Environment International 131 (2019) 105060

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Review article

Integrating *in silico* models and read-across methods for predicting toxicity of chemicals: A step-wise strategy

Emilio Benfenati[a,*], Qasim Chaudhry[b], Giuseppina Gini[c], Jean Lou Dorne[d]

[a] Department of Environmental and Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via La Masa 19, Milano, Italy
[b] University of Chester, Parkgate Road, Chester CH1 4BJ, United Kingdom
[c] Politecnico di Milano, piazza L. da Vinci 32, Milano, Italy
[d] Scientific Committee and Emerging Risks Unit, European Food Safety Authority, Via Carlo Magno 1A, Parma, Italy

ARTICLE INFO

Handling Editor: Da Chen

ABSTRACT

In silico methods and models are increasingly used for predicting properties of chemicals for hazard identification and hazard characterisation in the absence of experimental toxicity data. Many *in silico* models are available and can be used individually or in an integrated fashion. Whilst such models offer major benefits to toxicologists, risk assessors and the global scientific community, the lack of a consistent framework for the integration of *in silico* results can lead to uncertainty and even contradictions across models and users, even for the same chemicals. In this context, a range of methods for integrating *in silico* results have been proposed on a statistical or case-specific basis.

Read-across constitutes another strategy for deriving reference points or points of departure for hazard characterisation of untested chemicals, from the available experimental data for structurally-similar compounds, mostly using expert judgment. Recently a number of software systems have been developed to support experts in this task providing a formalised and structured procedure. Such a procedure could also facilitate further integration of the results generated from *in silico* models and read-across. This article discusses a framework on weight of evidence published by EFSA to identify the stepwise approach for systematic integration of results or values obtained from these "non-testing methods". Key criteria and best practices for selecting and evaluating individual *in silico* models are also described, together with the means to combining the results, taking into account any limitations, and identifying strategies that are likely to provide consistent results.

1. Introduction

In silico methods based on modelling structure-activity relationship (SAR) offer an alternative approach to generate estimates of chemical toxicity in the absence of experimental data. The availability of high-quality chemical property/effect databases, powerful data mining algorithms, and growing computational power over the past decades has led to more versatile and reliable computational tools and systems for assessing chemical toxicity. Such "non-testing methods" include predictive computational models based on SAR, quantitative SAR (QSAR), read-across extrapolations from measured data on analogous chemicals, and integrated expert systems that derive estimates from a combination of more than one model/approach. Considering the wide diversity of chemical structures, a number of methods and models may be needed to assess different chemicals and toxicological endpoints, and to interpret the results using a "weight of evidence" (WoE) approach.

Models based on (Q)SAR are mathematical descriptions of the biological/toxicological activity of a group of chemical compounds in terms of one or more of their physicochemical properties. A quantitative model may be based on linear or non-linear relationships between the property(ies) and the structural parameters. SARs on the other hand describe qualitative relationship(s) between a chemical structure and its property or biological activity. A simple SAR model may be based on a structural alert (SA), which is a distinctive moiety or a structural feature in the molecule related to its property or biological activity. These models can, however, be only as reliable as the data used to build them, and therefore approaches used in ensuring the quality of chemical and biological data used in model development have been the subject of a number of reviewsPrice and Chaudhry, 2014 (Price, 2014; Benfenati et al., 2007). The assessment of responses at the level of the whole organism also involves understanding of the complex biological processes. This needs a wide range of descriptors of molecular

* Corresponding author.
E-mail address: emilio.benfenati@marionegri.it (E. Benfenati).

https://doi.org/10.1016/j.envint.2019.105060
Received 14 April 2019; Received in revised form 26 June 2019; Accepted 25 July 2019
Available online 01 August 2019

properties to encode such complexities, in conjunction with advanced algorithms based on linear, non-linear or soft-computing techniques (Chaudhry et al., 2007). The selection and use of a particular model(s) and interpretation of the results therefore also require expert knowledge (Benfenati et al., 2013).

A fully tested and validated (Q)SAR model would generally yield a good predictive assessment of toxicity for untested chemicals as long as the compound is within the applicability domain (AD) of the model with regard to chemical space. Due to a different AD, each model will have certain limitations when applied to chemicals belonging to widely diverse classes. Such limitations can however be overcome through the use of several (Q)SAR models, and results can be further improved by combining with read-across extrapolations. Potentially, such an integrated approach would generate robust data, which, when integrated with other lines of evidence using a WoE approach (*e.g. in vitro*, *in vivo* data), could provide the basis for reliable characterisation of toxicological hazard - *e.g.* a reference point/point of departure.

Many (Q)SAR models are currently available for a range of complex toxicity endpoints. Some of these models have been developed for regulatory use - in line with the quality criteria and the validation principles specified in the OECD's Guidance Document on the validation of (Q)SAR Models (Organisation for Economic Cooperation and Development (OECD), Environmental Directorate, 2004). According to the Guidance, a (Q)SAR model should provide: a defined endpoint; an unambiguous algorithm; a defined domain of applicability; appropriate measures of goodness-of-fit, robustness and predictivity; and a mechanistic interpretation if possible.

The European Chemicals Agency (ECHA) has recently published a document describing how to use and report results from QSAR models (European Chemicals Agency, 2016). It includes practical examples on some *in silico* platforms, such as the OECD QSAR toolbox, EPISuite and VEGA.

Since individual QSAR models have become increasingly applied to a broad range of endpoints, and several models have become available for some endpoints, the emphasis has also shifted to the use of several models together to achieve more reliable results. However, this has also raised new questions - such as how to integrate different models, and how to resolve any conflicting outcomes. Finally, how the integration of results from QSAR models and read-across methods can be carried out in a consistent and standardised manner.

A number of integration methods for combining *in silico* results from QSAR models and application of read-across methods exist, and this article proposes a practical guide and stepwise approach that addresses first (Q)SAR models, then read-across, and finally the ways to integrating them through:

- Selection of appropriate (Q)SAR model(s) for a specific property/endpoint,
- Assessment of chemical toxicity using individual (Q)SAR model(s),
- Integration of results from different (Q)SAR models,
- Assessment of chemical toxicity using read-across procedures,
- Integration of results from (Q)SAR and read-across.

These steps that are described below, refer to the general framework provided by the recent EFSA Scientific Committee Guidance on “the use of Weight of Evidence (WoE) approach in scientific assessments” (Hardy et al., 2017). The Guidance constitutes our general theoretical basis; indeed, it is not limited to *in silico* models and proposes three key criteria for assessing, weighing and integrating lines of evidence in a WoE approach: relevance, reliability and consistency. These criteria can be applied to the lines of evidence that are heterogeneous in nature, as explained in the Guidance. In this case, the nature of the results provided using read-across approaches based on extrapolation from experimental data, can be rather different from that obtained from *in silico* modelling. Furthermore, the nature of the results of expert systems can also be different from modelling results that are statistically-based. For instance, models providing structural alerts related to adverse effect can be used for scientific reasoning, and this can inform on the relevance of the predicted value. These aspects will be discussed below, with practical examples.

2. Using individual (Q)SAR models

2.1. Selection of appropriate model(s)

Numerous free-access and commercial models are available – *e.g.* see lists at ANTARES (www.antares-project.eu), QSARDB (https://qsardb.org/), QMRF DB (http://qsardb.jrc.it/qmrf/), and ECHA (European Chemicals Agency, 2016). However, the fact that a model exists for a given property/endpoint does not necessarily mean it is also appropriate or reliable for *in silico* toxicity assessment of a given chemical. A number of criteria need to be taken into account, and these refer to specific features related to the target chemical. For example, most models do not work on inorganic chemicals, disconnected chemical structures, salts, or chemical mixtures. Moreover, a number of models may give different results depending on the format used for a chemical structure (*e.g.* CAS number, SMILES, name, *etc.*), whereas some models (*e.g.* VEGA – www.vegahub.eu) offer a solution by transforming any of the formats into an internal, consistent format. Another consideration for excluding a model may be the cost of the model, and users may prefer freely-available models. Comparative studies have shown that commercial models may not be more reliable than the free-access ones (Milan et al., 2011; Gonella Diaza et al., 2015).

If a commercial QSAR model is available, it should be used. However, it is more critical to provide full documentation, regarding for instance the algorithm and the training set used to build the model. Typically, the algorithm is not fully available, particularly for commercial models. Furthermore, confidentiality of the algorithm and the underlying data may also limit the possibility to develop networks between different systems. If the underlying data used to build the model are not indicated, as may happen for commercial models, this may limit the full and transparent use of the models with regards to their AD.

The inclusion criteria for models need to be based on two major considerations: the quality of each model, and the heterogeneity among different models. Regarding the model quality, it is advisable to start from an assessment based on the comparative performance of different models (for example the EU projects ANTARES - www.antares-life.eu/, and CALEIDOS - www.life-caleidos.eu/).

If available, results on the chemical category of interest should be evaluated and preference should be given to the model(s) that perform best for the category closest to the target chemical.

Where possible and feasible, more than one model should be used to add confidence to the assessment results. Ideally, all those models that fit with the inclusion criteria should be used, but this may be impractical, expensive and time-consuming. For instance, there are currently about 60 models available to predict mutagenicity (Ames test).

2.2. Assessment of individual QSAR results

Each model is based on the structural and activity data of a set of chemicals (the training set). ECHA has recommended taking into account the AD of the models, and whether the training set of the model contains compounds that are similar to the target chemical (European Chemicals Agency, 2017a). Furthermore, details on the output of a model can be variable. If the information provided as output is limited, it is difficult for the user to make a decision on the reliability of the results. Since each model has strengths and limitations, it is preferable to use models that are also more explanatory and provide detailed description of the reliability of the results, which typically is based on the AD.

Three broad categories of models are possible:

Table 1
Reproducibility of the experimental results for some endpoints.

Endpoints	Reproducibility	Reference
Mutagenicity (Ames test)	~80–84%	Piegorsch & Zeiger (1991)
Bioconcentration factor (BCF)	~ ± 0.6 log units	Lombardo et al. (2010)
Acute fish toxicity LC50 range	~ 3 log units	Hrovat & Segner (2009)
Carcinogenicity	~57%	Gottmann et al. (2001)
Developmental toxicity LOAEL	Geometric standard deviation 3.3	Janer et al. (2008)

1) Models for which the AD is not assessed;
2) Models for which the AD is poorly assessed;
3) Models for which the AD is assessed and described using a sound methodology.

The evaluation discussed below refers to the AD of the model as in the REACH regulation, Annex XI (Regulation (EC) no 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency.).

2.2.1. Models for which the AD is not assessed

Many models only provide a predicted value for the property/effect but no means to check whether the target substance is within the models' AD. For these models, a generic way to assess applicability is to refer to the performance on a general set of compounds. Another option is to identify the correct predictions and then to measure the similarity of the target compound compared to the predicted set of chemicals (Kulkarni et al., 2016). However, this process often requires manual handling of the data.

2.2.2. Models for which the AD is poorly assessed

Many models do not include a user-friendly tool for evaluating the AD. The rules for the AD definition are given, but they can only be applied manually. For example, EPISuite indicates that the results of the model are valid if the target chemical has some features within a certain range, and if a number of chemical fragments does not exceed a given values (www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface#what). Thus calculations have to be done manually for each chemical.

2.2.3. Models for which the AD is assessed and described using a sound methodology

Some models not only provide the prediction but also a detailed description of the AD check. If the output of the model provides a high level of detail, the expert can have greater confidence in the results. In any case, the expert has to verify and 'approve' any result produced by the software. Such an assessment has to address the three main components of a (Q)SAR model: the toxicological/biological part, the chemical part, and the algorithm used for modelling. Example of such implementation can be found in the VEGA platform (www.vegahub.eu), which also indicates critical issues that call for expert judgment. VEGA provides a quantitative value, termed as AD index (ADI), considering similar compounds in the model's training set. Automatic calculation of the ADI is based on a series of factors, *i.e.* accuracy of the results on similar compounds, the 'similarity' of the similar compounds, the concordance of the predicted value of the target compounds with the experimental values of the similar compounds, *etc.* Those values provide the user data that are useful for read-across. Unfortunately, in most other programs, the AD is only based on consideration of chemical similarity.

2.3. Uncertainty in the results of a QSAR model

Scientific advisory bodies and regulatory agencies have to deal with different levels of uncertainty in the risk assessment of chemicals. Uncertainties relating to the use of a model to derive toxicity estimates include:

- Parameter uncertainty: the parameters used to construct the model are computed, not experimentally measured, *e.g.* a computed value for LogP.
- Parametric variability: the input variables to the model, *e.g.* the descriptors, may have different values if computed by different programs.
- Structural uncertainty: each model has a bias based on its core datasets.
- Algorithmic uncertainty: a computer implementation may have numerical errors and/or numerical approximations.
- Experimental uncertainty: a measured property generally has a certain degree of observation error. Often this uncertainty is the largest, compared with other sources of uncertainty.

In read-across, the main source of uncertainty is the method itself, which is based largely on the assumptions for similarity scoring of the target molecule to the library molecules. Ideally, to integrate different components into a WoE strategy, the uncertainty associated with each component needs to be characterised. As the quality of a computational model will reflect the quality of the data used to build it, the uncertainty of a (Q)SAR estimate will also reflect the uncertainty of the experimental data. This should be taken into consideration in a QSAR model. Table 1 gives examples of the levels of uncertainty associated with experimental data.

These reproducibility values need to be considered when determining the prediction errors of a QSAR model. It needs noting that:

1) The result of a QSAR cannot be more predictive than the uncertainty of the measured values used to build it. If this is the case, the model is over-fitted. Thus, the uncertainty associated with a particular endpoint should be assessed.
2) The description of the predictivity of a QSAR model is related to the population of chemical substances used to build the model.
3) Different levels of predictivity may apply to different subsets of chemicals. It is useful to know whether results are for a specific chemical class, or a subset of compounds with a certain mode of action (Cappelli et al., 2015).
4) Besides the assessment of the prediction results of the model, it is always preferable to have a measure of the reliability of the prediction for each specific chemical. For example, the ANTARES (www.antares-life.eu/), CALEIDOS (www.life-caleidos.eu/) and PROSIL (www.life-prosil.eu) projects have evaluated several QSAR models for various endpoints, concluding that the use of models for chemicals within the AD improved the results.

3. Integrating results from different (Q)SAR models

3.1. Theoretical basis

The investigation of integrated systems is stimulated by the awareness that combined and integrated approaches are necessary to solve real world problems. Historically, these efforts started with seminal work about bagging classifiers (Breiman, 1996), which opened

the way to the general concept of ensembles (Dietterich, 2000). An ensemble is an algorithm trained with the results of a number of models, which is then used to make predictions. The ensemble itself, therefore, produces a hypothesis that is usually new compared with the hypotheses generated by the models used to build it. This empirical finding has been confirmed by a few theorems (Dietterich, 2000).

Ensembles tend to give better results when there is significant diversity among the models used to build them; in fact models of close similarity tend to have similar prediction errors that cannot be corrected. The prediction accuracy of models to integrate should be better than random choice and individual outputs of the models should be diverse. To the best of our knowledge, there is no explanatory theory about why and how diversity between component models contributes to the ensemble accuracy. One of the most used measures is the Q diversity (Polikar, 2006). Q assumes positive values if the same instances are correctly classified by both classifiers, and negative values otherwise. Maximum diversity is for Q = 0. The Q-value between the classifiers A and B can be computed, taking Nt and Nf to respectively represent the number of data correctly or wrongly predicted by both classifiers, where Na represents number of data correctly predicted by A but wrongly predicted by B, and Nb *vice versa*, as in Eq. (1).

$$Q = (NtNf - NaNb)/(NtNf + NaNb) \quad (1)$$

The Q value ranges between -1, $+1$; the interpretation of Q = 0 is that A and B are independent.

Model diversity can be achieved with various strategies, such as using different datasets or different training parameters to train individual classifiers. Alternatively, entirely different types of classifiers, such as linear regression, decision trees, and support vector machines, to cite a few, can be combined to enhance model diversity.

Two main strategies can be used to combine models: *classifier selection* and *classifier fusion (*Kuncheva, 2005). *Classifier selection* is based on the assumption that each classifier is an expert in a subpart of the domain, so the ensemble has to select the classifier best suited to predict a property or endpoint. *Classifier fusion* is based on the assumption that all the classifiers are trained in the entire space so the ensemble merges all the classifiers to obtain more accurate models. If the probability that each classifier gives the correct answer is better than 0.5, the accuracy of the ensemble approaches the value of 1 as the number of composing classifiers grows to infinity (Boland, 1989). In the case of the *in silico* models, a classifier is an algorithm that assigns the substance to a class, for instance a toxicity class. Two or more classes may be identified by a classifier.

Initial examples of *classifier fusion* were based on bagging and boosting. Using bagging, a number of training data subsets are randomly drawn—with replacement—from the training data. Each subset is used to train a different classifier of the same type; the classifiers are then combined using majority vote. A variant of bagging is random forests, which merges simple decision trees constructed with different parameters. The boosting technique uses a pool of classifiers that are sequentially trained on subsets of data, each time including data misclassified by the previous classifiers. The classifiers are combined using majority vote. Other combination strategies can be applied, such as algebraic combinations or voting. More complex combinations include the bucket and stacking methods. The bucket method takes the outputs of the different models as input to train a new classifier that learns how to combine them. Alternatively, the stacking method generates a combined output from all the outputs of the individual models (Wolpert, 1992). These strategies imply sequential or parallel use of the methods.

The no-free-lunch theorem (stating that any two models are equivalent when their performance is averaged across all possible problems) holds for ensembles as well as for single models (Polikar, 2006). It means that no single ensemble algorithm or combination rule is universally better than the others, and different combinations of the models are expected to provide quite similar results. The approaches discussed below have been shown to be effective to address real world scenarios.

The factors affecting the prediction errors of the integrated model $e(x)$ are related to two components, namely the average error across the different models $\bar{E}(x)$ and the variance $\bar{a}(x)$ of the models relative to the output of the integrated model as described in Eq. (2) (Li et al., 2007).

$$e(x) = \bar{E}(x) - \bar{a}(x) \quad (2)$$

The important variables that reduce the prediction errors of the integrated models are the quality and the diversity of the individual models. Both variables are inversely correlated; integrating individual models with good prediction score and of diverse nature will reduce prediction errors of the integrated models. In addition, the stability of the integrated model increases with the number of individual models that have been integrated.

3.2. Practical basis

The above described strategies are discussed in the context of QSAR. There it is quite common to use the term “consensus” instead of ensemble or integrated model. Indeed, consensus should be used when there is agreement (consensus comes from the Latin *cum* + *sentire*, to feel with, which obviously is not true for models that provide conflicting results).

Of course, integration is easier if there is agreement between the results of the different models. The advantage of using different models lies in the fact that confidence in the results increases when concordant outputs arise from the models. This applies to models that are different since results from closely related models are expected to be very similar, hence they only add redundant information, which does not increase the result reliability.

While most studies have reported better results with integrated approaches, a few studies have also reported similar or better results with a single model (Milan et al., 2011; Gómez-Carracedo et al., 2012; Lei et al., 2009; Katritzky et al., 2006). For such cases, most probably the best model is more predictive compared to other models which may bring “noise”, not novel information (Milan et al., 2011). Considering Eq. (2), if the quality of the additional models is modest, or their diversity is low, there is no advantage over the single model.

3.2.1. Dependent and independent models

The integration of results from independent models makes the overall results more robust for three main reasons:

1. Diverse set of compounds. Here, the larger the set of compounds, the larger is the statistical significance of the model.
2. Variety of chemical descriptors, or fragments used to describe the chemical information.
3. Variety of algorithms to build the model.

Two models are independent when they relate to different strategies, and are based on different algorithms, diverse sets of compounds or chemical descriptors. There is clearly added value when more independent models are integrated. For instance, the International Council for Harmonisation (ICH) M7 guidance on the use of QSAR models for predicting mutagenicity (Ames test) recommends using two models, one with SA, and one with a statistically-based approach (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICM), 2017).

Besides the statistical differences, as in Eq. (1), other factors also need attention. A model based on SAs has the advantage of greater confidence, because it can be represented as one single SA, activating the mechanism causing the toxic effect. However, the list of toxic fragments is by no means complete (Serafimova et al., 2010). Thus, models based on SAs can give false negatives; if no alert is identified, one cannot exclude that there may be a yet-uncharacterised fragment

responsible for the effect. For false positives, some of the alerts may be present for chemicals that are known to be non-toxic (Gini et al., 2014). It is likely that biological/toxicological effects involve multiple factors, and no single alert is sufficient to explain the effect.

The issue of dependent/independent models is even more critical when specific statistical approaches are used, such as Bayesian methods, for which a pre-requisite is that dependence does not exist (Buist et al., 2013).

4. Available methodologies for the integration of prediction results from individual QSAR models

Overall, five categories of methods are used to integrate results from individual QSAR models:

1. Algebraic and voting methods;
2. Weighing methods;
3. Hybrid methods;
4. Learning methods;
5. Expert-based methods.

Fig. 1 illustrates available integration method together with simple schemes for the integration procedures from the predictions of individual QSAR models.

4.1. Simple algebraic and voting methods

Algebraic and voting methods are simple integrations of multiple models, or programs that routinely process the output of several models, like TEST (www.epa.gov/chemical-research/users-guide-test-version-42-toxicity-estimation-software-tool-program-estimate). The user runs several models, and all results are treated as equally good. To assess the reliability of the output, users should also check whether the individual models are independent and, for this purpose, different strategies are applied depending on classifiers or regression-based models.

4.1.1. Classifier models

For *in silico* models defined as ***classifiers***, three combinations are available under the assumption that all model outputs are equally reliable:

(a) The majority vote;
(b) The conservative approach (worst case scenario);
(c) Unanimity: results are used only if they agree.

4.1.1.1. The majority vote. The majority vote takes as correct the most frequent situation, and thus the most probable one, and has been adopted by the US-EPA in TEST. Ruiz et al. applied majority vote to endocrine disruption (Ruiz et al., 2017). For the prediction of estrogen-receptor binding, an integrated model based on four models gave better prediction results compared with an integrated model based on three models, the latter being more predictive compared with the individual models. In contrast, for the prediction of androgen-receptor binding, no improvement of the prediction reliability was observed while comparing integrated and individual models, even considering unanimous predictions. Other studies applied different integration methods, such as simple majority, or plurality, in which case most of the classifiers gave highly reliable predictions (Kulkarni et al. 2016; Morales Helguera et al., 2013). The majority vote method has been applied in case of agreement of positive predictions of any two out of three models (Frid and Matthews, 2010). In this case, the results were

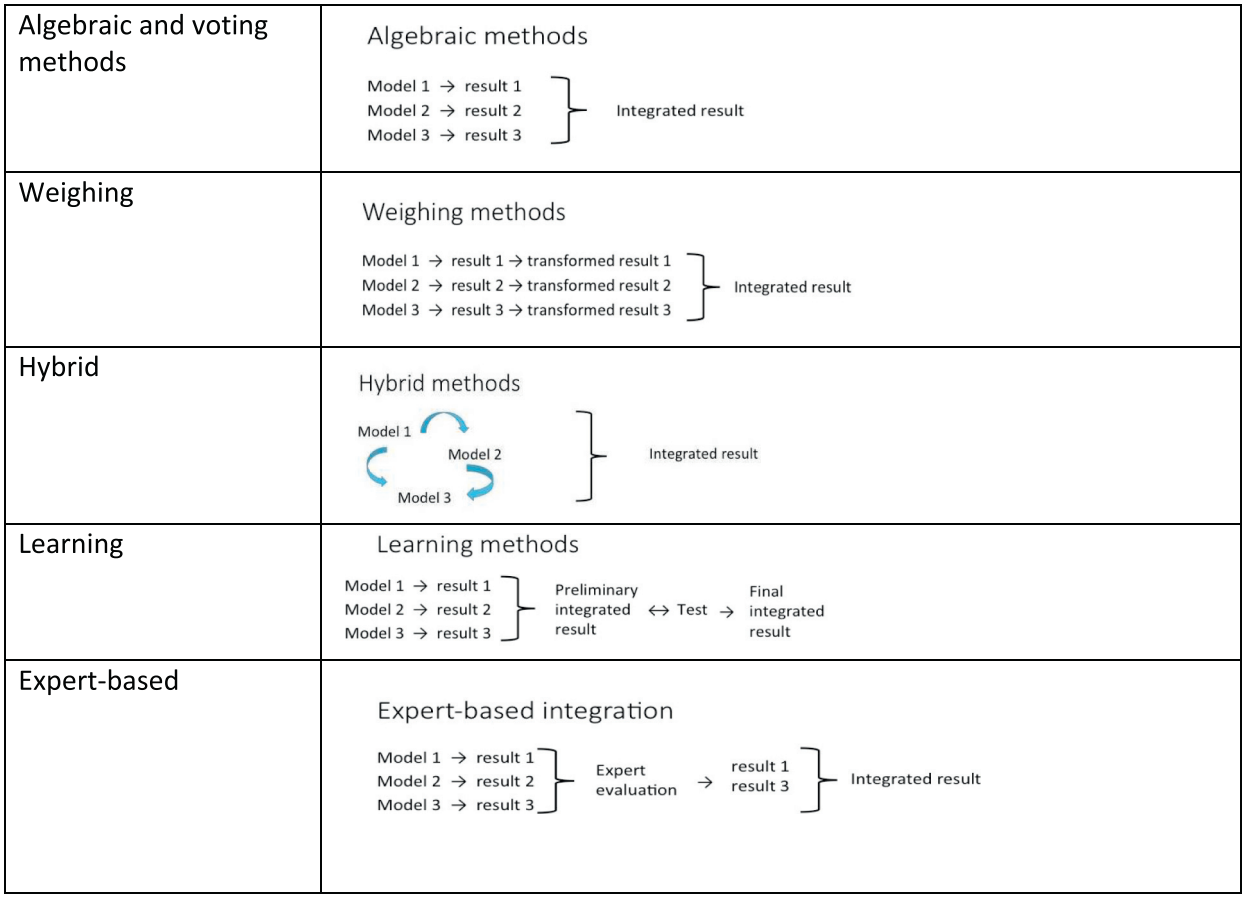

Fig. 1. Methodologies for the integration of prediction results from individual QSAR models. Each method is reviewed in Section 4.1.

more reliable compared with both the conservative and unanimity methods. The majority vote is usually applied with an odd number of models, as no majority exists for conflicting results from two models (Alves et al., 2016).

4.1.1.2. The conservative approach. The conservative approach intends to minimise false negatives. Hence, one disadvantage is that the results are likely to be over-conservative. With classifiers, this strategy provides a large number of false positives, increasing with the number of models used. The approach may be fit for purpose when a small number of models is used to assess large series of chemicals, to prioritise them and to identify chemicals of no-concern.

Jolly et al. applied the conservative approach to assess mutagenicity, using two or three models (Jolly et al., 2015). The authors also considered reliability of the predictions, and added a "no classification" zone when predictions gave ambiguous results. Asturiol et al. adopted the conservative approach to integrate models for skin sensitisation and commented "discordant positive predictions should correspond to low confidence predictions" (Asturiol et al., 2016). Contrera et al. compared prediction results from a number of integration strategies for carcinogenicity assessment, including the conservative approach, and an opposite strategy for minimising false positives (Contrera et al., 2007). The authors concluded that risk assessors should be more prone to accepting a lower risk for carcinogens, hence, avoiding false negatives.

4.1.1.3. Unanimity. The unanimity method has two major advantages: it provides the highest confidence in prediction results, and the most robust statistical results. The main disadvantage is that the coverage of the overall set of models is reduced. Several authors have applied the unanimity method and compared it with other methods (Ruiz et al., 2017; Jolly et al., 2015; Marzo et al., 2016). For instance, Marzo et al. reported that the unanimous approach gave the best accuracy for six models predicting developmental toxicity, but in some instances the approach could only be applied to less than 10% of the chemicals with experimental values (Marzo et al., 2016). Contrera et al. applied the unanimity approach to assess carcinogenicity, reporting more robust sensitivity, specificity and accuracy compared with other approaches, but with less coverage of chemical space (Contrera et al., 2007).

4.1.2. Regression models

The average of the predicted values is the most common approach, typically providing the most robust results. This is particularly relevant if the range of values is limited since the confidence of the results will increase. Conversely, if a large spread of the values is observed it may indicate that at least one of the predicted values is not reliable. This requires a more thorough evaluation of the quality of the models.

The arithmetic mean has been applied to predict water solubility (Muratov et al., 2010), pKa (Balogh et al., 2012), soil-sorption partition coefficient (Gramatica et al., 2007), histone deacetylase inhibitors (Zhao et al., 2013), apoptosis inducing substances (Sciabola et al., 2007), biotransformation in fish (Papa et al., 2014), algal toxicity (Erturk et al., 2012), Daphnia toxicity (Kar and Roy, 2010), and carcinogenicity (Kar and Roy, 2010). Balogh et al. noticed that not only the number of integrated models is important, but also their quality (Balogh et al., 2012). The similarity of the predicted values for soil-sorption partition coefficient depended on the chemicals; the authors therefore split chemicals in homogeneous or high variability classes (Papa et al., 2014). Santos-Filho et al. also generated more reliable predictions with the integration of different models for cell permeability coefficient (Santos-Filho and Hopfinger, 2008). Zhao applied a filter to select the preferred models to be integrated, based on their statistical quality (Zhao et al., 2013). Gaudin et al. generated hundreds of models to predict caffeine encapsulation (Gaudin et al., 2012). The authors explored the performance of the integrated model with increasing numbers of models (up to 100), and observed that 50 models produced optimal results, with minimal improvements by adding additional models. The models were selected based on best predictive performance and diversity. The results from these two strategies were similar; however, the authors favoured integration of the most diverse models to allow for a larger coverage of the chemical space. Li et al. also explored the impact of the number of models on the prediction of infrared spectra and noted an increase in the statistical performance between 10 and 100 models; 50 models gave optimal predictions (Li et al., 2007).

4.2. Weighing methods

Weighing *in silico* models prior to their integration is the second step of a WoE approach according to a recent EFSA Guidance (Hardy et al., 2017). This evaluation is important not only when the models agree but also when the results do not converge. When the results conflict, users often discard all results, which may waste useful datasets and predictions (Benfenati et al., 2013). A number of approaches are available to evaluate the reliability of the models. This evaluation should be done according to explicit criteria, and preferably under a formalised scheme, as discussed below.

Studies on weighing the results of different models and assessing reliability have been proposed to apply weights using considerations regarding (i) the specific case of the target substance, (ii) the results of the model for a certain category of substances (*e.g.* chemical class or mechanism of action), or (iii) the overall assessment of the results of a model in general, typically related to the test set or the total set of compounds available.

4.2.1. Applying substance-specific weights

Some studies have selectively applied weights to specific substances of interest as part of a group of chemicals. For example, the VEGA software (www.vegahub.eu) measures the AD index (ADI) and uses this weight to integrate the results of different models for mutagenicity (Cassano et al., 2014). Kulkarni et al. have integrated the results of several models to predict mutagenicity proving account that the target substance is similar to the chemicals that are correctly predicted by a given model (Kulkarni et al., 2016). Fernández et al. used molecular descriptors relative to the presence of a certain moiety in the molecule (Fernández et al., 2015). The substances in the training set were split into two subsets, depending on whether a certain chemical descriptor was present or not, and the performance of each model was evaluated separately for each of the two. The approach provided a priority list for all QSAR models and descriptors, and allowed the use of the best QSAR model.

Some QSAR studies report prediction performance for specific chemical classes and these can also provide predictions for chemical category(ies) of target compound(s) (Cappelli et al., 2015). More complex case studies have also been explored, and include integration of several model results dealing with several toxicological endpoints (Pizzo et al., 2016).

4.2.2. Applying model-specific weights

Several studies have investigated the integration of different models using a WoE approach with weights assigned to each model (the same weight is applied for all substances predicted by the model). For example, Mansouri et al. integrated with numerous approaches the results of 48 models to predict estrogen receptor activity (Mansouri et al., 2016) and observed that no single method outperformed the others. This conclusion agrees with the theory that prediction accuracy grows with the number of models and is not associated with a specific integration scheme.

Bayesian statistical approaches have been used in several studies. Rorije et al. applied a WoE approach to predict the skin sensitisation potential of chemicals using Bayesian statistics to calculate the probability of the predictivity of the approach and to estimate the reliability of a conclusion (Rorije et al., 2013). The probability whether a

substance was correctly predicted as a skin sensitiser with the first QSAR model was used as prior probability for the second QSAR model, and the process then iterated with each additional model, with the assumption that the models were independent. The study concluded that the application of two or more predictions from QSARs resulted in a higher probability of the conclusion to be reliable. The authors also adjusted the WoE for quality factors using a scale ranging between 0 and 1 – and set a threshold for the probability of the WoE conclusion for 'conclusive evidence'. Based on this study, van der Veen et al. proposed a tiered strategy, using a battery of QSAR models in the initial phase (Van der Veen et al., 2014). Buist et al. used Bayesian statistics to integrate models predicting mutagenicity (bacterial Ames test); the authors commented that models may be dependent, either because they are based on very similar datasets, or they may be related to the same mechanism of action (Buist et al., 2013). Fernández et al. applied Bayesian statistics for the integration of models predicting BCF using both discretized and continuous probability distributions as inputs for the Bayes theory (Fernández et al., 2012). The authors also considered a cost matrix to take into account the cost of misclassification; to provide a conservative assessment, the cost for false negatives was higher.

Information on the AD was used to apply weights to preferred predictions for human monoamine oxidase inhibitors, obtaining better performance of the integrated model (Morales Helguera et al., 2013). Alves et al. also noted that the use of AD improved the statistical performance of integrated models for skin sensitization (Alves et al., 2016).

Gupta et al. used a scoring function to integrate two models for the prediction of kinase inhibition activity, using an optimal coefficient for each model (Gupta et al., 2010).

In the case of models developed with the random forest approach, the integration of multiple models is a rule which has promoted the use of ensemble methods. For instance, Zhang et al. studied a range of ensemble methods, and identified a preferred method for imbalanced datasets Zhang et al., 2009 (Zhang & Hughes-Oliver, 2009).

4.3. Hybrid models

A number of *in silico* models have also been combined in the form of hybrid intelligent systems (Amaury et al., 2007). Compared with the approaches previously described, where pre-existing models are integrated *a posteriori*, the integration of different models is planned upfront. In this context, neural networks with fuzzy systems (Neagu, 2003), and neural networks with symbolic rules, have been the most used (Gini et al., 1998).

Hybrid intelligent architectures can be classified in two main categories: stand-alone, and integrated. Stand-alone systems combine different techniques in a single computational model, sharing data structures and knowledge representations. This is the case for many neuro-fuzzy systems. Integrated hybrid systems combine various techniques and focus on their interaction; for instance, genetic algorithms are used in the selection of the most relevant variables and are then further combined with learning methods to build a classifier (Amaury et al., 2007).

Well-known techniques may have limitations; however, these can be overcome through integration with complementary methods. Table 2 compares the pros and cons of the most common systems - *i.e.* expert systems (ES), neural networks (NN), fuzzy systems (FS), and genetic algorithms (GA) - and summarises their technical features.

Briefly, the integration of results generated from *in silico* models can be visualised at two levels. The first level includes individual models built in the classical manner or *in silico* models. Then, the second level takes their results as inputs into a software platform (Fig. 2).

The use of hybrid systems is particularly appealing since it can employ models that are already available and integrate them to create a broader system.

For example, Gini et al. integrated a neural network predicting the dose response relationship for the carcinogenicity of aromatic compounds with a rule-based system that assessed chemical fragments to predict the carcinogenicity classification (Gini et al., 1999). More detailed discussion on the use of hybrid methods in the QSAR field is given in Amaury et al. (Amaury et al., 2007)

Table 2
The main features of ES, NN, FS, and GA.

Property	ES	NN	FS	GA
Knowledge representation	Good	Very poor	Good	Poor
Knowledge discovery	Very poor	Good	Poor	Fair
Explanation ability	Good	Very poor	Good	Poor
Tolerance to imprecision	Bad	Good	Good	Good
Tolerance to uncertainty	Fair	Good	Good	Good
Adaptability	Very poor	Good	Poor	Good
Learning ability	Very poor	Good	Very poor	Good
Maintainability	Very poor	Good	Fair	Fair

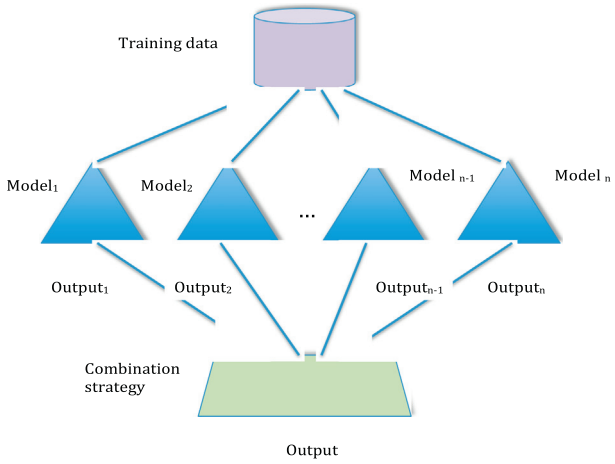

Fig. 2. Integration of individual models into a hybrid model.

Another case of integrated models includes prediction of ecotoxicological properties of pesticides (Benfenati, 2007). Five environmental endpoints were modelled using hybrid systems and outputs of each individual model were combined as inputs in an integrated software, providing more robust predictions.

Another example is the CAESAR mutagenicity (Ames test) model within VEGA. It is a hybrid model, composed of a statistical model predicting whether the substance is mutagenic or not; in order to increase the sensitivity, a second model is applied to substances that have been predicted to be negative using a set of rules (Ferrari and Gini, 2010). If the output of the second model also predicts mutagenicity, the software stops. In contrast, if the output of the second model predicts non-mutagenicity, a third model is applied based on another set of rules to obtain additional assessment regarding mutagenicity or non-mutagenicity.

4.4. Learning methods

Learning methods constitute a further step towards more complex integration of the results from different models. With the approaches described previously, only *a priori* information and prediction results could be integrated. In contrast, learning methods optimise the integration of different models based on *a posteriori* results when predicting new chemicals, so that data on chemicals not used in any of the separate models are needed. In most cases, this requirement is data demanding, because typically all chemicals with available activity data are already used in the development of individual models (*i.e.* the training and test sets).

In multiclass classification (Benfenati et al., 2002), the learned combination from five models with a neurofuzzy method gave better

prediction accuracy for predicting fish toxicity of organophosphorus pesticides. In another example, the authors used a neural network programmed to learn and combine models for the prediction of pesticide oral LD50 in quail for great improvement in accuracy (Gini et al., 2009).

4.5. Expert-based integration of in silico results from different models

In certain circumstances, expert judgment provides decision points on how to manage the results of the different models. Ellison et al. proposed a WoE approach for *in silico* methods for assessing skin sensitisation potential based on semi-quantitative scoring of the models using expert judgment (Ellison et al., 2010). Price and Chaudhry assessed the contribution of different individual input data to build QSAR models by a WoE approach for the evaluation of toxicity from chemicals migrating from food packaging, using a number of steps as follows: (Price and Chaudhry, 2014)

1) Data with low reliability from a (Q)SAR were excluded.
2) Data from "out of domain" compounds were excluded.
3) 'Positive' predictions were assigned if more than one method returned a positive prediction and fewer than two gave a negative or equivocal prediction.
4) 'Negative' predictions were assigned if more than one method returned a negative prediction and fewer than two gave a positive or equivocal prediction.
5) 'Equivocal' predictions were assigned if one of the above conditions was not fulfilled.
6) The final prediction of toxicity for each endpoint was based on a combination of results from read-across, QSAR models within VEGA, and the OECD QSAR Toolbox. The assessment showed that 65 of the 136 migrant packaging compounds were outside the domain of any of the relevant QSARs in the Toolbox. For these compounds, SAs were used for carcinogenicity or mutagenicity, together with read-across and VEGA results. These SARs are part of the profiling module of the Toolbox and comprise carcinogenicity and mutagenicity alerts from the expert system Toxtree (http://toxtree.sourceforge.net/). An unequivocal prediction was recorded when at least two results were in agreement.
7) For read-across purposes, QSARs were available for analogs of the migrant compounds in the first place since this was one of the criteria for their selection. To compare the individual models used on both sets of compounds to derive the WoE predictions, standard statistical parameters were used to assess the quality of the predictions.

Integration of models for the prediction of BCF has been applied by other studies for which a workflow was developed to identify the conditions producing more reliable results (Gissi et al., 2013). The components were the inputs of the different models, their relative diversity as well as the ADI (note that the ADI is different for each substance). Later on Grisoni et al. applied the integrated strategy and compared the results with other models for BCF prediction (Grisoni et al., 2015). The authors found that the integrated strategy did not provide better predictions but a broader AD.

4.5.1. Sequential integration of in silico models

In silico models may be processed in parallel or in a sequential manner. The latter requires the identification of exclusion criteria which may be applied as a filter, or to the identification of the preferred model.

Some results may be excluded for their low prediction reliability as described in Price and Chaudhry (Price and Chaudhry, 2014). Another strategy is to identify the preferred model, providing good prediction, and adding other models if necessary. For instance, if a model specific for a certain chemical class is available, it is generally preferred over a model developed for a broader chemical space.

When a large number of chemical classes is present in the dataset, specific QSARs can be developed for each class, and an "expert selection" approach can be applied to find the best QSAR model(s) for a specific substance. This approach gave good results for the prediction of the lethal concentration for 50% the animals (LC_{50}) in fathead minnow, with prediction errors 30% lower than that from a QSAR built on the entire dataset (Koenig et al., 2004). Another example is the above described CAESAR model for mutagenicity, which integrates statistics and expert-based rules (Ferrari and Gini, 2010).

5. Integration of prediction results from read-across methodologies: tools and results

In most cases reported so far, methods for the integration of prediction results from read-across methodologies have been based on expert judgment. This may give subjective results, because such an evaluation depends on personal expertise, and the parameters used to evaluate different lines of evidence may not be harmonised.

Some QSAR models provide the experimental values of the chemicals similar to the target compounds, as VEGA and TEST. This information can be used to facilitate read-across. The process is reproducible if it is run with automatic protocols so that these programs can be used for read-across, disregarding the predicted values. Furthermore, there are statistical QSAR models based on k-nearest neighbour (kNN), searching for the k most similar compounds, which also constitute a kind of automatic read-across (Manganaro et al., 2016).

A number of software have been specifically designed for read-across including:

- The OECD toolbox (www.qsartoolbox.org) allows the user to perform read-across, and provides several profilers to explore the possible mechanisms associated with a toxic effect or hazard property.
- ToxRead (www.toxread.eu) graphically shows chemicals similar to the query molecule, and provides a rationale for the effect in relation to the target compound. ToxRead represents all the elements for the evaluation on the same scheme - the similar substances, the rules that apply to the target compounds, and the similar compounds that have rules in common with the target compound. Based on the presence of active and inactive compounds, and of similar compounds, ToxRead calculates the effect value for the target compound (Gini et al., 2014).
- AMBIT offers another tool for read-across and includes the substances registered in REACH (http://cefic-lri.org/lri_toolbox/ambit/). However, the tool highlights legal disclaimers preventing the use of the available registered data for commercial purposes.
- Toxmatch is an open-source tool for the assessment of similarity for multiple properties which facilitates read-across and grouping (https://eurl-ecvam.jrc.ec.europa.eu/laboratoriesresearch/predictive_toxicology/qsar_tools/toxmatch). For a recent review see Patlewicz et al. (Patlewicz et al., 2017)

Recently, several studies recognised that the development of read-across strategies will benefit from the integration of multiple properties and criteria, which are not limited to the similarity based on the chemical structure. These include physicochemical, toxicokinetic and toxicodynamic properties which can also be used to evaluate similarity between chemicals (Schultz et al., 2015; Schultz and Cronin, 2017; Schultz & Richarz, 2019; Kuseva et al., 2019; Cronin et al., 2019). Relevant physicochemical properties may include logKow (the partition coefficient between octanol and water), water solubility, melting point, boiling point, hydrolysis rate, and Henry constant. These properties are often linked to bioavailability, and are closely related to a range of toxicological endpoints, both of ecotoxicological and human relevance. Toxicokinetics and/or (a)biotic transformation may strongly affect the

behaviour of the chemical substances, including metabolic pathways. The biological and toxicological similarities are also strongly linked to the read-across hypothesis. Recently, mechanistic data and information, assessed through mode of action or adverse outcome pathways (AOP) has raised interest for use in read-across (Schultz et al., 2015). Thus, these properties can also be applied to identify similar compounds.

Novel testing methods have also been indicated as useful support for read-across (Berggren et al., 2015). Read-across has also been described within guidelines published by regulatory authorities, such as ECHA, and these clearly shows that there is both a theoretical and a practical interest in such non-testing method (European Chemicals Agency, 2017b).

For read-across, typically only one assessment is performed following a single conceptual pathway. Indeed, read-across is context-dependent (Schultz and Cronin, 2017). Some studies have addressed the validation of read-across. For instance, one study has evaluated the results of read-across for 27 substances assessed by several experts for three endpoints: mutagenicity, BCF and fish acute toxicity and results using ToxRead provided good reproducibility (Benfenati et al., 2016).

6. Integrating evidence

6.1. Advantages and disadvantages of ensemble methods

Ensemble models have a number of advantages over a single model. First of all, the use of a single model implies rejecting *a priori* part of the evidence that may be available from other models. The second advantage is of statistical nature. It is hard to find the best model but it is possible to find several good models. Integrating their outputs may reduce the effect of random noise in data and avoid over- or under-estimation (Gaudin et al., 2012). A third advantage is data availability, since often there are too few data to build individual models with robust predictions. In this case, a resampling technique can be applied to generate different overlapping subsets, which can be applied to train the single model, and then create an ensemble of models.

Another advantage of ensemble models regards diversity of the available endpoints (*e.g. in vitro*, *in vivo*, OMICs, *etc.*), and the experimental test species (*e.g.* rat, mouse, dog, rabbit, *etc.*). Such diverse data do not allow building a single classifier or continuous model, but rather different classifiers or continuous models from each data/endpoint type of the same nature. These models can be integrated using ensemble models.

The disadvantages of ensemble methods are related to complexity of the integration of the numerous models to be executed, and diversity of the integration methods.

Besides these disadvantages, there are at least two other main issues: (Price and Chaudhry, 2014) the time required to run the models and to integrate them, and (Benfenati et al., 2007) the cost of the models, which may be high for commercial software. One partial solution is to use free tools and platforms which automatically integrate results of different models as in the already cited TEST and VEGA systems. Roy et al. has made available a free program for integrating multiple models in three possible ways, from all the predictions, or the weighted predictions, or selecting the best (Roy et al., 2018).

6.2. Integrating the results from individual (Q)SAR and read-across models

Fig. 3 illustrates major factors impacting the integration of the elements of non-testing methods such as QSAR and read-across. The expert uses her/his knowledge to assess the substance(s) for (eco)toxicological properties. The predicted values from *in silico* and read-across methods to be integrated should be relevant, reliable and consistent and those deserve attention to support the expert in the assessment.

Table 3 indicates the prevailing elements within the (Q)SAR and read-across approaches. (Q)SARs are general models for which detailed knowledge of the property may not be as explicit and detailed as a specific read-across. Indeed, read-across are often focused on specific chemicals, and the information applied for read-across relates to the chemicals (target and source chemical(s)), and can include other types of evidence (such as target organ, plausible mode or mechanism of action, presence of one or more SAs, *etc.*). All these details are not always considered in (Q)SAR models.

On the other hand, (Q)SAR models may use well-defined parameters (*e.g.* polarity, solubility, presence of a certain active fragment in the molecule) and thus may indicate a solid, general and transparent scientific basis associated with the prediction.

In general, read-across approaches apply to a narrow chemical space, and are appropriate for compounds of closely similar structure. It is not recommended to derive conclusions from substances that do not share similar properties. In addition, it is important to note that chemical similarity is one criterion to address similarity and there are also others such as physicochemical properties, exposure, source, biological and toxicological effect, use and regulatory criteria.

Referring to the reproducibility of the experimental value applied to read-across, results from a single compound are considered less reliable compared with those based on a range of compounds. Ideally, the quality of the data predicting the property values of the substances used for read-across should be higher than in the case of QSAR.

Various QSAR programs and software platforms include an assessment of chemicals similarity to the target compound, and the results can be directly used for read-across assessment. For example, ToxRead integrates the results of QSAR models and read-across evaluation for mutagenicity as a single output.

In any case, the use of optimised software tools for the integration of *in silico* tools with automated functions should not undermine the role of expert judgment for assessing the results for their consistency, relevance, limitations and the corresponding uncertainties particularly for conflicting outcomes. Overall, the evaluation resulting from the application of QSAR and read-across tools should be integrated based on the evidence available. This requires the evaluation of each individual model prior to integration, the assessment of the results of the QSAR models used, and finally the assessment of the combined QSAR and read-across results.

6.2.1. Integrating concordant results

If the results of different models/tools agree, the evaluation is simple, and the only issue is the level of reliability. If one of the values is sufficiently reliable, the others will contribute towards raising the overall level of confidence.

However, things may differ if results agree only for a series of models, but none of them is individually supported by the evidence that the prediction is valid. It may happen for results of models lacking the AD assessment. Anyhow, as all the results support each other, the confidence can increase by increasing the number of models.

Table 4 summarises elements of uncertainty analysis for the interpretation of the results from *in silico* methods in relation to increasing evidence (columns) and agreement between models (rows) (adapted from the IPCC guidance) (Mastrandrea et al., 2010). In principle, confidence in the model results is highest as agreement and evidence are highest and the reverse provides the lowest confidence. In the case of a very large number of QSAR models, the user may still obtain sufficient confidence in the results, even if the models are not fully transparent.

Factors increasing the confidence of the assessment include:

- Number of models in agreement;
- Information on the applicability domain of the models;
- Identification of the scientific criterion for the effect/property – *e.g.* SA/molecular feature associated with toxic properties;
- Number of similar compounds in agreement with experimental values;
- Agreement between the results from *in silico* models and read-across.

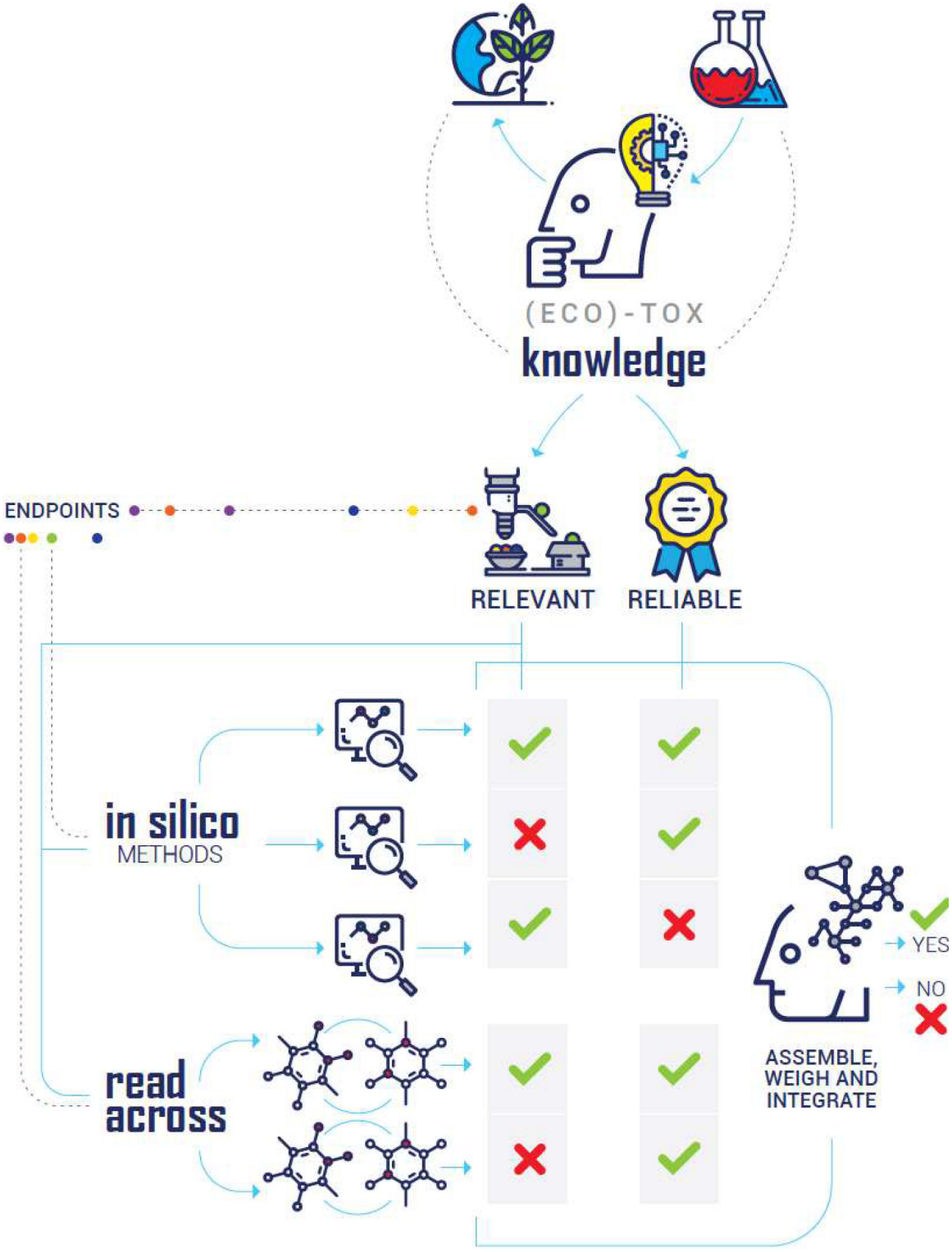

Fig. 3. Scheme for the integration of the results from *in silico* models and read-across methods. Experts evaluates the lines of evidence for each endpoint for their relevance, reliability and consistency. *In silico* methods may help experts in this process particularly to support the identification of the most robust datasets for the context of the assessment. Similarly, tools for read-across may identify datasets which may not be adequate for the purpose of the assessment. During the process, experts can integrate the results from multiple sources, applying weights related to the relevance, reliability and consistency of the results.

Conversely, factors increasing the uncertainty of the assessment include:

- Limited number of models;
- Lack of information on the applicability domain of the model;
- Limited number of similar compounds;
- Conflicting results between different models;
- Conflicting results between similar compounds;
- Lack of knowledge regarding the scientific criterion for the effect/property.

Table 3
Knowledge perspectives and range of approaches available in read-across and QSAR models.

	Local approach	General approach
Implicit knowledge	Similar compounds: Only one or few substances are used.	(Q)SAR models. At the basis of the model there is a large number of substances with the relative property values.
Explicit knowledge	SAs, mode-of-action, mechanism of action, *etc.* relative to a specific property. The user should check if the same theoretical mechanism is applicable to the target and the similar compounds.	Pharmacokinetic properties, polarity, *etc.* Ideally, the descriptors used within the model are associated with the mechanism involved in the process to be modelled.

6.2.2. Integrating conflicting results

In the case of conflicting results, the two separate lines of evidence obtained from results of QSAR model predictions and read-across approaches, should be carefully assessed. This often requires careful analysis of the scientific criteria to classify the effect or property dealt with by the QSARs or the read-across and it is useful to identify similar compounds as a starting point. This may be the preferred approach, since evidence associated with experimental values, *i.e.* read across, are considered more robust than that from the results of QSAR models. If the read-across also indicates an adverse (potent) effect or property, it may be difficult to find a solid scientific ground which supports lack of effects, even if the QSAR model predicts it. In this case, a strategy would be to document the scientific ground demonstrating that such adverse effect or property is not likely for the target compound, due for example to pharmacokinetic properties which would not apply to similar compounds.

For situations under which no closely similar compounds are available, the read-across evaluation may become weak, and results from QSAR models may be more informative.

For models predicting effects for which a sound scientific ground has been defined as an explicit rule, *e.g.* SA, this may provide relevant, reliable and consistent results. If the same SA is present in the target and source compound, this may overrule the result of the *in silico* model. Fig. 4 illustrates the approach using ToxRead to predict the output of the mutagenicity assessment for diethylcarbamic chloride.

In this graph, chemicals are shown as circles, and rules as triangles. The target compound (diethylcarbamic chloride), at the center, is connected to three triangles, representing read-across by three rules all indicating mutagenicity because the triangles point downwards and are red. The rule with the lowest *p*-value for toxicity is at the top, and the values increase clockwise. Thus, we start with rule R3.0, referring to the fragment N-C-Cl. Linked to the alerts are chemicals found in the database that are similar to the target and fall under the specific rule. These similar compounds are represented as circles and denoted by CAS numbers; all are shown in red because they have already been tested positive for mutagenicity. The size of the similar compounds depicts the level of similarity (larger ones are closer). From this picture, it can be seen that the most similar compound is CAS 79-44-7, which is dimethylcarbamic chloride (see Fig. 5). This substance is very similar to the target compound, the only difference being the carbon chain linked to the nitrogen, with one or two carbon atoms. Thus, this provides a strong line of evidence towards mutagenicity. Positivity is further confirmed by three other similar compounds that are also mutagenic. The combination of these two lines of evidence is important for the overall conclusion on mutagenicity of the query compound.

The other two SAs are rules R1.0 and R2.0, which are closely related - both being acyl halides - R1.0 includes different halogens, while R2.0 refers only to the acyl chloride. Linked to R.0, there are two mutagenic compounds, and one non-mutagenic. Chemical 79-44-7 is again associated with this rule. However, there is also a non-mutagenic compound (CAS 2941-64-2), which is S-ethyl chloridothiocarbonate; it contains no nitrogen and its sulfur is linked to the C═O moiety. This is a major difference compared to the target compound and, therefore, does not necessarily conflict with the fact that other substances with the acyl-chloride linked to nitrogen (and hence more similar to the query compound) are mutagenic. Very similar considerations apply to the rule R1.0, because these rules are closely related, and the similar compounds are the same. Thus, the read-across results provide a strong evidence for mutagenicity from rule R3.0 and the related compounds, and partial support from the evidence from rules R2.0 and 1.0.

ToxRead also shows similar compounds (Fig. 4), regardless of the presence of rules. The most similar compounds, shown in Fig. 3 as circles directly linked to the target compound, are dimethylcarbamic chloride (CAS 79-44-7, already discussed), *N*,*N*-diethylacetamide (CAS 127-19-5), and *N*-nitrosodiethylamine (CAS 55-18-5). Two of them are mutagenic whereas *N*,*N*-diethylacetamide is non-mutagenic but quite similar to the target compound. Its similarity index to the target compound is 0.84. The information provided by this substance does not contradict the line of evidence of mutagenic effect relating to rule R3.0, because this rule is absent in CAS 127-19-5, and therefore despite being structurally similar it is not relevant for the assessment.

Chemical CAS 55-18-5 poses a similar situation as it does not contain the same alerts for mutagenicity present in the target compound, even though it is quite similar (similarity index 0.817). Thus, this substance is not important for the assessment due to the absence of the relevant rule in the target compound. Table 5 summarises the elements used for the assessment. This shows that a simple evaluation of the most similar compounds, as in Fig. 4, without further expert scrutiny, may be misleading: one substance is mutagenic and relevant, whilst the other two are not relevant because neither has the relevant information related to mutagenicity. Thus, the user needs to look beyond simple chemical similarity as it may provide false indications.

In ToxRead, the results of VEGA QSAR models appear by clicking on the target compound. All four QSAR models predict the substance as

Table 4
The integration of different scenarios with varying levels of agreement and number of tools.

	Limited evidence	Medium evidence	High evidence
High agreement	**Many** QSAR models **agree**, but there is *limited evidence from read across*/similar compounds/alerts.	**Many** QSAR models **agree** and there is a *medium level of evidence from read-across*/similar compounds/alerts.	**Many** QSAR models **agree** and there is a *high level of evidence from read-across*/similar compounds/alerts.
Medium agreement	**Medium agreement** between **some** QSAR models and *limited evidence from read-across*/similar compounds/alerts.	**Medium agreement** between **some** QSAR models and a *medium level of evidence from read-across*/similar compounds/alerts.	**Medium agreement** between **some** QSAR models and a *high level of evidence from read-across*/similar compounds/alerts.
Poor agreement	**Low number** of QSAR models in general and/or **poor agreement** among the models. Also *limited evidence from read-across*/similar compounds/alerts.	**Low number** of QSAR models in general and/or **poor agreement** among the models. Also a *medium level of evidence from read-across*/similar compounds/alerts.	**Low number** of QSAR models in general and **poor agreement** among the models. Also a *high level of evidence from read-across*/similar compounds/alerts.

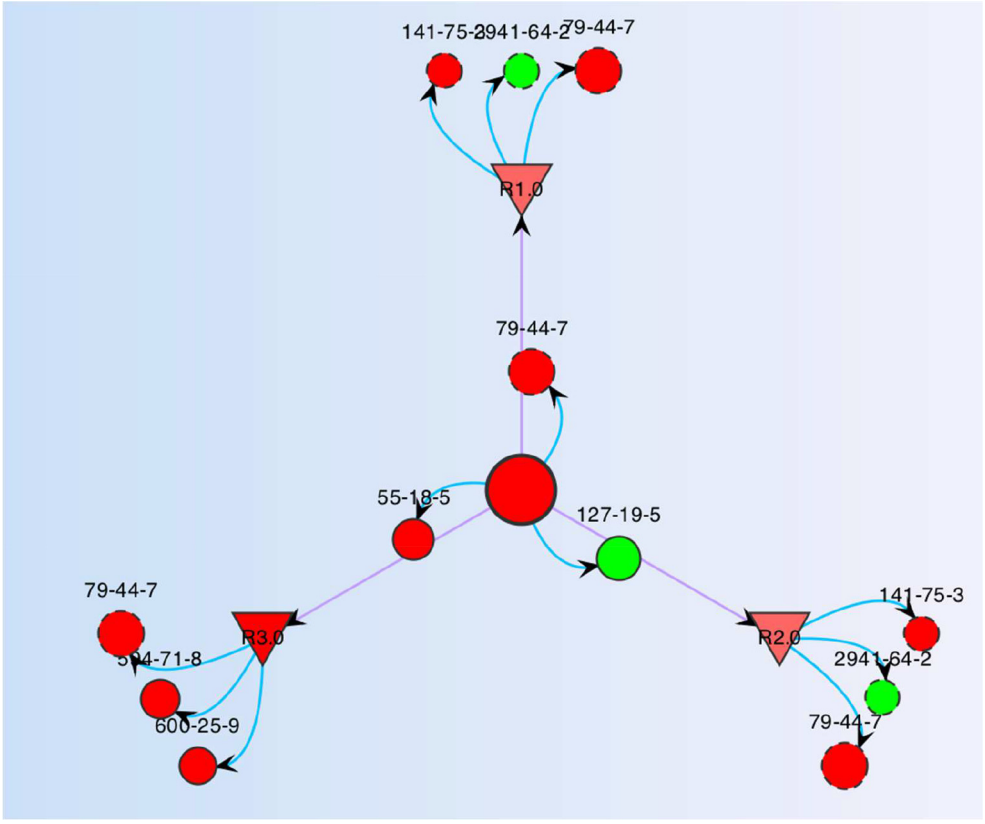

Fig. 4. Graph representing mutagenicity assessment for diethylcarbamic chloride, similar compounds and toxicity alerts using ToxRead. The numbers in the CAS number format (XXX-YY-Z) indicate the substances, which in most of the cases appear more than once. The numbers in the triangles indicate the rules, such as structural alerts. Important substances are discussed in Table 5.

mutagenic, with different levels of ADI; ToxRead combines the results of read-across and QSAR, and provides the overall assessment for the compound as mutagenic.

This case study demonstrates the importance of assessing different lines of evidence generated from *in silico* models and read-across approaches and specifically here three lines of evidence: (Price and Chaudhry, 2014) similar compounds; (Benfenati et al., 2007) presence and relevance of alerts; and (Chaudhry et al., 2007) results of an independent method (*e.g.* QSAR model). The availability of all these elements simplifies the interpretation of the results to conclude, whereas in contrast, 'weaker' evidence may be disregarded. For example, a combination of a structural alert with data from similar compounds greatly increases the confidence in the prediction. In such a situation, it may be possible to over-rule the conflicting result from a QSAR model.

Table 5
Elements in common with the target compound.

	Similar 1: 79–44-7	Similar 2: 127–19-5	Similar 3: 55-18-5
Mutagenic	Yes	No	Yes
Rule 3.0	Yes	No	No
Rule 2.0	Yes	No	No
Rule 1.0	Yes	No	No
Similarity with the target compound	0.861	0.840	0.817
Relevance	Yes	No	No

7. Conclusions

The assessment of complex biological properties, such as chemical toxicity, without involving the use of animals is not straightforward. In the absence of a living and functioning biological organism, alternative non-testing methods such as *in silico* models can at best provide pieces of the evidence. However, if several strands of the partial evidence are combined in a systematic way, they may provide sufficient overall evidence to support conclusions with a robust scientific basis. Among other alternative methods, *in silico* approaches stand out because they provide low-cost and rapid means for qualitative and quantitative assessment of chemical toxicity. *In silico* tools have also diversified hugely over the years and have become more stringent to provide relevant, reliable and consistent results. However, the outcomes of various *in silico* approaches still face the challenge of their wider acceptance, to

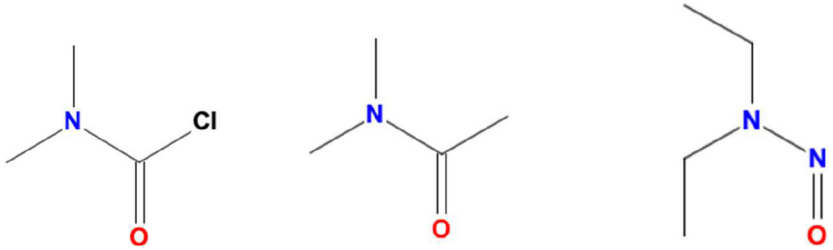

Fig. 5. From left to right - chemical structures of dimethylcarbamic chloride (CAS 79-44-7), N,N-diethylacetamide (CAS 127-19-5), and *N*-nitrosodiethylamine (CAS 55-18-5).

assemble sufficient evidence for hazard assessment and ultimately for their regulatory use in chemical risk assessment. The recent EFSA guidance document on the use of the WoE approach in scientific assessments provides a robust framework and methodology to assemble, weigh and integrate lines of evidence under different data-poor and data-rich scenarios (Hardy et al., 2017).

This review has focused on the methods and stepwise approaches to integrate results from available *in silico* models and read-across approaches which can be used to predict hazard or property values. Concerning such integration approaches, data gaps have been identified from the literature particularly in relation to recent QSAR and read-across models and software. It is widely recognised that integration of the results from different (Q)SARs and other *in silico* tools can improve the overall confidence in the predicted estimates. In doing so, it is critical to consider key criteria about evidence, agreement and uncertainty that will affect relevance, reliability and consistency of the results. As discussed in this article, several strategies for the integration of results from are available, and for QSAR models the predicted value, the model's AD and the information on structurally-similar compounds are of critical importance. It is also clear that integration of the results from different tools must not be taken as a mere aggregation of the numbers. A number of other aspects also need consideration, such as the underlying principles and rules, SAs, chemical similarity, similar physicochemical properties and relevant molecular descriptors. Each of these elements should be included in the overall integration of the results from different models/models. It is also important that the overall process of integration is transparent, and allows exploration and critical assessment of the various steps. Indeed, interpretation of the results needs to be performed by the user in the context of the assessment, as integration tools can only provide a means for assembling the relevant elements. In this context, the integration methods and steps presented here also aim to break down the current barriers between QSARs and read-across approaches and facilitate their acceptance and application in regulatory risk assessment.

The integration of different models potentially offers reliability and confidence in the assessments because of the consideration of multiple lines of evidence. However, this also requires the user to be familiar with QSAR modelling approaches and read-across, and to take into consideration the detailed information provided by each tool to interpret the output and draw conclusions with confidence.

The use of data regarding physicochemical properties will further improve the integration of QSAR models and read-across. These data may be of experimental or predicted nature, in the case of missing values. These data are not only relevant to analysing chemical structures, but also to exploring and understanding the biological and toxicological processes that are fundamental to the interaction of a substance with a biological target site to elicit an adverse effect.

The integration of the results from multiple modelling tools is expected to increase with the increased use of large collection of data collection, such as ToxCast (https://www.epa.gov/chemical-research/toxicity-forecasting) and the US-EPA computational dashboard. The availability of toxicological data in structured databases will offer greater possibilities to apply read-across approaches, and a basis to develop new *in silico* models. A recent example includes EFSA's chemical hazards database: OpenFoodTox (https://www.efsa.europa.eu/en/microstrategy/openfoodtox). OpenFoodTox contains EFSA's hazards data used in over 1650 risk assessments for over 4500 chemicals. The availability of *in silico* tools for toxicokinetic parameters, including those describing absorption, distribution, metabolism and excretion (ADME), represents another very important factor to be added to the weight of evidence. For instance, the use of toxicokinetic data may help to select data on chemicals and species relevant for the target compound, since the metabolism may vary between different species.

Thus, in the future, the general framework might remain valid, however, data complexity and steps may probably be added while increasing the level of relevance and reliability and consistency required to decipher the complexity of the biological and toxicological processes.

Declaration of Competing Interest

There are no conflicts to declare.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 681002 (EU-ToxRisk) and from EFSA, contract NP/EFSA/AFSCO/2016/01. The views expressed in this article are the authors only and do not reflect the views of the European Food Safety Authority.

References

Zhang, Q., et al., 2009. A model-based ensembling approach for developing QSARs. Journal of Chemical Information and Modeling, 49, 1857–1865. https://doi.org/10.1021/ci900080f.

V. Alves, E. Muratov, S. Capuzzi, R. Politi, Y. Low, R. Braga, A. V. Zakharov, A. Sedykh, E. Mokshyna, S. Farag, C. Andrade, V. Kúzmin, D. Fourches and A. Tropsha, Allarms about structural alerts, Green Chem., 2016, 18, 4348–4360. DOI:https://doi.org/10.1039/C6GC01492E.

N. Amaury, E. Benfenati, S. Bumbaru, A. Chana, M. Craciun, J. R. Crètien, G. Gini, G. Guo, F. Lemke, V. Minzu, J. Muller, D. Neagu, M. Pintore, S. A. Stroia and P. Trundle, Hybrid systems, in *Quantitative Structure-Activity Relationships (QSAR) for Pesticides Regulatory Purposes*, ed. E. Benfenati, Elsevier Science Ltd, Amsterdam, 2007, pp. 149–184.

D. Asturiol, S. Casati and A. Worth, Consensus of classification trees for skin sensitisation hazard prediction, Toxicol. in Vitro, 2016, 36, 197–209. DOI:https://doi.org/10.1016/j.tiv.2016.07.014

G. T. Balogh, A. Tarcsay and G. M. Keseru, Comparative evaluation of pK(a) prediction tools on a drug discovery dataset, J. Pharm. Biomed. Anal., 2012, 67-68, 63–70. DOI: https://doi.org/10.1016/j.jpba.2012.04.021

Benfenati, E. (Ed.), 2007. Quantitative Structure-Activity Relationship (QSAR) for Pesticide Regulatory Purposes. Elsevier Science Ltd, Amsterdam.

. E. Benfenati, P. Mazzatorta, C.-D. Neagu and G. Gini, Combining classifiers of pesticides toxicity through a neuro-fuzzy approach, in *Proceedings 3rd Int. Workshop on Multiple Classifier Systems*, ed. F. Roli and J. Kittler, LNCS vol. 2364, Springer-Verlag, Berlin, 2002, pp. 293–303.

E. Benfenati, M. Clook, S. Friday and A. Hart, QSARs for regulatory purposes: the case for pesticide authorization, in *Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes*, ed. E. Benfenati, Elsevier Science Ltd, Amsterdam, 2007, pp. 1–57.

Benfenati E, S. S. Pardoe, T. M. Martin, R. Gonella Diaza, A. Lombardo, A. Manganaro and A. Gissi, Using toxicological evidence from QSAR models in practice, Altex, 2013, 30, 19–40. DOI:10.14573/altex.2013.1.019.

E. Benfenati, M. Belli, T. Borges, E. Casimiro, J. Cester, A. Fernandez, G. Gini, M. Honma, M. Kinzl, R. Knauf, A. Manganaro, E. Mombelli, M. I. Petoumenou, M. Paparella, P. Paris and G. Raitano, Results of a round-robin exercise on read-across, SAR QSAR Environ. Res., 2016, 27, 371–384. DOI:https://doi.org/10.1080/1062936X.2016.1178171.

E. Berggren, P. Amcoff, R. Benigni, K. Blackburn, E. Carney, M. Cronin, H. Deluyker, F. Gautier, R. S. Judson, G. E. N. Kass, D. Keller, D. Knight, W. Lilienblum, C. Mahony, I. Rusyn, T. Schultz, M. Schwarz, G. Schüürman, A. White, J. Burton, A. Lostia, S. Munn and A. Worth, Chemical safety assessment using read-across: how can novel testing methods strengthen evidence base for decision-making. Environ. Health Perspect., 2015, 123, 1232–1240.

Boland, P.J., 1989. Majority systems and the Condorcet jury theorem. Statistician 38, 181–189. https://doi.org/10.2307/2348873.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1023/A:1018054314350.

H. Buist, T. Aldenberg, M. Batke, S. Escher, R. K. Entink, R. Kühne, H. Marquart, E. Pauné, G. Schüürmann and D. Kroese, The OSIRIS weight of evidence approach: ITS mutagenicity and ITS carcinogenicity, Regul. Toxicol. Pharmacol., 2013, 67, 170–181.

C. I. Cappelli, A. Cassano, A. Golbamaki, Y. Moggio, A. Lombardo, M. Colafranceschi and E. Benfenati, Assessment of in silico models for acute aquatic toxicity towards fish under REACH regulation, SAR QSAR Environ. Res., 2015, 26, 977–999. DOI:https://doi.org/10.1080/1062936X.2015.1104519.

A Cassano, G. Raitano, E. Mombelli, A. Fernández, J. Cester, A. Roncaglioni and E. Benfenati, Evaluation of QSAR models for the prediction of Ames genotoxicity: a retrospective exercise on the chemical substances registered under the EU REACH regulation, J. Environ. Sci. Heal. C, 2014, 32, 273–298. DOI:https://doi.org/10.1080/10590501.2014.938955

Q. Chaudhry, J. Chrétien, M. Craciun, G. Guo, F. Lemke, J.-A. Müller, D. Neagu, N. Piclin, M. Pintore and P. Trundle, Algorithms for (Q) SAR model building, in *Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes,* ed. E. Benfenati, Elsevier Science Ltd, Amsterdam, 2007, pp. 111–148.

J. F. Contrera, N. L. Kruhlak, E. J. Matthews and R. D. Benz, Comparison of MC4PC and MDL-QSAR rodent carcinogenicity predictions and the enhancement of predictive performance by combining QSAR model, Regul. Toxicol. Pharmacol., 2007, 49,

172–182. DOI: https://doi.org/10.1016/j.yrtph.2007.07.001
M. T. D. Cronin, A.-N. Richarz and T. W. Schultz, Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction. Regul. Toxicol. Pharmacol., 2019, 106, 90–104. doi: https://doi.org/10.1016/j.yrtph.2019.04.007.
. T. Dietterich, Ensemble methods in machine learning, in *Multiple Classifier Systems—1st Int. Workshop, MCS 2000*, vol. 1857, ser. Lecture Notes in Computer Science, eds. J. Kittler and F. Roli, Cagliari, 2000,pp. 1–15.
M. Ellison, J. C. Madden, P. Judson and M. T. D. Cronin, Using in silico tools in a weight of evidence approach to aid toxicological assessment, Mol. Inform., 2010, 29, 97–110. DOI:https://doi.org/10.1002/minf.200900006.
Erturk, M.D., Turker Sacan, M., Nivic, M., Minovski, N., 2012. Quantitative structure-activity relationships (QSARs) using the novel marine algal toxicity data of phenols. J. Mol. Graph. Model. 38, 90–100.
European Chemicals Agency, 2016. Practical Guide. How to Use and Report (Q)SARs. pp. 1–37.
European Chemicals Agency, 2017a. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.7a: Endpoint Specific Guidance. ECHA-16-G-10-EN, Helsinki.
European Chemicals Agency, 2017b. Read-across Assessment Framework. ECHA-17-R-01-EN, Helsinki.
A. Fernández, A. Lombardo, R. Rallo, A. Roncaglioni, F. Giralt and E. Benfenati, Quantitative structure-activity relationships (QSARs) using the novel marine algal toxicity data of phenols, Environ. Int., 2012, 45, 51–58.
Fernández, A., Rallo, R., Giralt, F., 2015. Prioritization of in silico models and molecular descriptors for the assessment of ready biodegradability. Environ. Res. 142, 161–168. https://doi.org/10.1016/j.envres.2015.06.031.
Ferrari, T., Gini, G., 2010. An open source multistep model to predict mutagenicity from statistic analysis and relevant structural alerts. Chem. Cent. J. 4 (Suppl. 1), S2. http://www.journal.chemistrycentral.com/.
A.A. Frid and E. J. Matthews, Prediction of drug-related cardiac adverse effects in humans—B: use of QSAR programs for early detection of drug-induced cardiac toxicities, Regul. Toxicol. Pharmacol., 2010, 56, 276–289. DOI:https://doi.org/10.1016/j.yrtph.2009.11.005.
C. Gaudin, D. Cunha, E. Ivanoff, P. Horcajada, G. Chevé, A. Yasri, O. Loget, C. Serre and G. Maurin, A quantitative structure activity relationship approach to probe the influence of the functionalization on the drug encapsulation of porous metal-organic frameworks, Micropor. Mesopor. Mat., 2012, 157, 125–130. DOI:https://doi.org/10.1016/j.micromeso.2011.06.011A.
G. Gini, T. Garg and M. Stefanelli, Ensembling regression models to improve their predictivity: a case study in QSAR (quantitative structure activity relationships) with computational chemometrics, Appl. Artif. Intell., 2009, 23, 261–281. DOI:https://doi.org/10.1080/08839510802700847.
Gini, G., Benfenati, E., Testaguzza, V., Todeschini, R., 1998. Hytex (Hybrid Toxicology Expert System): architecture and implementation of a multi-domain hybrid expert system for toxicology. Chemom. Intell. Lab. 43, 135–145. https://doi.org/10.1016/S0169-7439(98)00125-7.
G. Gini, E. Benfenati, P. Grasso, M. Lorenzini and A. Vittore, Some results for the prediction of carcinogenicity using hybrid systems, in *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*, ed. G. Gini and A. Katrizky, AAAI Press, Menlo Park, 1999, pp 138–143.
G. Gini, A. M. Franchi, A. Manganaro, A. Golbamaki and E. Benfenati, ToxRead: a tool to assist in read across and its use to assess mutagenicity of chemicals, SAR QSAR Environ. Res., 2014a, 25, 999–1011. DOI: https://doi.org/10.1080/1062936X.2014.976267
G. Gini, A. M. Franchi, A. Manganaro, A. Golbamaki and E. Benfenati, ToxRead: a tool to assist in read across and its use to assess mutagenicity of chemicals, SAR QSAR Environ. Res., 2014b, 25, 999–1011. DOI:https://doi.org/10.1080/1062936X.2014.976267.
A. Gissi, O. Nicolotti, A. Carotti, D. Gadaleta, A. Lombardo and E. Benfenati, Integration of QSAR models for bioconcentration suitable for REACH, Sci. Total Environ., 2013, 456-457, 325–332. DOI: https://doi.org/10.1016/j.scitotenv.2013.03.104.
M. P. Gómez-Carracedo, R. Fernandez-Varela, D. Ballabio and J. M. Andrade, Screening oil spills by mid-IR spectroscopy and supervised pattern recognition techniques, Chemom. Intell. Lab., 2012, 114, 132–142. DOI:https://doi.org/10.1016/j.chemolab.2012.03.013.
R. Gonella Diaza, S. Manganelli, A. Esposito, A. Roncaglioni, A. Manganaro and E. Benfenati, Comparison of *in silico* tools for evaluating rat oral acute toxicity, SAR QSAR Environ. Res., 2015, 26, 1–27. DOI:https://doi.org/10.1080/1062936X.2014.977819.
E. Gottmann, S. Kramer, B. Pfahringer and *C. helma*, Data quality in predictive toxicology: reproducibility of rodent carcinogenicity experiments, Environ. Health Persp., 2001, 109, 509–514. DOI:https://doi.org/10.1289/ehp.01109509.
P. Gramatica, E. Giani and E. E. J. Papa, Statistical external validation and consensus modeling: a QSPR case study for Koc prediction, Mol. Graph. Model., 2007, 25, 755–766. DOI:https://doi.org/10.1016/j.jmgm.2006.06.005.
F. Grisoni, V. Consonni, S. Villa, M. Vighi and R. Todeschini, QSAR models for bioconcentration: is the increase in the complexity justified by more accurate predictions?, Chemosphere, 2015, 127, 171–179. DOI:https://doi.org/10.1016/j.chemosphere.2015.01.047.
A. Hardy, D. Benford, T. Halldorsson, M. J. Jeger, H. K. Knutsen, S. More, H. Naegeli, H. Noteborn, C. Ockleford, A. Ricci, G. Rychen, J. R. Schlatter, V. Silano, R. Solecki, D. Turck, E. Benfenati, Q. M. Chaudhry, P. Craig, G. Frampton, M. Greiner, A. Hart, C. Hogstrand, C. Lambre, R. Luttik, D. Makowski, A. Siani, H. Wahlstroem, J. Aguilera, J.-L. Dorne, A. F. T. Dumon, M. Hempen, S. Valtena Martınez, I. Martino, C. Smeraldi, A. Terron, N. Georgiadis and M. Younes, Guidance on the use of the weight of evidence approach in scientific assessments, EFSA J. 2017; 15, e04971.
H. Hrovat, H. Segner and S. Jeram, Variability of in vivo fish acute toxicity data, Regul. Toxicol. Pharmacol., 2009, 54, 294–300. DOI:https://doi.org/10.1016/j.yrtph.2009.05.013.
International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICM), 2017. Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk. M7 (R1). http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Multidisciplinary/M7/M7_R1_Addendum_Step_4_2017_0331.pdf.
G. Janer, W. Slob, B. C. Hakkert, T. Vermeire and A. H. Piersma, A retrospective analysis of developmental toxicity studies in rat and rabbit: what is the added value of the rabbit as an additional test species?, Regul. Toxicol. Pharmacol., 2008, 50, 206–217. DOI: https://doi.org/10.1016/j.yrtph.2007.11.007
R. Jolly, K. Begam Riaz Ahmed, C. Zwickl, I. Warson and V. Gombar, An evaluation of in-house and off-the-shelf in silico models: implications on guidance for mutagenicity assessment, Regul. Toxicol. Pharmacol., 2015, 71, 388–397. DOI:https://doi.org/10.1016/j.yrtph.2015.01.010
S. Kar and K. Roy, QSAR modeling of toxicity of diverse organic chemicals to Daphnia magna using 2D and 3D descriptors, J. Hazard. Mater., 2010, 177, 344–351. DOI: https://doi.org/10.1016/j.jhazmat.2009.12.038.
S. Kar and K. Roy, First report on development of quantitative interspecies structure carcinogenicity relationship models and exploring discriminatory features for rodent carcinogenicity of diverse organic chemicals using OECD guidelines, Chemosphere, 2012, 87, 339–355. DOI: https://doi.org/10.1016/j.chemosphere.2011.12.019.
A. R. Katritzky, M. Kuanar, S. Slavov, D. A. Dobchev, D. C. Fara, M. Karelson, W. E. Jr Acree, V. P. Solove'v and A. Varnek, Correlation of blood-brain penetration using structural descriptors, Bioorg. Med. Chem., 2006, 14, 4888–4917. DOI:https://doi.org/10.1016/j.bmc.2006.03.012.
C. Koenig, G. Gini, M. Craciun and E. Benfenati, Multi-class classifier from a combination of local experts: toward distributed computation for real-problem classifiers, Int. J. Pattern Recogn., 2004, 18, 801–817.
S. A. Kulkarni, E. Benfenati and T. Barton-Maclaren, Improving confidence in (Q)SAR predictions under Canada's chemicals management plan – a chemical space approach, SAR QSAR Environ. Res., 2016, 27, 851–863. DOI:https://doi.org/10.1080/1062936X.2016.1243152.
A. Kumar Gupta, N. Sabarwal, Y. P. Agrawal, S. Prachand and S. Jain, Insights through AM1 calculations into the structural requirement of 3,4,6-substituted-2-quinolone analogs towards FMS kinase inhibitory activity Eur. J. Med. Chem., 2010, 45, 3472–3479. DOI: https://doi.org/10.1016/j.ejmech.2010.05.001.
Kuncheva, L.I., 2005. Combining Pattern Classifiers, Methods and Algorithms. Wiley Interscience, New York.
C. Kuseva, W. Schultz, D. Yordanova, H. Ivanova, K. Tankova, T. Pavlov, A. Chapkanov, G. Chankov, M. Georgiev, A. Gissi, T. Sobanski and O. G. Mekenyan, Category consistency in the OECD QSAR toolbox: assessment and reporting tool to justify read-across. Computational Toxicology, 2019, 11, 65–71.
B. Lei, L. Xi, J. Li, H. Liu and X. Yao, Global, local and novel consensus quantitative structure-activity relationship studies of 4-(Phenylaminomethylene) isoquinoline-1, 3 (2H, 4H)-diones as potent inhibitors of the cyclin-dependent kinase 4, Anal. Chim. Acta, 2009, 644, 17–24. DOI:https://doi.org/10.1016/j.aca.2009.04.019
Y. Li, X. Shao and W. Cai, A consensus least squares support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples, Talanta 2007, 72, 217–222. DOI:https://doi.org/10.1016/j.talanta.2006.10.022.
Li, Y., Shao, X., Cai, W., 2007. A consensus least squares support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples. Talanta 72, 217–222. https://doi.org/10.1016/j.talanta.2006.10.022.
Lombardo, A. Roncaglioni, E. Boriani, C. Milan and E. Benfenati, Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish, Chem. Cent. J., 2010, 4 (Suppl 1): S1. DOI:https://doi.org/10.1186/1752-153X-4-S1-S1.
A. Manganaro, F. Pizzo, A. Lombardo, A. Pogliaghi and E. Benfenati, Predicting persistence in the sediment compartment with a new automatic software based on the **k**-nearest neighbor (**k**-NN) algorithm, Chemosphere, 2016, 144, 1624–1630. DOI:https://doi.org/10.1016/j.chemosphere.2015.10.054.
K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A. M. Richard, C. M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E. B. Wedebye, F. Grisoni, G. F. Mangiatordi, G. M. Incisivo, H. Hong, H. W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N. G. Nikolov, O. Nicolotti, P. L. Andersson, Q. Zang, R. Politi, R. D. Beger, R. Todeschini, R. Huang, S. Farag, S. A. Rosenberg, S. Slavov, X. Hu and R. J. Judson, CERAPP: collaborative estrogen receptor activity prediction project, Environ. Health Perspect., 2016, 124, 1023–1033. DOI:https://doi.org/10.1289/ehp.1510267.
M. Marzo, S. Kulkarni, A. Manganaro, A. Roncaglioni, S. Wu, T. Barton-Maclaren, C. Lester and E. Benfenati, Integrating in silico models to enhance predictivity for developmental toxicity, Toxicology, 2016, 370, 127–137. DOI:https://doi.org/10.1016/j.tox.2016.09.015
Mastrandrea, M.D., Field, C.B., Stocker, T.F., Edenhofer, O., Ebi, K.L., Frame, D.J., Held, H., Kriegler, E., Mach, K.J., Matschoss, P.R., Plattner, G.-K., Yohe, G.W., Zwiers, F.W., 2010. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. Intergovernmental Panel on Climate Change (IPCC) Available at. http://www.ipcc.ch.
C. Milan, O. Schifanella, A. Roncaglioni and E. Benfenati, Comparison and possible use of in silico tools for carcinogenicity within REACH legislation, J. Environ. Sci. Heal. C, 2011, 29, 300–323. DOI:https://doi.org/10.1080/10590501.2011.629973.
A. Morales Helguera, A. Pérez-Garrido, A. Gaspar, J. Reis, F. Cagide, D. Vina, M. Natalia, D.S. Cordeiro and F. Borges, Combining QSAR classification models for predictive

modeling of human monoamine oxidase inhibitors, Eur. J. Med. Chem., 2013, 59, 75–90. DOI:https://doi.org/10.1016/j.ejmech.2012.10.035.

E. N. Muratov, V. E. Kuzmin, A. G. Artemenko, N. A. Kovdienko, L. Gorb, F. Hill and J. Leszczynski, New QSPR equations for prediction of aqueous solubility for military compounds, Chemosphere, 2010, 79, 887–890.

. C.-D. Neagu and G. Gini, Neuro-fuzzy knowledge integration applied in toxicity prediction, in *Innovations in knowledge engineering*, ed. R. Jain, A. Abraham, C. Faucher and B. J. ven der Zwaag, Advanced Knowledge International, Magill, 2003, pp. 311–342.

Organisation for Economic Cooperation and Development (OECD), Environmental Directorate, 2004. Report from the Expert Group on Quantitative Structure – Activity Relationships (Q)SAR on the Principles of Validation of (Q)SARs. (ENV/JM/MONO), Series on Testing and Assessment, No. 49.

E. Papa, L. van der Wal, J. A. Arnot and P. Gramatica, Metabolic biotransformation half-lives in fish: QSAR modeling and consensus analysis. Sci. Total Environ., 2014, 470-471, 1040–1046. DOI: https://doi.org/10.1016/j.scitotenv.2013.10.068

G. Patlewicz, G. Helman, P. Pradeep and I. Shah, Navigating through the minefield of read-across tools: a review of in silico tools for grouping, Computational Toxicology, 2017, 3, 1–18. DOI:https://doi.org/10.1016/j.comtox.2017.05.003.

Piegorsch, W.W., Zeiger, E., 1991. Measuring intra-assay agreement for the Ames salmonella assay. In: Hotorn, L. (Ed.), Statistical Methods in Toxicology, Lecture Notes in Medical Informatics. Springer-Verlag, pp. 35–41.

F. Pizzo, A. Lombardo, M. Brandt, A. Manganaro and E. Benfenati, A new integrated in silico strategy for the assessment and prioritization of persistence of chemicals under REACH, Environ. Int., 2016, 88, 250–60.DOI:https://doi.org/10.1016/j.envint.2015.12.019.

Polikar, R., 2006. Ensemble based systems in decision making. IEEE Circ. Syst. Mag. 6, 21–45. https://doi.org/10.1016/S0893-6080(05)80023-1.

N. Price and Q. Chaudhry, Application of in silico modelling to estimate toxicity of migrating substances from food packaging, Food Chem. Toxicol., 2014, 71, 136–141. DOI:https://doi.org/10.1016/j.fct.2014.05.022.

Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency).

E. Rorije, T. Aldenberg, H. Buist, D. Kroese and G. Schüürmann, The OSIRIS weight of evidence approach: ITS for skin sensitisation. Regul. Toxicol. Pharmacol., 2013, 67, 146–156. DOI:https://doi.org/10.1016/j.yrtph.2013.06.003.

K. Roy. P. Ambure, S. Kar and P. Kumar Ojha, Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models? J. Chemom., 2018, 32, e2992.

P. Ruiz, A. Sack, M. Wampole, S. Bobst and M. Vracko, Integration of in silico methods and computational systems biology to explore endocrine-disrupting chemical binding with nuclear hormone receptors, Chemosphere, 2017, 179, 99–109. DOI: https://doi.org/10.1016/j.chemosphere.2017.03.026.

O. Santos-Filho and A. J. Hopfinger, Combined 4D-fingerprint and clustering based membrane-interaction QSAR analyses for constructing consensus Caco-2cell permeation virtual screens J. Pharm. Sci., 2008, 97, 566–582. DOI:https://doi.org/10.1002/jps.21086.

T. W. Schultz and M. T. D. Cronin, Lessons learned from read-across case studies for repeated-dose toxicity. Regul. Toxicol. Pharmacol., 2017, 88, 185–191.

T. W. Schultz, A.-N. Richarz and M. T. D. Cronin, Assessing uncertainty in read-across: questions to evaluate toxicity predictions based on knowledge gained from case studies. Computational Toxicology, 2019, 9, 1–11.

T. W. Schultz, P. Amcoff, E. Berggren, F. Gautier, M. Klaric, D. J. Knight, C. Mahony, M. Schwarz, A. White and M. T. D. Cronin, A strategy for structuring and reporting a read-across prediction of toxicity. Regul. Toxicol. Pharmacol., 2015, 72, 586–601.

S. Sciabola, E. Carosati, L. Curucull-Sanchez, M. Baroni and R. Mannhold, Novel TOPP descriptors in 3D-QSAR analysis of apoptosis inducing 4-aryl-4**H**-chromenes: comparison versus other 2D- and 3D-descriptors, Bioorg. Med. Chem., 2007, 15, 6450–6462. DOI: https://doi.org/10.1016/j.bmc.2007.06.051.

R. Serafimova, M. F. Gatnik and A. Worth, *Review of QSAR Models and Software Tools for Predicting Genotoxicity and Carcinogenicity*, European Commission, Joint Research Centre, Institute for Health and Consumer Protection, Luxembourg: Publications Office of the European Union, EUR 24427 EN, 2010.

J. W. Van der Veen, E. Rorije, R. Emter, A. Natsch, H. van Loveren and H. Ezeridam, Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals, Regul. Toxicol. Pharmacol., 2014, 69, 371–379. DOI:https://doi.org/10.1016/j.yrtph.2014.04.018.

Price, N., Chaudhry, Q., 2014. Application of in silico modelling to estimate toxicity of migrating substances from food packaging. Food Chem. Toxicol. 71, 136–141. https://doi.org/10.1016/j.fct.2014.05.022.

Wolpert, D., 1992. Stacked Generalization. Neural Networks 5, 241–259.

Q. Zhang, J. M. Hughes-Oliver and R. T Ng, A model-based ensembling approach for developing QSARs. J. Chem. Inf. Model., 2009, 49, 1857–1865.

L. Zhao, Y. Xiang, J. Song and Z. Zhang, A novel two-step QSAR modeling work flow to predict selectivity and activity of HDAC inhibitors. Bioorg. Med. Chem. Lett., 2013, 23, 929–933. DOI: https://doi.org/10.1016/j.bmcl.2012.12.067.