective pre-reduction, descriptors that have constant or near-constant values (those with too limited variance in the data set) and those pair-wise correlated must be firstly eliminated to identify a more limited set of descriptors on which to apply the second variable selection step. While the deletion of constant descriptors is always applied, a comment is here necessary on the near-constancy of some descriptors in the studied data set. If this kind of descriptors that have a different value only for few chemicals (for instance counters of some functional group: 0 if absent and 1, 2 etc. if present) enter in a QSAR model, these descriptors are specifically selected to allow the good modeling of few high leverage chemicals which have influence in the selection of those specific descriptors, with such a low variance. It is evident that they are prominent in fitting the data, but the models could be not predictive for new compounds. In the preliminary objective reduction, these nearly constant descriptors must be also eliminated from the pool of the descriptors that are then used for the subsequent variable selection.

In theory, one would like to have predictors in a multiple regression model that each have a different influence on the response and are independent from each other. However, some molecular descriptors could be only different views of the same molecular aspect and for this reason can result in high correlation. Thus, in practice, some predictor variables can be inter-correlated; therefore, to avoid redundancy in descriptor information, one of the descriptors found to be pair-wise correlated with a percentage of generally 95%-98% (that can be chosen in QSARINS) must be excluded in this pre-reduction step. It is recommended to always publish the inter-correlation of modeling descriptors of the proposed QSAR models. The filtering of collinear or constant descriptors must also be done after the splitting of the available data for external validation: both for training and prediction set, because in each splitting the situation could change (this is present in the updated 2.2.4 version of QSARINS).

Secondly, modeling variable selection methods are applied to this pre-reduced set of descriptors to further reduce it to the true modeling set, not only in fitting but, most importantly, in prediction. This is the crucial step where the selection of the molecular descriptors that will be potentially the most relevant to model the endpoint must be selected and included in the QSAR relationship.

Such selection is performed in the "statistical/predictive" approach by alternative variable selection methods. Several strategies for variable subset selection have been applied in QSAR (stepwise regressions, forward selection, backward elimination, simulated annealing, evolutionary and genetic algorithms, among those most widely applied).

We apply Genetic Algorithms (GA) (Haupt & Haupt, 2004) which is a very effective procedure, widely and successfully applied in many QSAR approaches. In my experience, I have verified that, before the application of GA, it is highly useful to apply the All Subset procedure that explores all the possible combinations of few descriptors, normally 2 or 3 max depending on the data set dimension. This guarantees that the best modeling descriptors, which are generally the first to be selected, are not lost in the following application of GA. All subsets and GA procedures are both implemented for OLS within the QSARINS software.

The application of the Genetic Algorithm-Variable Subset Selection (GA-VSS) procedure to capture the most relevant variables in modeling a response provides a large set of possible models: a population of models, ordered according to their validation performances by the selected fitness function, for instance Q^2 (but other fitness functions could be preferred and chosen in QSARINS). Often different models have nearly equivalent predictive performance, because, being based on a variety of descriptors, they can reflect different aspects of molecular structure equally related to the response. The selection of the best model in this population can be done by applying different approaches together: i) the difference between fitting and cross-validation parameter must be the lowest (R^2 and Q^2 should be comparable); ii) all the statistical parameters for external validation on chemicals that haven't participated to the variable selection (prediction sets) with highest values in agreement; iv) lowest difference between Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) on training and prediction sets; iv) minimum number of outliers. These points will be better dealt in the following paragraphs on validation and on applicability domain.

In our works, particular attention was always paid also to the collinearity of the selected molecular descriptors: in fact, to avoid multicollinearity without, or with, "apparent" prediction power (due to chance correlation), regression is calculated only for variable subsets with an acceptable multivariate correlation with response, by applying the *QUIK* rule (Q under Influence of *K*) (Todeschini et al.,1999). When there are models of similar performances, those with a global correlation of [X+Y] block (K_{XY}) greater than the global correlation of the X block (K_{XX}) (X being the molecular descriptors in the model and Y the response variable) should be selected and further verified.

Check of Probability of Chance Correlation

The QSAR algorithm establishes the correlation between the studied response and the molecular descriptors, but some concerns have been raised in the literature (Doweyko, 2008; Rucker et al., 2007; Topliss & Costello, 1972; Topliss & Edwards, 1979) emphasizing that correlation between variables does not automatically imply that one causes the other and that chance correlation could occur, mainly in modeling small data sets starting from a large set of descriptors. However, correlation is a preliminary and fundamental requirement for causation, but must be controlled.

The most important and commonly applied ways to exclude chance correlation for each developed model are: a) to carefully verify the statistical predictivity of QSAR models by their validation (as requested by OECD principle 4), also externally on new chemicals, b) to scramble the response (Y scrambling) (Eriksson et al, 2003; Golbraikh & Tropsha, 2002; Tropsha et al. 2003). In Y-scrambling new models are recalculated for randomly reordered response using the same descriptors of the selected model. Evidence that the proposed model is well founded, and not just the result of chance correlation, is provided by obtaining new models, on the set with randomized responses, that have significantly lower R^2 and Q^2 than the original model. This is the most widely applied test to exclude chance correlation.

A deeper check of probability of chance correlation can be also applied (Rucker et al., 2007), by applying permutation test *a priori* while performing the variable selection procedure. In the new version 2.2.4 of QSARINS, to calculate the probability of possible chance correlation, the descriptors' selection procedure can be repeated many times (i.e. in multiple parallel populations of models) using randomized responses. The distribution of the performances of the best randomized models in each population is then used to calculate the probability of chance correlation for each model, comparing the performances of the original model and those on randomized responses, based on other descriptors. When a statistically relevant number of $R^2$and Q^2_{LOO} values, calculated for the best randomized and not randomized models, are similar, the quality of the not randomized model (proposed) is dubious. In this case, if this "dubious" model has been developed using correctly filtered descriptors (avoiding collinearity), has been deeply validated also externally, and had passed the Y-scrambling test, a careful analysis and possibly interpretation of the selected descriptors is necessary to guarantee the reliability of the proposed model, even if not substantially different from a possible chance correlation.

However, if the correlation is confirmed after rigorous verifications/validations, it has a reason of existing and it is a problem of the human mind if the hidden cause, captured by the molecular descriptors, is not discovered or understood. I fully agree with the words of Katritzky:

QSPRs using large descriptor pools have been criticized for their increased sensitivity to chance correlations. However, the possibility of chance correlations can be minimized to be negligible by using appropriate procedures…. Molecular descriptors are not random numbers. The descriptors encode features of molecular structure which can influence and control observed biological or physicochemical effects. While the interpretation of descriptors is not always clear or unique, in most cases they encode useful information: how successfully this information will be extracted depends on the skill of the researcher. (Katritzky et al. 2008)

Ratio Between Chemicals and Descriptors

The number of descriptors in a QSAR model is highly dependent on the quality of the studied data set, both in terms of number and type of compounds and of the modelled end-point. It is evident that, in general, biological end-points are more complex than physico-chemical properties and their modeling could require a higher number of descriptors. It is also clear that any data sets must be modelled by a few numbers of descriptors as possible.

Restrictions in the number of variables that should be used in a QSAR model as compared to the number of observations must be applied. In literature, Topliss and Costello (Topliss & Costello, 1972) and later on Cronin and Schultz (Cronin & Schultz, 2003) indicated that:

The ratio of observations to variables should be as high as possible, and at least 5:1.

In my opinion based on my modeling experience, this value is too low and cannot be generalized. The addition of too many descriptors is driven by the specific information included in the training set, some of low significance and necessary only for modeling some particular chemicals. In my experience, for small/medium data sets (20- 60 compounds) the ratio compounds/descriptors should be always higher than 10-15:1. However, this ratio becomes even senseless when big data sets are modelled. For data set of hundreds of chemicals this ratio should be at least 50:1, or better higher when the number of compounds increases. All models available in QSARINS-Chem (Gramatica et al. 2014) follow these ratios.

Any model based on more than 7-9 descriptors, mainly if counters with low variance in the data set, is, in my opinion, suspicious; it is probably a pure fitting model and must carefully checked for robustness and external predictivity. The overfitting is not useful for subsequent application of the model for prediction of new chemicals.

In this context, it is also obvious that the best number of descriptors must be not determined by the increasing of R2 value while adding new descriptors, but by the check of stabilization of the validation parameters (various Q^2). For this reason, the validation parameters are the fitness functions for GA in QSARINS.

Unfortunately, in my reviewing experience I have seen in many papers models based on a too high number of descriptors even for small data sets, following the above suggested ratio of 5:1, sometimes with a counter descriptor with a value, different from the majority of the set, only for 2-3 compounds. As said before, these low variance descriptors (near constant) must be deleted in the *a priori* objective variable reduction.

Moreover, this kind of fitting models could be probably interesting for interpreting some trend in the data set in the mechanistic approach, but they are not certainly useful for producing reliable predictions for new chemicals in the predictive approach. I believe that it is wrong to define "predicted data" the data calculated by this kind of models.

Importance of Descriptors in a Model

The influence of each descriptor appearing in a multilinear regression model in increasing or diminishing the modelled response is clearly understandable by the positive or negative sign in the QSAR equation. But the importance of each modeling descriptor cannot be directly determined simply by the model coefficients, as sometimes erroneously done. For assessing the relative importance of variables in a model, these variables must be easily compared to each other. Different descriptors have different scales, thus the standardized regression coefficients must be verified. Despite the name, it isn't actually the coefficients that get standardized, but the descriptor values, to be directly comparable. The descriptor values are calculated by subtracting the mean from the variable and dividing by its standard deviation. Descriptors with higher absolute value of their standardized coefficients explain greater part from the variance of the modelled response and should be better explained.

Model Reproducibility

The important issue of model reproducibility is also covered by the OECD principle 2.

Indeed, QSAR models must be reproducible to be practically applicable for the prediction of new chemicals that have no experimental data.

The need for regular check and updating of published QSAR models is particularly evident, if these models are to be useful for practical applications, for instance in regulation, and not just for scientific purposes. However, in the literature there is a plethora of QSAR models, daily published, that have no practical applicability because the molecular descriptors are no more available, or because they were calculated by not accessible software or by a different version of commercial software (Gramatica & Papa, 2005; Roy et al., 2011).

For this reason, some Insubria models, originally based on the commercial DRAGON descriptors (Todeschini et al., 2007), have been redeveloped by free online PaDEL-Descriptors (Yap, 2011). In the module QSARINS-Chem of QSARINS some of these models are available with the corresponding QMRF (QSAR Model Reporting Format), allowing the model application to new chemicals calculating the molecular descriptors exactly by the same PaDEL version used during the redevelopment. In this way, the reproducibility for future application of the models is guaranteed. Recently, a selection of these models has been implemented in the new Standalone version of QSARINS-Chem, free downloadable from our website and therefore more easily and widely applicable.

In the QSAR Application Toolbox (https://qsartoolbox.org), useful mainly for grouping chemicals into categories based on mechanism, and in the QsarDB (Ruusmann & Maran, 2013), models that guarantee the reproducibility can be applied. The QsarDB repository aims to make the processes and outcomes of *in silico* modelling work transparent, reproducible and accessible and makes data and models FAIR (Findable, Accessible, Inter-operable, Re-usable) (Ruusmann et al, 2015). The Insubria models are, so far, among the most numerous in the QsarDB repository because they are transparent and reproducible. Some models from Insubria, with the corresponding QMRF, are also included in the JRC QSAR model database. (http://qsardb.jrc.ec.europa.eu/qmrf)

Combined Modeling: Predictions by Consensus

The wider utility of QSAR models in environmental context is certainly for screening big datasets of chemicals without experimental data, in order to update priority lists, therefore the best predictions can derive by combining the results of various models in a consensus approach (Cassani et al., 2013a; Cassani & Gramatica, 2015; Gramatica et al. 2004b, 2007, 2012, 2015, 2016b; Kovatcheva et al., 2004; Papa et al. 2014; Sangion & Gramatica, 2016b; Zhu et al., 2008; Bhhatarai et al., 2011).

As mentioned above, the application of the Genetic Algorithm-Variable Subset Selection (GA-VSS) procedure to capture the most relevant variables in modeling the response provides a large set of possible models, based on molecular descriptors of different kind, but sometimes with nearly equivalent predictive performance. Thus, there could be many possible "best" models, all valid. So, how to select the final "best"? And why select only one? Different models can reflect different aspects

of molecular structure that are related to the modelled response. Or else, different descriptors are simply alternative viewpoints to represent same structural features, while taking into account same characteristics in different ways; thus, they lead to not perfectly equivalent, but similar, modeling descriptions of the studied end-point. Each individual QSAR model may overemphasize some structural aspects related to the modelled end-point and underestimate others.

Thus, it seems reasonable that a consensus (or combined) QSAR model, which can be derived by calculating an average result from good individual models obtained by GA, might provide better predictive ability than the majority of each individual model (Gramatica et al., 2004b, 2007, 2012; Papa et al., 2014). This combined modeling might take into account several peculiar aspects of some particular structures contemporaneously.

For the comparison of different QSAR models in a GA population it is useful to examine the variability in predicting responses of similarly reliable models. In fact, the models to be combined should be the most diverse. A comparison can be done from the loading plot derived by applying PCA to residuals of several models. Different models (e.g. with dissimilar residual profiles) appear distant in the PCA plot, while similar models (e.g. giving similar predicted values) are clustered. In my approach, the models selected as being the more representative for the combined modeling are those which are the most dissimilar in the PCA graph. It can be verified immediately that the difference in the residuals derives, as expected, from the difference in the molecular descriptors: indeed, the selected models are the most dissimilar in molecular descriptor composition. In conclusion, on the basis of different structural descriptions the predictions by an average/combined model can be derived from models corresponding to different prediction schemes.

It is also interesting to note that often some chemicals are predicted with practically the same response by all the models, selected for the combined modeling, while for other chemicals the response is predicted differently by each selected model. The chemicals in the former case can be defined as "prediction safe", being independent from the molecular description, while those in the latter case can be defined as "prediction sensitive", being selectively related to the structural information included in each different model. The range of predicted values among the compared models can be used to highlight "prediction sensitive" chemicals and therefore those of greatest concern (higher Delta). If the modelled response is detected experimentally *a posteriori,* "prediction sensitive" compounds could be useful in selecting the best among several possible models, and in interpreting the model's molecular descriptors in terms of the particular structural features characterizing these compounds. In conclusion, a selection of the most dissimilar models in the population of GA-models allows the proposal of consensus predictions and the highlighting of "prediction safe" or "prediction sensitive" chemicals depending on their independence, or not, of the model choice.

I applied this procedure in modeling the atmospheric reactivity of Volatile Organic Compounds (VOC) with OH radicals (Gramatica et al., 2004b) and Koc of pesticides (Gramatica et al., 2007), but other choices are possible. For instance, in QSARINS, in addition to PCA of residuals, it is possible to perform and verify the PCA of molecular descriptors in each model of the GA-population, allowing the direct selection of the most diverse models in term of structural information.

The consensus or combinatorial approach can be also applied in combining the predictions obtained by different modeling approaches: local and global, linear and not linear (Cassani et al., 2013a; Cassani & Gramatica, 2015; Gramatica et al, 2015, 2016b; Kovatcheva, 2004; Sangion & Gramatica, 2016b; Zhu et al., 2008).

Model Validation - The OECD Principle 4: "Appropriate Measures of Goodness-of-Fit, Robustness and Predictivity"

I have exchanged the order of the OECD Principles 3 and 4 in this discussion, because I'm convinced that it is necessary that a QSAR model, which, as all mathematical models, is a simplification of the studied phenomena, must be verified for its reliability, before to verify the applicability domain, mainly for new chemicals' prediction.

Fitting

Firstly, a QSAR model must be able to learn from available data and reproduce them well (goodness of fit, verified by R^2, the coefficient of multiple determination). R^2 shows how well the data points fit the regression line. It must be as high as possible: 1 or 100% for a perfect calculation. However, this is never possible because the error of calculation cannot be smaller than the experimental error of the modelled end-point, otherwise this indicates an over-fitted model.

It is most important to avoid overfitting by adding more and more descriptors. In fact, the value of R^2 can generally be increased by adding additional predictor variables to the model, even if the added variable does not contribute to reduce the unexplained variance of the dependent variable. This inconvenience can be avoided by using the so-called adjusted R^2 (R^2_{adj}) that is adjusted for the number of terms (number of compounds and descriptors) in the model. The value of R^2_{adj} decreases if an added variable to the equation does not reduce the unexplained variance: thus, when more and more useless descriptors are added. On the contrary, if more useful descriptors are added, R^2_{adj} will increase. Adjusted R^2 will always be less than or equal to R^2.

Internal Validation

Stability or robustness of a good fitting model must be verified by internal validation trough iterated cross-validations: Leave-One-Out (LOO), Leave-Many-Out (LMO) or bootstrap and calculating the corresponding cross-validated correlation coefficients (Q^2_{LOO}, Q^2_{LMO}). The chemicals iteratively put aside, to verify how well the model predict them, constitute the test sets, generally around 30% of the complete data set in LMO. At this level, it is highly important that the difference between R^2 and Q^2 is low (no more than around 10%), otherwise the model could be simply a fitting model, but unstable in predicting test set chemicals.

The Root Mean Square Error (RMSE) or the Standard Deviation Errors in Calculation (SDEC) summarize the overall error of the model: they are calculated as the root square of the sum of squared errors in calculation divided by the total number of chemicals.

These parameters must be calculated and compared both on training set, in Leave-One-Out cross-validation and on external chemicals ($RMSE_{TR}$ vs $RMSE_{LOO}$ vs $RMSE_{EXT}$ or SDEP). The more similar are these compared values the more the model has a general applicability. Similarly, the simpler MAE, Mean Absolute Error, can be compared. The different quality of these two parameters is commented in the literature (Chai & Draxler, 2014; Roy et al., 2016).

However, it is here important to highlight that RMSE and MAE, which are both dependent on the measure scale of the end points values, are useful mainly to compare the quality of models developed for the same end point and for this reason should not be considered useful parameters for the comparison of models for different end points. As these two parameters are not equivalent and there are no concordant opinions on the superiority of one parameter over the other, in my opinion, they should be verified together, always in addition to other statistical parameters.

External Validation

A QSAR model must be not only able to learn from available data and reproduce them well (goodness of fit, verified by R^2), be stable or robust (verified by internal cross-validations: leave-one-out, leave-many-out or bootstrap, verified by corresponding Q^2), but, most importantly, must be also reliable in making predictions outside the training set.

If a QSAR cannot be used to make reliable predictions for new chemicals, then it is of no practical use. An optimistic paraphrase of the famous George Box sentence should be: "Most models are wrong, but some are useful." The key question is: useful for what specific purpose?

In the specific context of environmental chemicals, the field of my researches, a useful QSAR model must exploit at most the available information from the limited existing knowledge included in the training set, being finally able to reliably predict data for new chemicals, not involved in model

development in any step. To verify this, a thorough process of validation that includes external validation must be undertaken.

A rigorously validated QSAR model can help: a) in screening large data sets and discovering dangerous unknown properties of chemicals, b) in setting priorities for compounds that require deep *in vitro* and *in vivo* investigations, c) in planning better experiments for a more rational use of resources and limited animal tests, d) in designing safer alternatives in a Green Chemistry philosophy. If all these points are satisfied by a reliable QSAR model, that QSAR model is not just "useful," it is also "right" for all the cited purposes.

The check of real predictivity is obviously the most important and primary aspect of "predictive QSARs". The chemometric approach to QSAR modeling, always applied in my researches, is the "predictive" one and has a focus on validation. I was really surprised at my first QSAR meeting in Bulgaria (2000) to see that few models were statistically validated and I had interesting discussions with several researchers on this point. The discussions in that meeting have highly influenced my experience in QSAR. At the subsequent QSAR meeting in Ottawa (2002) I presented the steps of the chemometric approach that I applied in my models (Figure 1) and I asked to Alex Tropsha and Vijay Gombar, famous QSAR modellers, who had already expressed my same ideas on the strong need of external validation and rigorous check of the Applicability Domain to guarantee the reliability of QSAR models, to write our common paper: "The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models" (Tropsha, Gramatica & Gombar, 2003). Some years later, after my experience in collaborating to fix the OECD Principles, I dealt again with this important topic in two specific papers: "Principles of QSAR models validation: internal and external" (Gramatica, 2007), a next necessary clarification in a following paper (Gramatica, 2014), and then always in all my publications.

Model validation, or better verification/evaluation, has been, mainly since then, the subject of wide debate in the scientific and regulatory communities. The crucial point that internal validation is necessary, but not sufficient, and only external validation can guarantee predictivity is stressed in many papers, since the famous "Kubini paradox" (Kubini, 2002), among others the most cited: Golbraikh & Tropsha, 2002; Tropsha, Gramatica & Gombar, 2003; Gramatica, 2007. A lot of interesting scientific papers have been published with different underlying ideas on the "best" way to validate QSAR models, by using different tools (internal cross-validation, double cross-validation, external validation verified by different statistical parameters). I list only some examples here: Baumann & Baumann, 2014; Baumann & Stiefl 2004; Doweyko, 2008; Eriksson et al., 2003; Gramatica, 2014; Hawkins, 2004; Roy, 2007; Roy et al., 2011.

However, even after this wide debate, unfortunately a lot of authors use the word "predictions" for their calculated responses, obtained from models verified only by R^2. I hope that the erroneous use and abuse in these cases of the word "prediction" will disappear in the future in QSAR literature.

Several published models are validated only internally by cross-validation (reporting only Q_{2LOO}) or are not robust having too high difference between R^2 and Q^2. Even stronger internal validation, such as Leave-Many-Out (LMO) cross-validations, is not often applied.

It is important to highlight that sometimes in populations of models, developed using evolutionary techniques for the selection of the descriptors, it is not unusual to find models with high internal predictivity, verified by internal validation methods (LOO, LMO, Bootstrap), but externally less predictive or even absolutely not predictive. This was the core of my paper: "External Evaluation of QSAR Models, in Addition to Cross-Validation: Verification of Predictive Capability on Totally New Chemicals" (Gramatica, 2014). I proposed to name "prediction set" the set of chemicals used only once for external validation, to distinguish it from the sets of chemicals used in internal cross-validation, that is an iterative process: the "test" sets.

In the last years, different researchers have proposed different metrics to find "the best" statistical parameter to characterize the external predictivity of a QSAR model by various modification of Q^2 formula: Q^2_{F1} (Shi et al., 2001), Q^2_{F2} (Schuurmann et al., 2008; Alexander et al., 2015), Q^2_{F3} (Consonni

et al., 2009, 2010), or comparing in different ways the experimental values of chemicals in the prediction set against the corresponding values predicted by the model (Chirico & Gramatica, 2011, 2012; Golbraikh &Tropsha, 2002; Ojha et al., 2011; Roy, 2007; Roy et al., 2009). I don't repeat here the formulas of these statistical parameters, reported in several papers and summarized in my paper: "A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology" (Gramatica & Sangion, 2016), where I had tried to clarify some confusion on the formulas in the literature.

In the papers where I had proposed to use the Concordance Correlation Coefficient (CCC) of Lin (Lin, 1989, 1992) also for validation of QSAR models, a rigorous comparison of the behaviour of all the more frequently used parameters had been done in different situations of realistic and also extreme data by means of extensive simulations (Chirico & Gramatica, 2011, 2012). CCC measures both precision (how far the observations are from the regression line) and accuracy (how far the regression line deviates from the slope 1 line passing to the origin, i.e., the diagonal concordance line). Thus, CCC quantifies the similarity of the predicted and experimental values (the agreement) as a single criterion, achieving the same goal as the Golbraikh and Tropsha method, which, however, requires several conditions to be met. Drawbacks for few parameters in some situations and the need to always verify the distribution of points in the plot experimental vs predicted values have also been highlighted.

Moreover, this analysis allowed us to also propose, for the same data scatter, new inter-comparable thresholds for each criterion in defining a QSAR model as really externally predictive, in a more precautionary approach. We have verified that the inter-comparable values for good predictive models should be higher than 0.70 for Q^2F_n (the previously proposed values of 0.50 (Tropsha et al. 2003) is absolutely too low), corresponding to 0.65 for r2 m (0.50, proposed by Roy, should be raised), and to 0.85 for CCC. In our simulations (Chirico & Gramatica, 2011, 2012) we have showed that Q^2_{F3} and CCC are the most reliable and stable parameters in all the studied situations.

An analysis of the results revealed that the scatter plot of experimental vs predicted external data must always be evaluated to support the statistical criteria values: in fact, in some cases the publications of models reporting only high statistical parameter values could hide models with unacceptable predictions visible only in looking to the scatter plots.

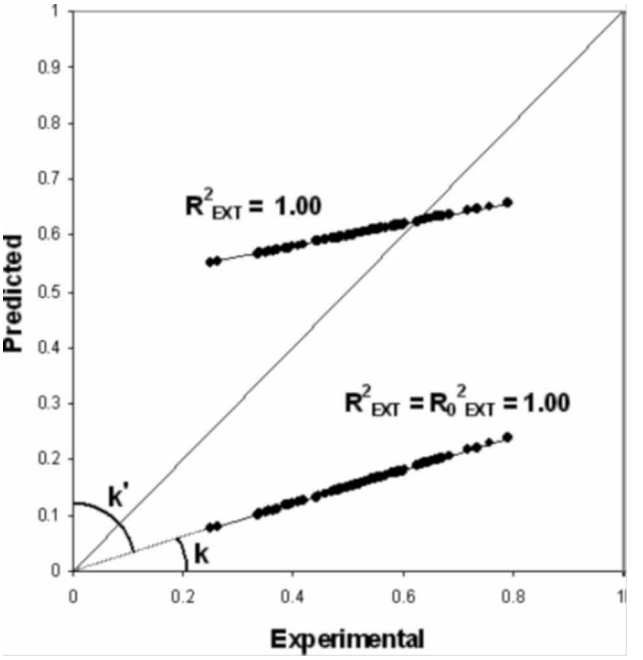
Moreover, as already said, RMSE and MAE on external chemicals, in comparison with the corresponding values on training set and LOO-CV, are other useful parameters for this check, but only various Q^2, r^2m and CCC parameters allow the comparison of the predictivity of models developed for different end points.

However, due to the demonstrated different behaviours of the various external validation parameters, and the tendency to give even contradictory results, especially on small data-sets, it is practically impossible to accept the superiority of one specific parameter for QSAR modeling or to select the best applicable in each specific situation. I'm convinced that no one of these Q^2_{EXT} is the "best" (while each proponent wishes to demonstrate the superior quality of their proposal, someone basing only on few examples); therefore, we have already suggested (Chirico & Gramatica, 2011, 2012) to verify and compare more than a single criterion to assess the real external predictivity of QSAR models (Gramatica & Sangion, 2016). For this reason, the calculation of all the above-reported validation criteria was implemented in QSARINS.

The demonstration that different external validation parameters can give different information on the real predictivity of a model (Chirico & Gramatica, 2011, 2012) requires that the specific parameter that has been calculated must be clearly reported. Unfortunately, in several publications the clear indication of which statistical parameter has been used to verify the external predictivity is not reported, simply indicating a generic R^2_{EXT} value. In this context, a crucial point that is useful to remember that, even though the linear relationship between experimental and predicted data can be perfect (the coefficient of determination on external set $R^2_{EXT} = 1$ and also $R_0{}^2_{EXT} = 1$, i.e., the regression line passing through the origin; see Figure 2 that shows extreme theoretical examples, luckily

not occurring in practice in good QSAR models), it does not automatically mean that the predicted data perfectly match the experimental ones (as already highlighted by Golbraikh & Tropsha, 2002). In fact, the data are well predicted only if they lie on the diagonal of the experimental vs predicted data in the scatter plot of Figure 2.

Figure 2. Examples of aligned external prediction data with R^2_{EXT} and/or $R_0{}^2_{EXT}$=1 for not predictive models (permission from American Chemical Society, JCIM)



If, in published papers, R^2_{EXT} (sometimes called R^2_{pred}), calculated from the coefficient of determination of the regression of observed on predicted values, is reported as the measure of predictivity, and visual inspection of the corresponding scatter plot of experimental vs predicted data is not possible, the reader should be not completely confident on the real external predictivity of the model. It is evident that correct statistical parameters, able to verify the real predictivity, must be calculated and reported.

In conclusion, after the wide debate of the last years, it is nowadays widely accepted (even if, unfortunately, not always) that, to be reliably applicable, a QSAR model must be not only able to reproduce well the training set (goodness of fit), be stable or robust (verified by internal cross-validations), but, most importantly, must be finally able to predict trustworthy data for new chemicals, not involved in model development in any step (external validation).

Splitting of Data for External Validation

But, how to validate externally?

The limiting problem for efficient external validation of a QSAR model is, obviously, data availability. Given the availability of a sufficiently large number (possibly no less than 20% of training) of blind new experimental data, the best proof of the accuracy of an already developed model is to test the model performance on these additional data, at the same time checking the AD. However, it is generally difficult to have data available (in useful quantity and quality) for new experimentally

tested compounds, thus, external validation can be usefully applied by QSAR modellers, after their model development and before its proposal, on chemicals that have been put aside before the modeling in what I call "prediction set" (to distinguish it from the "test sets" of iterative cross-validation).

Practically, the unique way is to split the complete set of the available data in a training set, which is used to select the modeling descriptors for model development, and an external prediction set (with known experimental response, and into the domain of the training set). This prediction set must be never used for variable selection, but is used exclusively once, and only after model development, for immediately verifying the reliability of the model predictions. Various procedures have been proposed to adequately split the available data set (Golbraikh et al., 2003; Leonard al., 2006).

Three different splitting methods are generally applied in my group: a) splitting by chemical structure similarity (Self Organizing Maps (SOM) or Kohonen map-ANN or PCA as in QSARINS) to guarantee a balance between structure representativity in the two sets (Prediction set I); b) splitting by response, thus, without any bias for structure (selection by response of compounds by arranging them in terms of ordered response, putting always the highest and lowest value in the training set, to cover the experimental range) (Prediction set II); c) random splitting, more similar to a potential reality, but with a higher possibility to produce more outliers out of the training domain (Prediction set III). The commonly applied splitting leaves 70-80% on training set for variable selection and model development and 20-30% on prediction sets for external validation after model development.

When the same common set of molecular descriptors is selected from each training sets, and is verified as predictive for different prediction sets, then this combination of descriptors is considered to be the best for modeling the response for the studied compounds, independently on the splitting criteria, thus unbiased of structure and response. Finally, this combined set of descriptors is used to derive a FULL model in order do not lose any available information: this QSAR model is our final proposal for future practical applications.

A particular effective example of rigorous external validation is a MLR model we developed of soil sorption coefficient (Koc) of 643 heterogeneous chemicals (mainly pesticides) (Gramatica et al., 2007) where only less than 15% of chemicals were used in training for the selection of four theoretical modelling descriptors and the high external predictivity, comparable with the fitting ($Q^2_{EXT(F1)}$= 0.79 in comparison with R^2=0.82), was verified on 550 chemicals, *a priori* split in the prediction set.

Selection of the "BEST" Validated Models

How to select the BEST model in a GA population of models?

I have already suggested selecting a model that, verified to be robust by internal validation, has also all the statistical parameters for external validation in agreement and with good values higher than their specific inter-comparable thresholds (defined in Chirico and Gramatica, 2012). My suggestion was also that the "best" QSAR model is the one which, from among the models with highest external predictivity, have the smallest difference between internal and external predictivity, RMSE and MAE values similar between training and prediction sets, and the fewest chemicals outside the chemical domain (both outliers for response with high residuals and structural outliers with high leverage value).

Another possibility, implemented in QSARINS, is to apply the Multi-Criteria Decision Method (MCDM). MCDM is a technique that takes into consideration and summarizes the performances of a certain number of different criteria simultaneously. In QSARINS, the MCDM of fitting (maximizing R^2 and R^2_{adj} and CCC_{TR}), cross validation (maximizing Q^2_{LOO}, Q^2_{LMO} and CCC_{cv} while minimizing $R^2_{Y\text{-}SCRAMBLE}$), external validation (maximizing Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and CCC_{EXT}) are automatically calculated using all the corresponding criteria. The best model is that with the best MCDM compromise among the selected validation criteria. Racz et al. (2015) have verified the good performances of this approach in selecting models with a good balance between fitting of the training set, robustness and external predictivity for chemicals not included in model development.

The statistical parameters here commented are those used in MLR, which is the core of this paper. For classification models, different parameters are calculated, such as those of the Cooper

statistics: Concordance or Accuracy, Sensitivity (true positive rate), Specificity (true negative rate), False Positive (over-classification rate), False Negative (under-classification rate). Also, the Receiver Operating Characteristic (ROC) curve can be usefully applied to visually compare the predictive abilities of different classification models.

Applicability Domain (AD) - The OECD Principle 3: "A Defined Domain of Applicability"

I have postponed the OECD Principle 3 on Applicability Domain (AD) after the OECD Principle 4 on validation, because, as already said, it is absolutely not useful, and also dangerous, to apply any QSAR model if is not preliminary deeply validated, internally for its robustness but also externally for predictivity. Thus, after the validation, and in particular if the aim is to apply the model in predicting response for new chemicals, it is highly important to remember that even a robust, significant, and validated QSAR model cannot be expected to reliably predict the modelled property for the entire universe of chemicals.

Applicability Domain (AD) of a QSAR model is a theoretical spatial region defined by the specific molecular descriptors of the model and the studied response and is thus defined completely by the nature of the chemicals in the training set, but also on the values of specific descriptors used in the model itself. In the context of reliable application of a QSAR model, given a peculiar training set, on which a model is constructed, it is generally felt that if a new molecule is somehow similar, or is in the "domain" or "space" of the training set, it is likely to be well-predicted as an interpolation, otherwise there is significant "extrapolation" and the prediction could be unreliable. Good prediction or possible unreliability must be verified on each compound.

A key aspect of the model applicability is the definition of the chemical space and the way to measure chemical similarity, as similarity is a relative concept. Similarity check depends on both the type of molecular representation (the descriptors in the model) and the similarity/distance measure used. Due to this lack of invariance of chemical space it is not unusual that two compounds that are neighbours in one representation of the chemistry-space (by the descriptors in each model) may not be close in another. For this reason, each model must be verified for its specific AD, which could be different even for the same set of chemicals.

So far there is no generally accepted or even standardized approach for defining the chemical space of QSAR models and there is no reason to state that a method is absolutely the best. However, even within these uncertainties, AD check must be always applied to make decision whether or not a QSAR prediction could be more or less reliable. This crucial and hot topic was dealt with at a JRC Workshop, where several different approaches for linear and non-linear models were proposed (Netzeva et al., 2005), in relation to different model typologies and is a topic in various publications (among others: Eriksson, 2003; Golbraikh & Tropsha, 2002; Roy et al., 2015a; Sahigara et al., 2012; Tropsha et al. 2003).

It is important to specify here that two different domains must be verified: one for the modelled end-point (Response AD), which is commonly and easily verified by the standardized residuals between experimental and predicted values in the training and prediction sets, and the other for the chemical structure (Structural AD). If the domain of the model is studied only for the chemicals in the training set it can be simply defined as the Model Domain (MD) that must be named Applicability Domain when the chemicals that are checked are external: or those in a split prediction set or those that are completely new entries.

In all my modelling works and in QSARINS the leverage approach for structural MD/AD is applied. A population of MLR models of similar good quality, developed by variable selection performed by Genetic Algorithm, can include hundreds of different models developed on the same training set but based on different descriptors: each model has its MD/AD that is dependent on the specific descriptors used in each model.

Leverage, based on Hat value of the descriptors' matrix diagonal in a regression model, represents a sort of compound "distance" from the chemical space of the model (in particular, from the structural centroid of the training set). In a study of Model Domain, the Hat value (h) of a training compound is a measures of the influence that a particular chemical's structure has on the model, in selecting specific descriptors useful to include it in the model. Training chemicals close to the centroid of the data sets (low h) are less influential in model building (in the descriptor selection) than extreme points. A chemical with high leverage (high h) in the training set could have great influence in the selection of the descriptors in the regression (good leverage): the fitted regression line will be forced near to the observed value and the residual (experimental-predicted value) is generally small, so the chemical does not appear to be an outlier for the response (it will be well predicted), even though it may actually be outside the structural MD. If a chemical is located outside the model structural domain a warning must be given.

The critical value for "warning leverage" is generally accepted as h*=3p_/n, where p_ is the number of model variables plus one and n is the number of the objects used to calculate the model). The training chemicals with h>h* must be always considered with attention because they are anomalous in comparison to the majority of the training chemicals. In my opinion, they could be eliminated, if one or few isolated, and the model redeveloped without them when the aim is to develop a more stable model on more similar chemicals (but probably the *a priori* structural PCA, which I had suggested as explorative analysis of the data set distribution, already highlighted them). On the contrary, if they are not few, they could be hold in the training set because they could be useful in enlarging the AD for future application to new chemicals.

In the selection of the best model in a population of models developed by GA, I suggested to choose the models with highest performances in validation checked by various statistical parameters in agreement, but also those with the lowest number of outliers (both for response and structure: high leverage). If the same compound appears as outlier for response in the majority of the models in GA population, based on different descriptors, the experimental value of this compound is dubious and an error could be present, thus I suggest deleting it from the training set. Similarly, if a same compound has high hat value in the majority of the models this is too anomalous respect the other compounds in the training set and too influent, causing instability. In QSARINS this check can be done for each model in GA population in order to have a better choose of the model to be proposed, taking into account also this point on the quality of chemicals in training set.

In the check for AD to new chemicals: if a chemical of the prediction set, or completely new, has a leverage value lower than the critical value (h<h*) it is into the structural domain of the model, the predicted data can be considered as interpolated and with reduced uncertainty, thus more reliable, because the probability of accordance between predicted and actual values is as high as that for the training set chemicals. Conversely, a high-leverage chemical (bad leverage) is structurally distant from the majority of training chemicals, thus it can be considered outside the structural AD of the model: the predicted data are extrapolated by the model and must be considered of increased uncertainty, and less reliable.

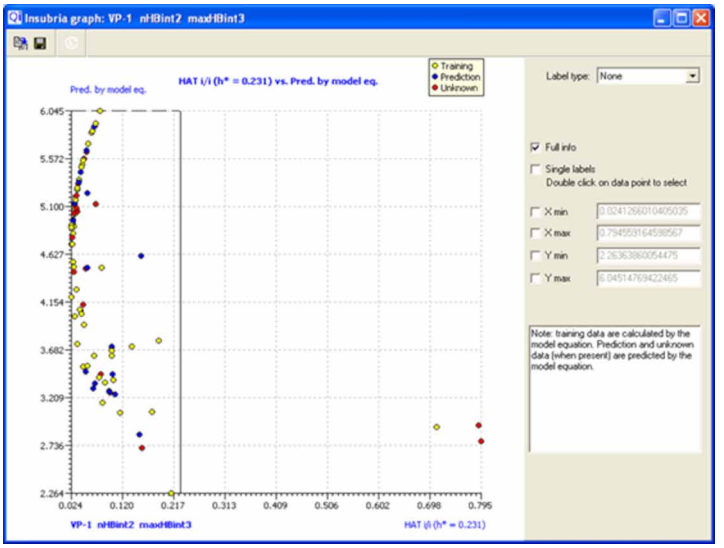
To visualize MD/AD of a QSAR model and identify chemicals that are outside the domain (both response and structural outliers) the Williams plot is a simple graphical detection on the Y-axis of the response outliers (i.e., compounds with cross-validated standardized residuals greater than 2.5-3 standard deviation units) and on X-Axis of structurally anomalous compounds with leverage values h >h*. I have always used the Williams plot since my first QSAR models, and since then, it is now widely applied by other authors and also in commercial software.

It is important to note that the AD of a model cannot be verified by studying only a few chemicals, as in such cases it is impossible to obtain reliable conclusions that can be generalized on the applicability of the model itself.

In the assessment of a new chemical for which a prediction can be made by applying a QSAR model, but for which there is no experimental value, the AD for the response can be only defined by

the range of the experimental data of the training, but it is obviously not possible to determine the standardised residuals for the new compounds. While, by the leverage approach, which is based only on structural descriptors, it is possible to identify which new compounds are structurally into the chemical space of the training set. This check on new chemicals without experimental data can be done by the Insubria graph, my proposal from a modification of the Williams plot, where the predicted data are plotted in Y axis and the leverage values in X. The Insubria graph (see Figure 3) is very useful

Figure 3. Example of Insubria graph for applicability domain on training (yellow), prediction set (blue) and chemicals with unknown experimental value (red)



for verifying *a priori* the AD of a QSAR model to any chemical, also without data, therefore also to chemicals not yet tested or synthesized, simply designing a planned structure (chemical design).

Activity Cliffs in QSAR and Read-Across

A big problem for QSAR modeling has been highlighted by Maggiora (Maggiora, 2006): very similar molecules may in some cases possess very different activities leading to what is called activity cliffs, so, significant mis-predictions of activity could arise among similar molecules, even in cases where overall predictivity of the proposed model is statistically high. Thus, perfectly valid data points located in cliff regions may appear to be outliers.

This big pitfall can be smoothed in QSAR modelling when big datasets are modelled for screening, because the information derives from many different structures and additionally outliers can be more easily identified, but it could be particularly dangerous in the read-across approach, where new data are predicted directly by interpolation from known data of very few chemicals, recognized as "similar" by different tools of categorization (structural similarity, mechanism similarity, etc.). In these cases, the recognized similarity on few compounds can hidden serious activity cliffs and produce wrong data that cannot be identified, with more probability than in QSAR models.

A Mechanistic Interpretation, if Possible: The OECD Principle 5

This is the most controversial Principle and the highest source of disagreement and discussions between mechanistic and statistical QSAR modellers, also in environmental field.

As it is well known, according to the Hansch approach, followed by the mechanistic modellers, the descriptor selection is guided by the modeller's conviction to have *a priori* knowledge of the mechanism of the studied activity/property, and the presumption to assign mechanistic meaning to any used molecular descriptor. The descriptors are personally selected from among a limited pool of potential modeling variables, normally well-known and frequently used (for instance: logKow used in a lot of toxicity models and various partition coefficients etc.; HOMO/LUMO always selected for modeling chemical reactivity, etc.). As said before, this is the "descriptive" approach.

On the other hand, the "statistical/predictive" approach is based on fundamental conviction that the descriptor selection should not be based on any *a priori* mechanistic assumptions of the QSAR modeller. Thus, the modeling descriptors should not be selected personally, but only by application of unbiased mathematical tools. The number and typology of the available input descriptors must be as wide and different as possible in order to guarantee the possibility of representing any aspect of the molecular structure. Rigorous validation, exclusion of chance correlation and check of Applicability Domain are mandatory for guaranteeing reliable predictivity on new chemicals.

Obviously, for the "mechanistic/descriptive" approach QSAR modeling must be mechanism-based, while for the "statistical/predictive" approach mechanistic interpretation is the last point in the QSAR modeling (Stanton, 2003; Tropsha et al., 2003; Gramatica, 2007). In my opinion, it is right that it is the last OECD Principle. But interpretation can be done only "If possible", as I have personally suggested at the OECD meeting in 2004, when these principles were fixed. In fact, it is not always possible (and often only a presumption) to understand the mechanism of the studied end-point from the modelling molecular descriptors. It is surely relevant to find the physical/structural relationship with the end point (Katritzky et al, 2008; Stanton, 2003), remembering always that S in QSAR means chemical Structure.

In relation to this point, Livingstone, in an interesting Perspective paper (Livingstone, 2000) states:

The need for interpretability depends on the application, since a validated mathematical model relating a target property to chemical features may, in some cases, be all that is necessary, though it is obviously desirable to attempt some explanation of the "mechanism" in chemical terms, but it is often not necessary, per se.

A crucial question I have on this point is: are we really sure to known the overall mechanism of any end-point of environmental interest?

Even in the simplest cases of physico-chemical properties, the mechanisms are quite complex and their understanding is only possible at certain levels of approximation. The biological response of one organism to a xenobiotic (such as for instance, skin sensitisation, carcinogenicity, mutagenicity, or ecotoxicological endpoints such as LD50 in different species, etc.) is the result of a combination of different biological processes that depends on several mechanisms of penetration, distribution, metabolism, and interactions of this chemical with organs, cells, receptors, etc. This complex behaviour is what is experimentally measured, in many cases, and condensed in a unique numerical value of activity, toxicity, etc.

How we could guarantee that a particular mechanism is the only one determining the studied biological response?

As the majority of the biological response measures are not specific for one particular mechanism, the selection of one (or more) physico-chemical variable *a priori* for their mechanistic meaning in relation to one assumed mechanism is very risky (Stanton, 2003), as we can ignore important variables that influence the other mechanisms participating in such a response.

As already discussed above, among those "mechanistic" descriptors that some authors prefer log P is a sort of "magic" descriptor, useful for modelling a lot of end-points. Lipophilicity plays certainly a crucial role in biological activities, but as well as electronic or steric effects. LogKow is highly correlated with a lot of dimensional descriptors, even with MW, as it is actually informative

on molecular size and can be frequently substituted by other structural descriptors. Estrada (Estrada & Patlewicz, 2004) has pointed out that there is no reason for assigning mechanistic meaning to logKow when other more complex theoretical descriptors are able to model in the same way an end-point, but are not accepted because considered "difficult to interpret mechanistically". Estrada says:

It is the consequence of the wrong application of the Occam's razor philosophy. The same that can conduce to deny general relativity because it makes more assumptions that Newton's gravitational law, and it is far more complex.

Luckily, the molecular structure can be represented in different ways, some more comprehensible for all, some others less, but all with a structural meaning, even if not always clearly understood by the users in their QSAR modelling. The molecular descriptors have normally detailed analysis of their structural meaning in specific papers of the researchers that had proposed them (normally physico-chemists), but these papers are not frequently accessed by some QSAR models' applicants. This is the main reason for which some easier descriptors are considered historically interpretable or "mechanistic" and some other defined alternatively as incomprehensible or not transparent.

However, as already highlighted before, if the models are correctly and deeply validated, molecular descriptors cannot be considered as random numbers. The descriptors selected in a model, even if not always clear for mechanistic meaning, encode useful information because are related to specific features of molecular structure which influence and control the modelled biological activity or physico-chemical property (Katritzky et al. 2008).

Moreover, it is important to highlight that, even if some descriptors are more obvious and easy interpretable, while others are more difficult to interpret, none of the descriptors could explain the observed response singularly and independently. Indeed, it is only the combination of all the descriptors selected in a model that is able to model the studied end-point and ensure the high overall predictive power of the models. This makes difficult to mechanistically interpret the specific individual role of each single descriptor in the QSAR model, where the descriptors "work" together in the description of different mechanisms simultaneously.

In conclusion and in my experience, the planned use of QSAR model predictions is one important aspect to take into consideration: for physico-chemical properties prediction and for virtual screening of heterogeneous chemicals the primary focus should be the predictive ability and mechanistic interpretations are secondary, even not necessary if the models are deeply validated.

QSAR MODELS FOR RANKING INDEXES: CUMULATIVE END-POINTS BY PCA

My researches on QSAR modeling, applied to organic chemicals that are /or could be of high concern for the environment for their already found/or potential hazard, have dealt various environmental topics modelled by regression and classification models: in addition to the already cited topics, mainly Endocrine Disruptors Chemicals (among others: Liu et al., 2006, 2007; Li & Gramatica, 2010a, b, c; Kovarich et al., 2011, 2012) and different end points for emerging pollutants studied by the European Project CADASTER: for instance, (benzo)-triazoles (Bhhatarai & Gramatica, 2011a; Cassani et al., 2013b), flame retardants (Kovarich et al., 2011; Papa et al., 2010), per-fluorinated compounds (Bhhatarai & Gramatica, 2011b, c; Kovarich, et al., 2012) and fragrances (Ceriani et al., 2015).

But, as environmental chemist, I'm convinced that the environment is a highly complex system in which many different variables, such as physico-chemical properties, chemical reactivity and biological activity of chemicals are of contemporaneous relevance. Most of these properties are in some way related to each other and it is only their cumulative effect that contributes to the environmental fate and biological effects of chemicals. Therefore, it is crucial to find an effective

way to understand, rationalize, and interpret the correlation and the covariance among the individual properties characterizing the studied system to have a holistic view of the problem.

The main peculiarity of my researches is the idea to model cumulative end-points and not only single specific response. In this complex context, multivariate statistical analysis methods are fundamental to extract the meaningful information from data. In particular, the application of explorative methods (such as for instance PCA) to various end points of environmental concern allows a combined view that generates ordination, and grouping, of the studied chemicals according to various properties, reactivities, or activities, analysed together, in addition to the discovering of relationships among the variables.

The scores of the most informative component (first Principal Component, PC1), i.e. the coordinates of chemicals along this component, can be used to rank compounds along the direction of maximum variance by linear combination of the information encoded into the variables used to feed the PCA. This linear combination describes a complex holistic behaviour of the studied system (i.e. it is a new macro-variable), which depends on the co-variances of the original variables and on their weight along each component.

The ranking of chemicals according to the studied properties, reactivities or activities contemporaneously along PC1, i.e. their relative position, was proposed as cumulative Ranking Indexes. Each Index, represented by the value of the PC1 score, condenses the main information related to the combined properties. If it explains a reasonably significant variance of the studied variables, it can be usefully modelled as a new macro-end-point by QSAR approaches. These models allow to predict behaviours of chemicals which depend on the variation of the cumulative index (for instance, overall persistence in the environment or overall aquatic toxicity), and not of each single variable (persistence in each different environmental medium or toxicity on only one aquatic organism).

In Figure 4 an example of definition and modeling of the overall Aquatic Toxicity Index (ATI) for pharmaceuticals is reported (Sangion & Gramatica, 2016a).

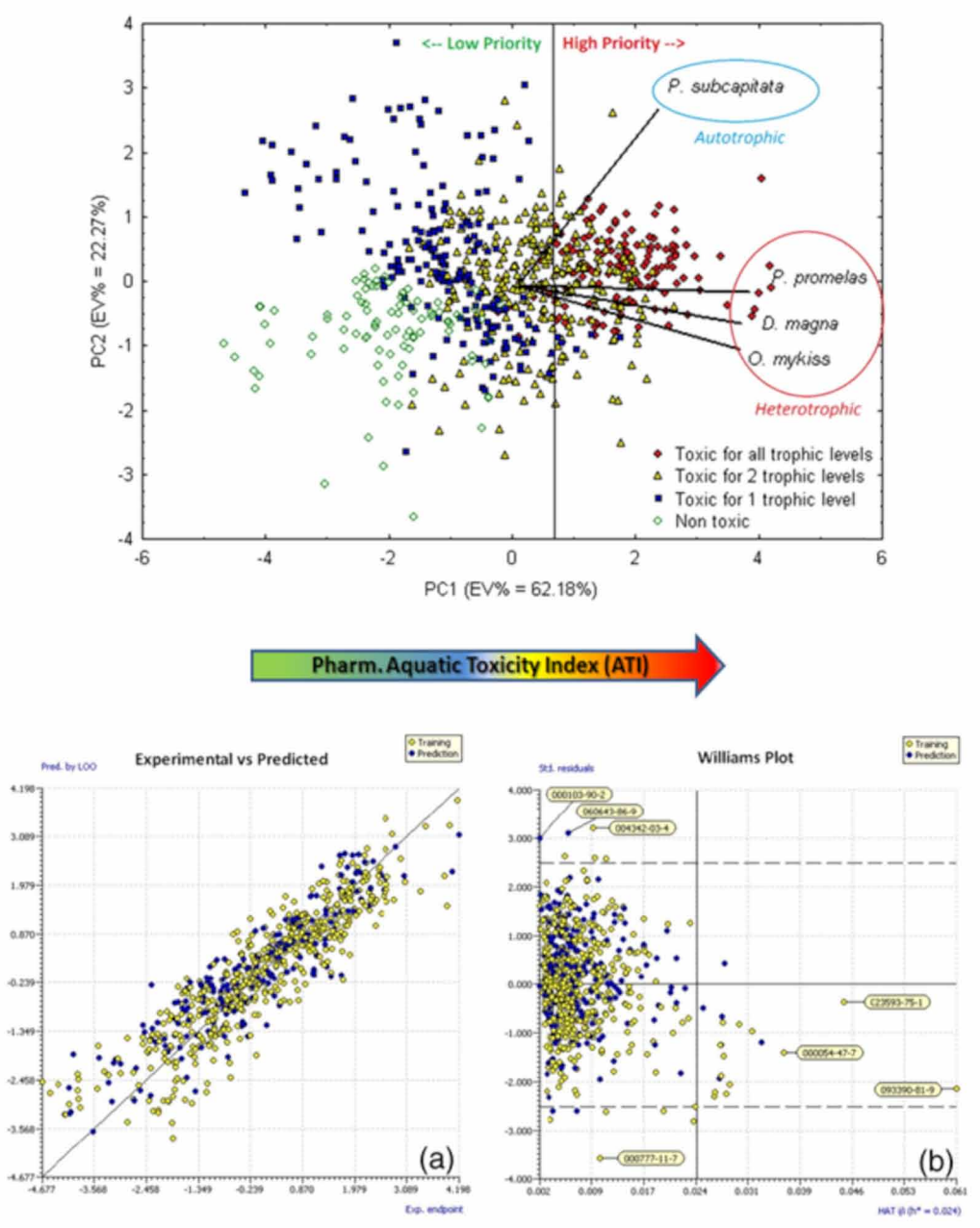
The QSAR modelling of various cumulative end-points is the central and peculiar core of most of my twenty-year research at Insubria University, reviewed also in book chapters (Gramatica, 2009, 2012, 2016) and in a recent paper (Gramatica et al., 2018).

I'll cite here only some of the most significant and/or recent papers reporting this kind of approach of cumulative Indexes: a) environmental partitioning of pesticides (Gramatica et al., 2004a; Leaching Index (LIN) in Gramatica & Di Guardo, 2002) and of (benzo)-triazoles (Bhhatarai & Gramatica, 2011a); b) degradability of Volatile Organic Compounds (VOCs) by tropospheric oxidants: Atmospheric Persistence Index (ATPIN) (Gramatica et al. 2004c); c) persistence of Persistent Organic Pollutants (POPs) by Global Half Life Index (GHLI), combining half-lives in 4 environmental compartments (Gramatica & Papa, 2007; Papa & Gramatica, 2008); d) rat/mouse toxicity of per-fluorinated compounds (PFCs) (Bhhatarai & Gramatica, 2010; Bhhatarai & Gramatica, 2011b); e) aquatic toxicity of Personal Care Products (PCPs) (Gramatica et al., 2016a) and pharmaceuticals (Sangion & Gramatica, 2016a) with definition of a global Aquatic Toxicity Index (ATI); f) PBT (Persistence, Bioaccumulation, Toxicity) screening of various compounds by Insubria PBT Index (Papa & Gramatica, 2010; Gramatica et al., 2015; Cassani & Gramatica, 2015; Gramatica et al., 2016b; Sangion & Gramatica, 2016b).

Some of the here listed QSAR models of cumulative ranking end-points, applied to several classes of chemicals of emerging concern (CEC), are implemented in the module QSARINS-Chem (Gramatica et al. 2014) for easy applicability on screening new chemicals.

Taking into account the fundamental assumption that the hazard of any chemical is an inherent property of the molecular structure and with prioritization aims, in my researches on proposing QSAR models of various ranking indexes I have showed that the fundamental information inherent in the chemical structure (the molecular descriptors in the QSAR models) can be useful in predicting a potential cumulative chemical hazard that derives from a contemporaneous combination of different end-points of concern (obtained by PCA). This prioritization, applied to chemicals without

Figure 4. Definition of the Aquatic Toxicity Index (ATI) of pharmaceuticals, plot of model for ATI (a); and Williams plot (b) (with permission of Elsevier, Environment International)



experimental data in screening big data sets, can allow to concentrate experiments on prioritized chemicals, thus reducing time, costs and animal test, but also to avoid the synthesis, and introduction to the market and into the environment, of harmful compounds, which could be recognized dangerous only after evidence of human health concerns have been manifested. This is the basis of the "benign by design" approach of Green Chemistry. In fact, the possibility to continuously contaminate the

environment with "regrettable substitutions" could be highly reduced if *a priori* screenings will be more widely applied. This was my dream when I started to apply QSAR to environmental topics.

REVIEWERS' ROLE

During my life, I have seen several published papers with big mistakes on QSAR modeling and this has serious consequences on the acceptability and practical utility of QSAR, mainly by not experts. Therefore, I believe that a great role in the publication of good QSAR models is that of Editors, which should select very carefully the reviewers with pertinent expertise. I receive often the requests for reviewing papers that are out of my specific expertise, even if in QSAR: I always decline to review these papers motivating with my not sufficient experience on the specific method dealt and suggesting more competent researchers. However, it is evident that some reviews are done by not really competent reviewers, otherwise many bad QSARs would not be published. Some absurd comments of reviewers, if accepted, give no useful improvement of the work, and often result in damage.

Obviously, the reviewers, if correctly selected by the Editors, have a most relevant role in improving the papers. They should review with great attention (obviously spending time…) and giving detailed suggestions for the best revision and improvement of the paper. I had reviewed hundreds of papers in my life (and this is now my main role) and I have often written many pages of suggestions. In my opinion, it is not useful to suggest simply the addition of some papers of the reviewer in the references: this is a sort of signature and is not useful for the improvement of the paper.

Moreover, a reviewer should respect the approach to QSAR modeling of the authors. As said at the beginning of this paper, it is evident that there are two completely different philosophies in approaching QSAR modeling: "predictive" selecting variables from a large descriptor pool and "mechanistic" using well known descriptors. If a QSAR model is well validated, it is reliable and will be useful, independently of the approach: thus, any reviewer must respect the authors' view. I had a really disappointing experience in my first years of QSAR modeling when a reviewer declined one of my papers, simply stating that other "good" models were already available for the same end-point and in particular because my proposed models on aquatic toxicity were not based on logKow: in fact, I had specifically demonstrated in that work that log Kow was not the best descriptor because other models, based on structural descriptors, had better performances. Luckily, I had then the opportunity to publish the same work on a journal with higher Impact Factor, where one of the reviewers wrote: "publish as it is." That paper has now many citations.

I stress again the point that both approaches, if correctly performed with rigorous model validation and check of AD for future applications, are useful and could also be used combining the relative predictions by consensus. This was always my idea that I had also presented firstly in the Final Round Table of QSAR 2012 in Tallin (Estonia) and finally in my last plenary in the Opening of QSAR 2018 in Bled (Slovenia) (Figure 5).

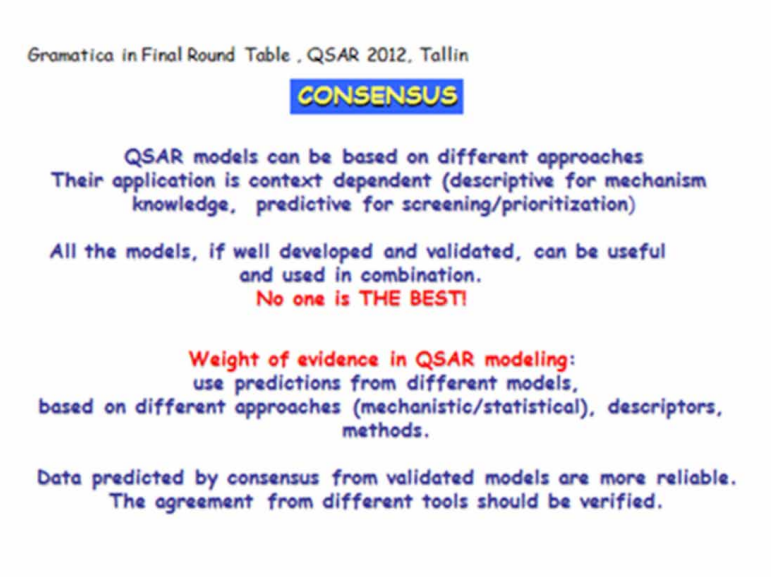
I was particularly happy to see that also Fujita and Winkler were on this same line, because a previous version of the paper "Understanding the roles of the "two QSARs."" (Fujita & Winkler, 2016) had the preliminary title: "Reconciling the "two QSARs"". This means also Consensus approach.

CONCLUSION

In this commentary paper, I have presented my ideas and suggestions to guarantee good development and validation of QSAR models, as well as some anecdotes based on my personal experience of "predictive approach" applied to environmental topics. At the end, it is a sort of excursus of "my life in QSAR".

I'm strongly convinced that QSAR modeling is a "fantastic" field with several useful applications, but I'm also sure that QSAR models must be applied with expertise.

Figure 5. Slide of Paola Gramatica during the final round table of 2012 QSAR meeting in Tallin (Estonia)



As expertise is requested in any field, why not be confident only in QSAR experts for QSAR models application?

The simplicity of use of some QSAR software can hide most of the problems that lie in the development of reliable QSAR models, with the consequence that QSAR is perceived by some non-experts as a sort of "magical" predictive approach applicable, without competent control, to any compound. Indeed, simply pushing a button, in any QSAR software that applies good or not so good models, a number always results. However, an inexperienced researcher would soon come to the conclusion that, if the value predicted by simply "pushing a button", but from an insufficiently validated model, is not in agreement with the observed experimental value, then the QSAR approach is, in general, not correct ("QSAR doesn't work well"). A conclusion reached just because the possible unreliability of a specific QSAR model prediction (for instance, due to extrapolation outside the model domain or for insufficient validation) has not been taken into consideration. Simply they are applying to the chemicals of their interest not validated QSAR models and out of their applicability domain, obtaining wrong predictions (just unreliable numbers) for their inexperience in using QSAR. They should apply better and more appropriate models to their molecules. For this reason, I'm afraid that the continuous diffusion of QSAR tools, in a "push a button" approach, with the aim to allow easy application of QSAR models, could cause not reliable applications by not experts. Caution should be applied, because QSAR models can be highly useful, but, on the contrary even dangerous, if not correctly developed and applied.

I have implemented all my ideas in QSARINS to leave this software for development and validation of OLS models as my legacy to QSAR community, but it is not "push a button" software: for this reason, it is not freely downloadable, but the license is distributed after mandatory approval from software owners.

In conclusion, I strongly believe in the great potentialities of QSAR modeling, but its correct development, validation and application are mandatory to guarantee utility and exclude any possible drawback.

ACKNOWLEDGMENT

I wish to thank all my collaborators in these years of my "second life" in QSAR modeling: Roberto Todeschini and Viviana Consonni of University of Milano-Bicocca, my heir Ester Papa and all my young fellows and coauthors at University of Insubria: Barun Bhhatarai, Stefano Cassani, Lidia Ceriani, Nicola Chirico, Elisa Giani, Simona Kovarich, Jiazhong Li, Huanxiang Liu, Mara Luini, Pamela Pilutti, Partha Pratim Roy, Alessandro Sangion.

REFERENCES

Alexander, D. L. J., Tropsha, A., & Winkler, D. A. (2015). Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling*, *55*(7), 1316–1322. doi:10.1021/acs.jcim.5b00206 PMID:26099013

Baumann, D., & Baumann, K. (2014). Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *Journal of Cheminformatics*, *6*(1), 47. doi:10.1186/s13321-014-0047-1 PMID:25506400

Baumann, K., & Stiefl, N. (2004). Validation tools for variable subset regression. *Journal of Computer-Aided Molecular Design*, *18*(7-9), 549–562. doi:10.1007/s10822-004-4071-5 PMID:15729854

Benfenati, E., Gini, G., Piclin, N., Roncaglioni, A., & Varı', M. R. (2003). Predicting logP of pesticides using different software. *Chemosphere*, *53*(9), 1155–1164. doi:10.1016/S0045-6535(03)00609-X PMID:14512120

Bhhatarai, B., & Gramatica, P. (2010). Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat and Mouse Using QSAR Modeling. *Chemical Research in Toxicology*, *23*(3), 528–539. doi:10.1021/tx900252h PMID:20095582

Bhhatarai, B., & Gramatica, P. (2011a). Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. *Water Research*, *45*(3), 1463–1471. doi:10.1016/j.watres.2010.11.006 PMID:21112604

Bhhatarai, B., & Gramatica, P. (2011b). Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse. *Molecular Diversity*, *15*(2), 467–476. doi:10.1007/s11030-010-9268-z PMID:20803170

Bhhatarai, B., & Gramatica, P. (2011c). Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. *Environmental Science & Technology*, *45*(19), 8120–8128. doi:10.1021/es101181g PMID:20958003

Bhhatarai, B., Teetz, W., Liu, T., Oberg, T., Jeliazkova, N., Kochev, N., & Gramatica, P. et al. (2011). CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. *Molecular Informatics*, *30*(2-3), 189–204. doi:10.1002/minf.201000133 PMID:27466773

Cassani, S., & Gramatica, P. (2015). Identification of potential PBT behavior of personal care products by structural approaches. *Sustainable Chemistry and Pharmacy*, *1*, 19–27. doi:10.1016/j.scp.2015.10.002

Cassani, S., Kovarich, S., Papa, E., Roy, P. P., Rahmber, M., Nilsson, S., & Gramatica, P. et al. (2013a). Evaluation of CADASTER QSAR models for aquatic toxicity of (benzo-)triazoles and prioritization by consensus. [ATLA]. *Alternatives to Laboratory Animals*, *41*(1), 49–64. doi:10.1177/026119291304100107 PMID:23614544

Cassani, S., Kovarich, S., Papa, E., Roy, P. P., van der Wal, L., & Gramatica, P. (2013b). Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity-activity modelling. *Journal of Hazardous Materials*, *258-259*, 50–60. doi:10.1016/j.jhazmat.2013.04.025 PMID:23702385

Ceriani, L., Papa, E., Kovarich, S., Boethling, R., & Gramatica, P. (2015). Modeling Ready Biodegradability of Fragrance Materials. *Environmental Toxicology and Chemistry*, *34*(6), 1224–1231. doi:10.1002/etc.2926 PMID:25663647

Chai, T., & Draxler, R. R. (2014). Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Arguments against Avoiding RMSE in the Literature. *Geoscientific Model Development*, *7*(3), 1247–1250. doi:10.5194/gmd-7-1247-2014

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., & Tropsha, A. et al. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, *57*(12), 4977–5010. doi:10.1021/jm4004285 PMID:24351051

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. *Journal of Chemical Information and Modeling*, *51*(9), 2320–2335. doi:10.1021/ci200211n PMID:21800825

Chirico, N., & Gramatica, P. (2012). Real External Predictivity of QSAR Models. Part 2. New inter-comparable thresholds for different validation criteria and the need for scatter plot inspection. *Journal of Chemical Information and Modeling*, *52*(8), 2044–2058. doi:10.1021/ci300084j PMID:22721530

Consonni, V., Ballabio, D., & Todeschini, R. (2009). Comments on the Definition of the Q(2) Parameter for QSAR Validation. *Journal of Chemical Information and Modeling*, *49*(7), 1669–1678. doi:10.1021/ci900115y PMID:19527034

Consonni, V., Ballabio, D., & Todeschini, R. (2010). Evaluation of Model Predictive Ability by External Validation Techniques. *Journal of Chemometrics*, *24*(3-4), 194–201. doi:10.1002/cem.1290

Cronin, M. T. D., & Schultz, T. W. (2003). Pitfalls in QSAR. *Journal of Molecular Structure THEOCHEM*, *622*(1-2), 39–51. doi:10.1016/S0166-1280(02)00616-4

Dearden, J.C. (2016). The history and development of quantitative structure-activity relationships (QSARs). *International Journal of Quantitative Structure-Property Relationships, 1*(1), 1-44. doi:l0.40IS/IJQSPR.201601010t.

Dearden, J. C., Cronin, M. T. D., & Kaiser, K. L. E. (2009). How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research*, *20*(3-4), 241–266. doi:10.1080/10629360902949567 PMID:19544191

Doweyko, A. M. (2008). QSAR: Dead or alive. *Journal of Computer-Aided Molecular Design*, *22*(2), 81–89. doi:10.1007/s10822-007-9162-7 PMID:18189157

DRAGON for Windows (Software for Molecular Descriptor Calculations) v.5.5, Todeschini, R., Consonni, V., Mauri, A., Pavan, M. 2007, Talete srl, Milan, Italy. Retrieved from http://www.talete.mi.it

Ecological Structure Activity Relationships (ECOSAR) Predictive Model. Retrieved from: https://www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model

EPI Suite™-Estimation Program Interface. (n.d.). Retrieved from https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface

Eriksson, L., Jaworska, J., Worth, A., Cronin, M., McDowell, R. M., & Gramatica, P. (2003). Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs. *Environmental Health Perspectives*, *111*(10), 1361–1375. doi:10.1289/ehp.5758 PMID:12896860

Estrada, E., & Patlewicz, G., G. (2004). On the usefulness of graph-theoretic descriptors in predicting theoretical parameters. Phototoxicity of polycyclic aromatic hydrocarbons (PAHs). *Croatica Chemica Acta*, *77*(1-2), 203–211.

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, *50*(7), 1189–1204. doi:10.1021/ci100176x PMID:20572635

Fujita, T., & Winkler, D. A. (2016). Understanding the roles of the "two QSARs." *Journal of Chemical Information and Modeling*, *56*(2), 269–274. doi:10.1021/acs.jcim.5b00229 PMID:26754147

Gadaleta, D., Lombardo, A., Toma, C., & Benfenati, E. (2018). A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. *Journal of Cheminformatics*, *10*(60), 1–13. doi:10.1186/s13321-018-0315-6 PMID:30536051

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y. D., Lee, K. H., & Tropsha, A. (2003). Rational selection of training sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design*, *17*(2), 241–253. doi:10.1023/A:1025386326946 PMID:13677490

Golbraikh, A., & Tropsha, A. (2002). Beware of q(2)! *Journal of Molecular Graphics & Modelling*, *20*(4), 269–276. doi:10.1016/S1093-3263(01)00123-1 PMID:11858635

Gramatica, P. (2007). Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science, 26*(5), 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. In T. Puzyn, J. Leszczynski, & M. T. D. Cronin (Eds.), *Recent Advances in QSAR Studies* (pp. 327–366). New York, USA: Springer-Verlag Pub.

Gramatica, P. (2012). Modeling Chemicals in the Environment. In D. J. Livingstone & A. M. Davies (Eds.), *Drug Design Strategies-Quantitative Approaches* (pp. 458–478). UK: Royal Society of Chemistry Pub.

Gramatica, P. (2013). On the Development and Validation of QSAR Models. In B. Reisfeld & A. N. Mayeno (Eds.), Computational Toxicology: Volume II, Methods in Molecular Biology, vol. 930 (pp. 499-526). New York, USA: Springer Science+Business Media. doi:10.1007/978-1-62703-059-5_21

Gramatica, P. (2014). External Evaluation of QSAR Models, in Addition to Cross-Validation: Verification of Predictive Capability on Totally New Chemicals. *Molecular Informatics*, *33*(4), 311–314. doi:10.1002/minf.201400030 PMID:27485777

Gramatica, P. (2016). Prioritization of chemicals based on chemoinformatic analysis. In J. Leszczynski & T. Puzyn (Eds.), *Handbook of Computational Chemistry* (Vol. 5, pp. 1–33). Netherlands: Springer Science Pub.; doi:10.1007/978-94-007-6169-8_58-1

Gramatica, P., Cassani, S., & Chirico, N. (2014). QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS. *Journal of Computational Chemistry*, *35*(13), 1036–1044. doi:10.1002/jcc.23576 PMID:24599647

Gramatica, P., Cassani, S., Roy, P. P., Kovarich, S., Yap, C. W., & Papa, E. (2012). QSAR Modeling Is Not "Push a Button and Find a Correlation": A Case Study of Toxicity of (benzo-)triazoles on Algae. *Molecular Informatics*, *31*(11-12), 817–835. doi:10.1002/minf.201200075 PMID:27476736

Gramatica, P., Cassani, S., & Sangion, A. (2015). PBT Assessment and Prioritization by PBT Index and Consensus Modeling: Comparison of Screening Results from Structural Models. *Environment International*, *77C*, 25–34. doi:10.1016/j.envint.2014.12.012 PMID:25617903

Gramatica, P., Cassani, S., & Sangion, A. (2016a). Aquatic Ecotoxicity of Personal Care Products: QSAR models and ranking for prioritization and safer alternatives' design. *Green Chemistry*, *18*(16), 4393–4406. doi:10.1039/C5GC02818C

Gramatica, P., Cassani, S., & Sangion, A. (2016b). Are some "safer alternatives" hazardous as PBTs? The case study of new flame retardants. *Journal of Hazardous Materials*, *306*, 237–246. doi:10.1016/j.jhazmat.2015.12.017 PMID:26742016

Gramatica, P., Chirico, N., Papa, E., Cassani, S., & Kovarich, S. (2013). QSARINS: A new software for the development, analysis and validation of QSAR MLR models. *Journal of Computational Chemistry*, *34*(24), 2121–2132. doi:10.1002/jcc.23361

Gramatica, P., & Di Guardo, A. (2002). Screening of pesticides for environmental partitioning tendency. *Chemosphere*, *47*(9), 947–956. doi:10.1016/S0045-6535(02)00007-3 PMID:12108701

Gramatica, P., Giani, E., & Papa, E. (2007). Statistical external validation and consensus modeling: A QSPR case study for Koc prediction. *Journal of Molecular Graphics & Modelling*, *25*(6), 755–766. doi:10.1016/j.jmgm.2006.06.005 PMID:16890002

Gramatica, P., & Papa, E. (2003). QSAR Modeling of Bioconcentration Factor by Theoretical Molecular Descriptors. *QSAR & Combinatorial Science*, *22*(3), 374–385. doi:10.1002/qsar.200390027

Gramatica, P., & Papa, E. (2005). An update of the BCF QSAR model based on theoretical molecular descriptors. *QSAR & Combinatorial Science*, *24*(8), 953–960. doi:10.1002/qsar.200530123

Gramatica, P., & Papa, E. (2007). Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure. *Environmental Science & Technology*, *41*(8), 2833–2839. doi:10.1021/es061773b PMID:17533846

Gramatica, P., Papa, E., & Battaini, B. (2004a). Ranking and classification of non-ionic organic pesticides for environmental distribution: A QSAR approach. *International Journal of Environmental Analytical Chemistry*, *84*(1-3), 65–74. doi:10.1080/0306731031000149732

Gramatica, P., Papa, E., & Sangion, A. (2018). QSAR modeling of cumulative environmental end-points for the prioritization of hazardous chemicals. *Environmental Science: Processes & Impact* [Special Issue]. *Modeling in Environment*, *20*(1), 38–47. doi:10.1039/c7em00519a PMID:29226926

Gramatica, P., Pilutti, P., & Papa, E. (2004b). Validated QSAR Prediction of OH Tropospheric degradability: Splitting into training-test set and consensus modeling. *Journal of Chemical Information and Computer Sciences*, *44*(5), 1794–1802. doi:10.1021/ci049923u PMID:15446838

Gramatica, P., Pilutti, P., & Papa, E. (2004c). A tool for the assessment of VOC degradability by tropospheric oxidants starting from chemical structure. *Atmospheric Environment*, *38*(36), 6167–6175. doi:10.1016/j.atmosenv.2004.07.026

Gramatica, P., & Sangion, A. (2016). A historical excursus on the statistical validation parameters for QSAR models: A clarification concerning metrics and terminology. *Journal of Chemical Information and Modeling*, *56*(6), 1127–1131. doi:10.1021/acs.jcim.6b00088 PMID:27218604

Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models, (2007) OECD Environ. Health and Safety Pub., (2). Retrieved from https://www.oecd.org/env/guidance-document-on-the-validation-of-quantitative-structure-activity-relationship-q-sar-models-9789264085442-en.htm

Hanley, J. A. (1989). Receiver Operating Characteristic (ROC) Methodology: The State of the Art. *Critical Reviews in Diagnostic Imaging*, *29*, 307–335. PMID:2667567

Haupt, R. L., & Haupt, S. E. (2004). *Practical Genetic Algorithms*. New Jersey: Wiley-Interscience.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. doi:10.1021/ci0342472 PMID:14741005

HyperChem(TM), 2002. Hypercube, Inc., Gainesville,Florida 32601, USA.

JRC QSAR model database. (n.d.). Retrieved from http://qsardb.jrc.ec.europa.eu/qmrf

Kaiser, K. L. E. (2003). The use of neural networks in QSARs for acute aquatic toxicological endpoints. *Journal of Molecular Structure THEOCHEM*, *622*(1-2), 85–95. doi:10.1016/S0166-1280(02)00620-6

Katritzky, A. R., Dobchev, D. A., Slavov, S., & Karelson, M. (2008). Legitimate Utilization of Large Descriptor Pools for QSPR/QSAR Models. *Journal of Chemical Information and Modeling*, *48*(11), 2213. doi:10.1021/ci8002073 PMID:18956833

Kovarich, S., Ester Papa, E., & Gramatica, P. (2011). QSAR classification models for the prediction of endocrine disrupting activity of brominated flame retardants. *Journal of Hazardous Materials*, *190*(1-3), 106–112. doi:10.1016/j.jhazmat.2011.03.008 PMID:21454014

Kovarich, S., Papa, E., Li, J., & Gramatica, P. (2012). QSAR classification models for the screening of the Endocrine Disrupting activity of perfluorinated compounds. *SAR and QSAR in Environmental Research*, *23*(3-4), 207–220. doi:10.1080/1062936X.2012.657235 PMID:22352429

Kovatcheva, A., Golbraikh, A., Oloff, S., Xiao, Y., Zheng, W., Wolschann, P., & Tropsha, A. et al. (2004). Combinatorial QSAR of Ambergris Fragrance Compounds. *Journal of Chemical Information and Computer Sciences*, *44*(2), 582–595. doi:10.1021/ci034203t PMID:15032539

Kubini, H. (2002). From Narcosis to Hyperspace: The History of QSAR. *Quantitative Structure-Activity Relationships*, *21*(4), 348–356. doi:10.1002/1521-3838(200210)21:4<348::AID-QSAR348>3.0.CO;2-D

Leonard, J. T., & Roy, K. (2006). On Selection of Training and Test Sets for the Development of Predictive QSAR models. *QSAR & Combinatorial Science*, *25*(3), 235–251. doi:10.1002/qsar.200510161

Li, J., & Gramatica, P. (2010a). The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders. *Molecular Diversity*, *14*(4), 687–696. doi:10.1007/s11030-009-9212-2 PMID:19921452

Li, J., & Gramatica, P. (2010b). Classification and Virtual Screening of Androgen Receptor Antagonists with Various Methods and Consensus Approach. *Journal of Chemical Information and Modeling*, *50*(5), 861–874. doi:10.1021/ci100078u PMID:20405856

Li, J., & Gramatica, P. (2010c). QSAR classification of estrogen receptor binders and identification of pleiotropic EDCs. *SAR and QSAR in Environmental Research*, *21*(7), 657–669. doi:10.1080/1062936X.2010.528254 PMID:21120754

Lin, L. I. K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, *45*(1), 255–268. doi:10.2307/2532051 PMID:2720055

Lin, L. I. K. (1992). Assay Validation Using the Concordance Correlation Coefficient. *Biometrics*, *48*(2), 599–604. doi:10.2307/2532314

Liu, H., Papa, E., & Gramatica, P. (2006). QSAR Prediction of Estrogen Activity for a Large Set of Diverse Chemicals under the Guidance of OECD Principles.

Liu, H., Papa, E., & Gramatica, P. (2006, November). QSAR Prediction of Estrogen Activity for a Large Set of Diverse Chemicals under the Guidance of OECD Principles. *Chemical Research in Toxicology*, *19*(11), 1540–1548. doi:10.1021/tx0601509

Liu, H., Papa, E., John Walker, J., & Paola Gramatica, P. (2007). In silico Screening of Estrogen-Like Chemicals Based on Different Nonlinear Classification Models. *Journal of Molecular Graphics & Modelling*, *26*(1), 135–144. doi:10.1016/j.jmgm.2007.01.003

Livingstone, D. J. (1995). *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*. Oxford, England: Oxford University Press.

Livingstone, D. J. (2000). The characterization of chemical structures using molecular properties. A survey. *Journal of Chemical Information and Computer Sciences*, *40*(2), 195–209. doi:10.1021/ci990162i PMID:10761119

Livingstone, D. J. (2004). Building QSAR models: a practical guide. In M. T. D. Cronin & D. J. Livingstone (Eds.), Predicting chemical toxicity and fate (pp. 151-170). Boca Raton, FL: CRC Press. doi:10.1201/9780203642627.ch7

Maggiora, G. M. (2006). On outliers and activity cliffs-Why QSAR often disappoints. *Journal of Chemical Information and Modeling*, *46*(4), 1535–1535. doi:10.1021/ci060117s PMID:16859285

Mansouri, K., Grulke, C. M., Richard, A. M., Judson, R. S., & Williams, A. J. (2016). An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. *SAR and QSAR in Environmental Research*, *27*(11), 911–937. doi:10.1080/1062936X.2016.1253611 PMID:27885862

Netzeva, T. I., Worth, A., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., & Yang, C. et al. (2005). Current Status of Methods for Defining the Applicability Domain of (quantitative) Structure-Activity Relationships. Report and Recommendations of ECVAM Workshop 52. [ATLA]. *Alternatives to Laboratory Animals*, *33*(2), 155–173. doi:10.1177/026119290503300209

OECD principles for the validation, for regulatory purposes, of (Q)SAR models. (2004). OECD. Retrieved from https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm

Ojha, P. K., Mitra, I., Das, R. N., & Roy, K. (2011). Further exploring r2m metrics for validation of QSPR models. *Chemometrics and Intelligent Laboratory Systems*, *107*(1), 194–205. doi:10.1016/j.chemolab.2011.03.011

PaDEL-Descriptors. (n.d.). Retrieved from: http://padel.nus.edu.sg/software/padeldescriptor/index.html

Papa, E., & Gramatica, P. (2008). Screening of persistent organic pollutants by QSPR classification models: A comparative study. *Journal of Molecular Graphics & Modelling*, *27*(1), 59–65. doi:10.1016/j.jmgm.2008.02.004 PMID:18387326

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 Chemistry*, *12*(5), 836–843. doi:10.1039/b923843c

Papa, E., Kovarich, S., & Gramatica, P. (2010). QSAR modeling and prediction of the endocrine disrupting potencies of brominated flame retardants. *Chemical Research in Toxicology*, *23*(5), 946–954. doi:10.1021/tx1000392 PMID:20408563

Papa, E., Kovarich, S., & Gramatica, P. (2011). On the Use of Local and Global QSPRs for the Prediction of Physico-chemical Properties of Polybrominated Diphenyl Ethers. *Molecular Informatics*, *30*(2-3), 232–240. doi:10.1002/minf.201000148 PMID:27466776

Papa, E., van der Wal, L., Arnot, J. A., & Gramatica, P. (2014). Metabolic biotransformation half-lives in fish: QSAR modeling and consensus analysis. *The Science of the Total Environment*, *470*, 1040–1046. doi:10.1016/j.scitotenv.2013.10.068 PMID:24239825

Papa, E., Villa, F., & Gramatica, P. (2005). Statistically validated QSARs and theoretical descriptors for the modelling of the aquatic toxicity of organic chemicals in *Pimephales promelas* (Fathead Minnow). *Journal of Chemical Information and Modeling*, *45*(5), 1256–1266. doi:10.1021/ci050212l PMID:16180902

Puzyn, T., Gajewicz, A., Rybacka, A., & Haranczyk, M. (2011). Global versus local QSPR models for persistent organic pollutants: Balancing between predictivity and economy. *Structural Chemistry*, *22*(4), 873–884. doi:10.1007/s11224-011-9764-5

QSAR Application Toolbox. (n.d.). Retrieved from https://qsartoolbox.org

QsarDB. (n.d.). QSAR Data Bank of University of Tartu. Retrieved from https://qsardb.org/

Racz, A., Bajusz, D., & Heberger, K. (2015). Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters. *SAR and QSAR in Environmental Research*, *26*(7-9), 683–700. doi:10.1080/1062936X.2015.1084647 PMID:26434574

Raevsky, O. A., Polianczyk, D. E., Grigorev, V. Y., Raevskaja, O. E., & Dearden, J. C. (2015). In silico Prediction of Aqueous Solubility: a Comparative Study of Local and Global Predictive Models. *Molecular Informatics*, *34*(6-7), 417–430. doi:10.1002/minf.201400144 PMID:27490387

REACH. European REACH Regulation. (2012). Retrieved from: https://ec.europa.eu/environment/chemicals/reach/legislation_en.htm

Ren, S. (2003). Ecotoxicity prediction using mechanism-and non-mechanism-based QSARs: A preliminary study. *Chemosphere*, *53*(9), 1053–1065. doi:10.1016/S0045-6535(03)00573-3

Renner, R. (2002). The Kow controversy. Doubts about the quality of basic physicochemical data for hydrophobic organic compounds could be undermining many environmental models and assessments. Environmental Science & Technology, 36(21), 410A-413A. doi:10.1021/es022457+

Roy, K. (2007). On Some Aspects of Validation of Predictive Quantitative Structure-Activity Relationship Models. *Expert Opinion on Drug Discovery*, *2*(12), 1567–1577. doi:10.1517/17460441.2.12.1567 PMID:23488901

Roy, K., Das, R. N., Ambure, P., & Aher, R. B. (2016). Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *152*, 18–33. doi:10.1016/j.chemolab.2016.01.008

Roy, K., Kar, S., & Ambure, P. (2015a). On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *145*, 22–29. doi:10.1016/j.chemolab.2015.04.013

Roy, K., Kar, S., & Das, R. N. (2015b). *A primer on QSARlQSPR modeling. Fundamental concepts*. Heidelberg, Germany: Springer.

Roy, K., Kar, S., & Das, R. N. (2015c). *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. London, UK: Academic Press.

Roy, P. P., Kovarich, S., & Gramatica, P. (2011). QSAR model reproducibility and applicability: A case study of rate constants of hydroxy radical reaction models applied to Polybrominated Diphenyl Ethers and (Benzo-) Triazoles. *Journal of Computational Chemistry*, *32*(11), 2386–2396. doi:10.1002/jcc.21820 PMID:21541967

Roy, P. P., Paul, S., Mitra, I., & Roy, K. (2009). On Two Novel Parameters for Validation of Predictive QSAR Models. *Molecules*, *14*(5), 1660–1701. doi:10.3390/molecules14051660 PMID:19471190

Rucker, C., Rucker, G., & Meringer, M. (2007). y-Randomization and Its Variants in QSPR/QSAR. *Journal of Chemical Information and Modeling*, *47*(6), 2345–2357. doi:10.1021/ci700157b PMID:17880194

Ruusmann, V., & Maran, U. (2013). From data point timelines to a well curated data set, data mining of experimental data and chemical structure data from scientific articles, problems and possible solutions. *Journal of Computer-Aided Molecular Design*, *27*(7), 583–603. doi:10.1007/s10822-013-9664-4 PMID:23884706

Ruusmann, V., Sild, S., & Maran, U. (2015). QSAR DataBank repository: Open and linked qualitative and quantitative structure–activity relationship models. *Journal of Cheminformatics*, *7*(1), 32. doi:10.1186/s13321-015-0082-6 PMID:26110025

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, *17*(5), 4791–4810. doi:10.3390/molecules17054791 PMID:22534664

Sangion, A., & Gramatica, P. (2016a). Hazard of pharmaceuticals for aquatic environment: Prioritization by structural approaches and prediction of ecotoxicity. *Environment International*, *95*, 131–143. doi:10.1016/j.envint.2016.08.008 PMID:27568576

Sangion, A., & Gramatica, P. (2016b). PBT assessment and prioritization of Contaminants of Emerging Concern: Pharmaceuticals. *Environmental Research*, *147*, 297–306. doi:10.1016/j.envres.2016.02.021 PMID:26921826

Schüurmann, G., Ebert, R.-U., Chen, J., Wang, B., & Kuehne, R. (2008). External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. *Journal of Chemical Information and Modeling*, *48*(11), 2140–2145. doi:10.1021/ci800253u PMID:18954136

Scior, T., Medina-Franco, J. L., Do, Q. T., Martinez-Mayorga, K., Rojas, J. A. Y., & Bernard, P. (2009). How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Current Medicinal Chemistry*, *16*(32), 4297–4313. doi:10.2174/092986709789578213 PMID:19754417

Shi, L. M., Fang, H., Tong, W. D., Wu, J., Perkins, R., Blair, R. M., & Sheehan, D. M. et al. (2001). QSAR Models Using a Large Diverse Set of Estrogens. *Journal of Chemical Information and Computer Sciences*, *41*(1), 186–195. doi:10.1021/ci000066d PMID:11206373

Stanton, D. T. (2003). On the physical interpretation of QSAR models. *Journal of Chemical Information and Computer Sciences*, *43*(5), 1423–1433. doi:10.1021/ci0340658 PMID:14502475

Todeschini, R., & Consonni, V. (2000). *Handbook of molecular descriptors*. Weinheim, Germany: Wiley-VCH. doi:10.1002/9783527613106

Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*. Weinheim, Germany: Wiley-VCH. doi:10.1002/9783527628766

Todeschini, R., Consonni, V., & Maiocchi, A. (1999). The K Correlation Index: Theory Development and its Application in Chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *46*(1), 13–29. doi:10.1016/S0169-7439(98)00124-5

Topliss, J. G., & Costello, R. J. (1972). Chance correlations in structure-activity studies using multiple regression analysis. *Journal of Medicinal Chemistry*, *15*(10), 1066–1068. doi:10.1021/jm00280a017 PMID:5069775

Topliss, J. G., & Edwards, R. P. (1979). Chance factors in studies of quantitative structure-activity relationships. *Journal of Medicinal Chemistry*, *22*(10), 1238–1244. doi:10.1021/jm00196a017 PMID:513071

Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, *29*(6-7), 476–488. doi:10.1002/minf.201000061 PMID:27463326

Tropsha, A., & Golbraikh, A. (2007). Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Current Pharmaceutical Design*, *13*(34), 3494–3504. doi:10.2174/138161207782794257 PMID:18220786

Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science*, *22*(1), 69–77. doi:10.1002/qsar.200390007

Waldman, M., Fraczkiewicz, R., & Clark, R. D. C. (2015). Tales from the war on error: The art and science of curating QSAR data. *Journal of Computer-Aided Molecular Design*, *29*(9), 897–910. doi:10.1007/s10822-015-9865-0 PMID:26290258

Yap, C. W. (2011). PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466–1474. doi:10.1002/jcc.21707 PMID:21425294

Young, D., Martin, T., Venkatapathy, R., & Harten, P. (2008). Are the chemical structures in your QSAR correct? *QSAR & Combinatorial Science*, *27*(11–12), 1337–1345. doi:10.1002/qsar.200810084

Zefirov, N. S., & Palyulin, V. A. (2001). QSAR for boiling points of "small" sulfides. Are the "high-quality structure-property-activity regressions" the real high quality QSAR models? *Journal of Chemical Information and Computer Sciences*, *41*(4), 1022–1027. doi:10.1021/ci0001637 PMID:11500119

Zhao, L. L., Wang, W. Y., Sedykh, A., & Zhu, H. (2017). Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do. *American Chemical Society (ACS). Omega*, *2*(6), 2805–2812. doi:10.1021/acsomega.7b00274 PMID:28691113

Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., & Tetko, I. V. et al. (2008). Combinational QSAR Modeling of Chemical Toxicants Tested against *Tetrahymena pyriformis. Journal of Chemical Information and Modeling*, *48*(4), 766–784. doi:10.1021/ci700443v PMID:18311912

Paola Gramatica is Senior Professor of University of Insubria (Varese, Italy), where she was Full Professor of Environmental Chemistry in the Department of Theoretical and Applied Sciences (DiSTA) and Delegate of two Rectors for International Relationships, till the retirement (March 2018). At the beginning of her research activity she worked on synthesis and structure of natural organic compounds at University of Milano. Then since 1995 she was Associate Professor of Organic Chemistry and then Full Professor of Environmental Chemistry at University of Insubria, working on QSAR modeling. She founded and was leader of the QSAR Research Unit in Environmental Chemistry and Ecotoxicology of University of Insubria since 1997. She was also Director of the Structural and Functional Biology Department (DBSF). PG is author of 155 scientific papers on ISI international journals, highly cited (near 10.500 citations, h index= 43 in ISI web of Science; 11.400, h=44 in Scopus; more than 12.400 in Research Gate and 14.400 in Google Scholar, h= 52, January 2020), four chapters in scientific books and more than 300 presentations to international meetings (some invited plenary conferences). PG is included in the World Top Scientists list (into the top 15% of 105.000 scientists, selected among near 7 million of researchers in 176 scientific fields) (source PLOS-Biology, Aug.2019). The main research topic is the development and applications of QSAR models to various issues of environmental chemical pollution for prioritization of hazardous organic chemicals (for instance: Persistent, Bioaccumulative, and Toxic (PBT) compounds) and validation of QSAR models. Since 2001, when the EU-White Paper on the new regulation of chemicals was published (now REACH regulation), PG has worked as QSAR expert for the European Joint Research Centre (Ispra) on validation of QSAR models for regulatory applicability and was a selected member of the OECD Task Force of QSAR Experts for the definition of the famous "OECD principles for QSAR validation for regulation applicability". The QSARINS software for MLR models development and validation and QSARINS-Chem software for Insubria databases and application of Insubria models to new chemicals are the main outcomes of the last 20 years of experience in validated QSAR modeling. QSARINS will be the legacy of Paola Gramatica to the QSAR community.