



Using Toxicological Evidence from QSAR Models in Practice

Emilio Benfenati¹, Simon Pardoe², Todd Martin³, Rodolfo Gonella Diaz¹, Anna Lombardo¹, Alberto Mangano¹, and Andrea Gissi¹

¹Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy; ²PublicSpace Ltd, Lancaster, UK; ³US Environmental Protection Agency, Cincinnati, OH, USA

Summary

Leading QSAR models provide supporting documentation in addition to a predicted toxicological value. Such information enables the toxicologist to explore the properties of chemical substances as well as to review and to increase the reliability of toxicity predictions. This article focuses on the use of this information in practice. We explore the supporting documentation provided by the EPISuite, T.E.S.T. and VEGA platforms when evaluating the bioconcentration factor (BCF) of three example compounds. Each compound presents a different challenge: to recognize high reliability, analyze complex evidence of reliability, and recognize uncertainty. In each case, we first describe and discuss the supporting documentation provided by the QSAR platforms. We then discuss the judgments on reliability across sectors from 28 toxicologists who used this supporting information and commented on the process. The article demonstrates both the use of QSAR models as tools to reduce or replace in vivo testing, and the need for scientific expertise and rigor in their use.

Keywords: bioconcentration, QSAR, REACH

1 Introduction

With its strong focus on encouraging innovation in toxicity evaluation and its explicit demand to use existing experimental data where possible, the 2007 EU legislation on the “Registration, Evaluation, Authorisation and restriction of Chemicals” (REACH) has stimulated constructive discussions in Europe on the development and use of QSAR models for regulatory purposes. In the US and Canada, QSAR models have been used for decades, and the US Environmental Protection Agency has itself developed a number of models for regulatory use (including EPISuite and T.E.S.T.).

QSAR models are tools for the toxicologist to explore the properties of chemical substances. In effect, they maximize the insight from existing experimental data to enable predictions about further compounds that have not been tested. QSAR models developed by US regulators and by EU funded research are made freely available online.

The evolving regulatory demands in the EU and elsewhere have prompted development in the reliability and the supporting evidence provided by leading QSAR platforms. Such development is not only vital to meet regulatory demands, but is also essential to the role of QSAR models in providing toxicological evidence to reduce and replace costly *in vivo* testing.

In this article we focus critically on aspects of the evidence and information provided by QSAR models and their use *in practice*.

In three case studies, we explore the information provided by the EPISuite, T.E.S.T., and VEGA platforms for evaluating the bioconcentration factor (BCF) for three example compounds. We address the central practical challenges of how to determine whether a toxicity prediction for a compound is reliable, as well as how to discuss the predictions and supporting information so that the evidence and reasoning can be understood, reviewed, and potentially accepted by others. This investigation is the first inter-laboratory and cross-institutional review exercise on the reliability of QSAR results in the evaluation of an animal model. As far as we know, it is the first exercise of this nature.

In sections 2-6 below, we start by explaining the rationale for this exercise: why the practical case-by-case use of QSAR models is an important focus for investigative review, especially in light of the EU regulatory demands from REACH. We then describe the methods of this investigation, including the significance of the three example compounds as case studies, the features of the QSAR platforms used, and, finally, the issue of reliability for experimental BCF results and predictions.

Taking each compound as a separate case study, we describe the BCF predictions offered by each QSAR platform and the range of information provided to support those predictions. We discuss and illustrate the use of this information by toxicologists to reach a decision on the reliability of the prediction. We then discuss the decisions on reliability made by the 28 participants and the comments they made about their decisions. Having

Received July 20, 2012; accepted in revised form November 02, 2012.



reviewed the predictions and information from each platform separately, we then illustrate the potential value of exploring a “consensus” across platforms.

The article demonstrates the practical use of QSAR models as tools to develop predictions of toxicity that can potentially reduce or replace *in vivo* testing. In doing so, it shows the need for scientific expertise and rigor to ensure quality and reliability in their use. The intention is to stimulate discussion on the acceptable use of predictions from QSAR models in practice. The article forms part of a wider project (<http://www.orchestra-qsar.eu>) in which our focus is on promoting an understanding of QSAR models to inform and develop good scientific practice in their use.

2 Rationale: why analyze the practical use of QSAR models?

2.1 QSAR models

The concept of “Structure-Activity Relationship” (SAR) is that the biological activity of a chemical in the human body and the environment (including its toxicity) can be related to its molecular structure and physical-chemical properties. When quantified, this relationship is known as “QSAR.” The acronym “(Q)SAR” is used to cover both. A QSAR model makes use of existing experimental toxicity data for a series of chemicals. By using potentially complex algorithms, the model correlates experimentally observed toxicity with aspects of molecular structure and physical-chemical properties across a series of related compounds in order to predict the toxicity of further chemicals with related molecular structures.

When toxicologists use one of the leading platforms they receive predictions and supporting information from several QSAR models relevant to the particular compound and the particular toxicological endpoint. The reason for employing several QSAR models within a software program is that each model uses a different correlation algorithm; together, these algorithms can increase and cross-check the reliability of the prediction. Crucially, it means that difficulties in prediction, and consequent lower levels of reliability, are revealed to the user by a lack of consistency across the values from each model. We illustrate this in the case studies below.

2.2 The EU regulatory focus on use in practice, case-by-case

The European Chemicals Agency (ECHA) is responsible for the implementation of the REACH regulations and has produced detailed guidance documents on the regulatory use of QSAR models. Its focus is particularly on the use of QSAR models within a “weight of evidence” approach, in which toxicological assessment moves away from the traditional “single *in vivo* test” approach. (See Benfenati et al., 2011 for an outline of issues and challenges in the regulatory acceptance of *in silico* methods.)

Under REACH, ECHA requires industry to evaluate all existing information, including QSAR results, and to use further *in vivo*, *in vitro*, and *in silico* assessments where necessary. The policy emphasis is that *in vivo* tests are to be approved on verte-

brates only “as a last resort” when the necessary evidence cannot be produced by alternative methods.

The REACH legislation specifies four *requirements* for the use of (Q)SAR models for regulatory purposes (ECHA, 2008):

- 1) results are derived from a (Q)SAR model whose scientific validity has been established,
- 2) the substance falls within the applicability domain of the (Q)SAR model,
- 3) results are adequate for the purpose of classification and labeling and/or risk assessment,
- 4) adequate and reliable documentation of the applied method is provided.

The regulatory focus, therefore, is not only on (1) the scientific validity of the model but on three further issues that relate to the way the model is being used in a particular case. These are: (2) what chemical substance is being investigated and therefore whether the chosen model is suitable, (3) what regulatory purpose the results are being used for and therefore whether the results meet the demands, and (4) whether the documentation of both the model and its use are sufficient to enable the regulator to make an independent and informed judgment on the reliability of the prediction. To be able to ensure each of these requirements is met, ECHA does not intend to approve a list of QSAR models for use within REACH (e.g., ECHA, 2009). Instead, acceptance will depend on their use in practice on a “*case-by-case*” basis.

Acceptance case-by-case reflects the fact that a QSAR model is developed from experimental results for a particular toxicological endpoint, for a particular series of chemicals. A model, therefore, is intended for use for that same endpoint and for a range of similar chemicals. This is its “applicability domain” (2). If a model is used outside its applicability domain, the reliability of a prediction will be lower.

Case-by-case acceptance also addresses an evident fear that QSAR models might be used uncritically and without an understanding of toxicology. It is a fear sometimes summarized by the phrase “just press a button and get an answer.” While this phrase is presented by some toxicologists as if it constitutes a critique of QSAR models, it is important to recognize that it is actually an anxiety about the human use of models *in practice*. Hence the policy of acceptance *case-by-case* demands quality not only within the QSAR models but also in their professional and scientific use.

QSAR models are no exception in demanding expertise in their use. *In vivo* and *in vitro* tests require expertise and adherence to protocols. Across the sciences, accurate results require method, expertise, and care; the interpretation of results to draw conclusions requires knowledge, expertise, caution, and judgment. Ensuring rigor requires some systematic skepticism and scrutiny of each result and each conclusion. Hence using QSAR models, like using any other models in science, is certainly not about the mere acceptance of a number produced by a computer. It requires expert assessment and judgment.

2.3 Increasing model quality and regulatory acceptance by a focus on practice

The regulatory focus on the quality of QSAR predictions and of the use of QSAR models *in practice* demands a similar fo-

cus from model developers and from advocates of their regulatory use. To this end, leading QSAR models provide detailed supporting information in addition to a predicted toxicity value or classification. This information enables the toxicologist to explore the properties of chemical substances and therefore to review and increase the reliability of toxicity predictions. The intention is that this supporting information be carefully analyzed by the toxicologist who is ultimately responsible for the assessment and for the regulatory dossier submission.

The supporting information is provided in recognition that there is a regulatory demand for three categories of documentation:

- a) The predicted value or class produced by the QSAR model;
- b) The further documentation provided by the QSAR model;
- c) A document prepared by the human expert, analyzing and concluding from the first two items.

These three demands have several implications. First, the old and misguided antagonism towards computers by some toxicologists finally needs to be put to rest. To put it simply, the QSAR model *helps* the experts in their job; it does not *do* their job. QSAR models have to be understood as a valuable tool that offers the expert advanced ways of exploring the properties and features of chemical substances.

A second implication is that the human expert has to analyze the prediction and the other material provided by the model and take responsibility for the final judgment. In particular, given the importance of the applicability domain, it is the responsibility of the user to ensure that the target compound is within the domain and to evaluate the reliability of the outputs for the specific compound. A third implication within the EU regulatory process is that if the applicant submits QSAR predictions without suitable explanation and discussion within the documentation, then it is likely that the conclusions will not be accepted by regulators. (The same applies to results from read-across methods.)

The three demands also help to clarify the *areas of research activity* that are needed by developers of QSAR models to ensure their quality and to increase regulatory acceptance of their use in practice:

- i) A first vital area is to prove the scientific validity of models (REACH requirement 1, see section 2.2) by reviewing their predictive performance. For this reason, for example, the EC funded project ANTARES (<http://www.antes-life.eu>) is carefully checking the performance of about 50 QSAR models for different endpoints (including BCF) using large datasets to verify the model predictivity.
- ii) A second area of work is to develop QSAR models and platforms oriented to regulatory demands. A platform has to provide clear information on the applicability domain (REACH requirement 2), provide results that are adequate for regulatory decisions of risk and classification (3), and provide the level of supporting information and documentation that is necessary for the user and regulator to be able to evaluate that adequacy in each case (4). That is the aim, for example, in developing the VEGA QSAR platform.
- iii) A third area of work is to review and demonstrate the current “state-of-the-art” in terms of what supporting information QSAR models provide for the user. This involves comparative analysis of the outputs from leading platforms and in-

cludes reviewing the extent to which they provide complementary evidence based on different experimental data and algorithms. It can increase expectations and so generate new “benchmarks” in what QSAR platforms are expected to provide for regulatory use.

- iv) A fourth area of work is to review user experience of the QSAR models *in practice* and, specifically, to review user evaluations of the predictions and confidence in them. Given the case-by-case nature of reliability, this must be done through case studies evaluating particular compounds and particular endpoints. Such work can contribute to identifying practical feedback on QSAR platforms, with associated demands, but also can help to identify the kinds of toxicological expertise and rigor needed from users.

In this article, our focus is on the third and fourth areas of work.

3 Materials and methods: three case studies

3.1 Three case studies using three example compounds

The intention of the “review exercise” was to generate three potentially valuable case studies of the *use* of supporting information to judge the reliability of QSAR predictions. The three example compounds, therefore, were chosen to present different challenges to raise and illustrate different issues for the user in reviewing the prediction. Each compound offers a useful and different case study.

Chemical 1: a case study in recognizing and documenting *high reliability*

Chemical 1 is predicted for BCF with high reliability. It was chosen for this study (i) to provide participant users with an example of what very high reliability looks like, and (ii) to provide an important case study of users recognizing and documenting a prediction with very high reliability. It enables us to investigate whether the information provided by the QSAR models is sufficient to convince the wisely cautious toxicologist of that reliability, and if not, what further information is needed. (Note: while all three platforms offer a prediction, an experimental result for this compound was actually used in the development of the T.E.S.T. and EPISuite models as part of the training sets.)

Chemical 2: a case study in analyzing and documenting *complex reliability*

Chemical 2 presents a more typical example for the use of QSAR models in practice, where the compound lies within the applicability domain and where the user reviews a range of detailed evidence to make an informed judgment. We judge the BCF predictions for Chemical 2 to be reliable. However, different elements within the evidence suggest unreliability and reliability, so the user has to analyze the material using toxicological understanding. It therefore provides a valuable case study in analyzing and documenting complex reliability.

Chemical 3: a case study in recognizing and documenting *uncertainty*

Chemical 3 has a more difficult molecular structure in terms of BCF prediction, and all three platforms communicate



the considerable difficulty in predicting consistent values. We judge the predictions for this compound to be unreliable. It therefore provides an important case study of whether the current models enable users to recognize and document such a lack of reliability.

The predictions and supporting information for the three compounds are described and partly reproduced below. The full information provided to the users is included in supplementary files 1, 2, and 3 at <http://www.altex-edition.org>. (The annexes can be used in combination with this article to produce a useful training resource.)

3.2 Terminology: QSAR “platforms,” “programs,” and “models”

There is often some fluidity in the ways in which the term “model” is used in the literature, from referring to the software platform as a whole, to the software program designed for a particular endpoint, to the particular models that make up that program. This fluidity is useful when simply referring to QSAR models in contrast to other methods, but it can become ambiguous and potentially confusing when discussing models in more detail. It then becomes necessary to differentiate the levels.

In this article, therefore, we will refer to EPISuite, T.E.S.T., and VEGA as three QSAR model “platforms,” rather than “models.” Each platform contains a range of software “programs,” each of which is designed to be used to process the chemical structure, to generate chemical descriptors, to measure the similarity between different compounds, to evaluate the applicability domain, or to predict a chemical property with a certain QSAR model or set of models. Each platform, therefore, offers a range of QSAR *models* and may include one or more QSAR models for each endpoint (such as BCF, skin sensitization, carcinogenicity, or mutagenicity). Typically, each QSAR model has been developed with its own training set, algorithm, and test set and therefore has its own intended applicability domain. In some cases, a QSAR model is made applicable to a range of chemicals by being composed of sub-models that each address specific chemical categories.

3.3 The BCF model platforms used

The review exercise used QSAR models designed for evaluating BCF on the following three platforms:

- EPISuite (<http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm>), explained at <http://www.epa.gov/oppt/exposure/pubs/episuite.htm>
- T.E.S.T. (<http://www.epa.gov/nrmrl/std/qsar/qsar.html>), explained at <http://www.epa.gov/nrmrl/std/qsar/TEST-user-guide-v41.pdf>
- VEGA (<http://www.vega-qsar.eu>), explained at http://www.vega-qsar.eu/guidelines/VEGA_model_guide_BCF_2_1_8.pdf

These three platforms were chosen because (i) they are all freely available to users, (ii) they represent leading platforms in Europe and the US, (iii) they all provide models for evaluating BCF, and (iv) we are interested in reviewing the use of the newer VEGA platform alongside other leading platforms. The T.E.S.T. and EPISuite platforms have been developed by the US EPA within different centers. The VEGA platform has been

developed by a European consortium coordinated by the Istituto Mario Negri in Milan, and it incorporates the earlier established CAESAR platform.

EPISuite

Of the three platforms, EPISuite offers the oldest set of QSAR models for BCF, although the program has been updated periodically. The program (called BCFBAF) includes models based on logP (the partition coefficient between octanol and water). LogP is calculated through a series of fragments present in the molecule, so the BCF value is calculated using these fragments. The first kind of model (called Meylan model) used within EPISuite for BCF is based on a few sub-models based on logP. Some of these apply for ionic compounds. The sub-models are selected and used by the system depending on the range of the logP value. The second kind of model, based on a model by Arnot-Gobas, takes into account the metabolic transformation in fish. Coefficients are used to optimize the regression curve using fragments that have been associated with different levels of metabolic rate.

Within EPISuite, the user has to select and use the appropriate model for a particular class of compounds. To enable this, the applicability domain of each model is defined by the range of the logP and molecular weight values. It has to be checked manually by the user, and where this is potentially uncertain, the user needs to make a judgment.

T.E.S.T.

T.E.S.T. has several models for BCF that are selected automatically for the user. T.E.S.T. then also has a consensus model that integrates the results from the individual models. The system checks the reliability of each model automatically, and if one or more of the models has a poor performance for a certain compound, then it is not included within the consensus model. For example, the hierarchical clustering method employs three different applicability domain measures to determine whether a prediction can be made by each cluster model. The model ellipsoid constraint checks whether the leverage of the test chemical is less than the maximum leverage for the chemicals in the model (in terms of the descriptors appearing in the model). The R_{\max} constraint checks if the distance from the test chemical to the centroid of the cluster is less than the maximum distance for any chemical in the cluster to the cluster centroid (in terms of the entire set of descriptors). Finally, the fragment constraint checks whether the compounds in the cluster have at least one example of each of the fragments contained in the test chemical. For example, if one was trying to make a prediction for ethanol, the cluster must contain at least one compound with a methyl fragment ($-\text{CH}_3$ [aliphatic attach]), one compound with a methylene fragment ($-\text{CH}_2$ [aliphatic attach]), and one compound with a hydroxyl fragment ($-\text{OH}$ [aliphatic attach]). Different models are based on different algorithms, and each employs a particular series of molecular and physical-chemical descriptors. The intentional differences between the models therefore mean that they function together within the consensus model to increase reliability. LogP is just one of the descriptors used by T.E.S.T.. T.E.S.T. also has implemented the nearest neighbors model, which predicts the

BCF value of the target compound on the basis of the experimental values of the most similar compounds in its database. This allows the user to determine whether the training set possesses chemicals similar to that of the chemical being analyzed. Furthermore, the T.E.S.T. platform lists predicted values for those similar compounds in the test set (with a similarity value >0.5) alongside their experimental values, so that the user can assess the quality of the prediction for the target compound.

VEGA

VEGA has incorporated the BCF models from the CAESAR platform (<http://www.caesar-project.eu/>) and further developed the information provided. For BCF, CAESAR uses two models, and then a third model which uses the input of these two models to calculate the final BCF value. (The final value is not simply the average of the results from the two models, or simply the worst case; instead it is predicted through a more analytical process.) VEGA improves the evaluation of the position of the target compound within the applicability domain by (i) plotting the BCF results against the logP values for all the compounds in the training set, by (ii) offering detail for the three most similar compounds, and by (iii) showing particular molecular fragments of interest. (There are two kinds of molecular fragments of interest: fragments which identify a chemical class for which the prediction is more uncertain, and fragments which can be useful when evaluating the results because they are associated with a certain role in BCF behavior, such as polarity.) Like T.E.S.T., VEGA also shows the experimental and predicted values for the most similar compounds in the data set.

VEGA includes an “applicability domain index” which automatically checks whether there are critical issues that may produce error and/or reduce the reliability of the prediction. These checks include (a) whether there is sufficient similarity between the target compound and the most similar compounds in the dataset, (b) whether the experimental values of the similar compounds are close to the prediction for the target compound to be evaluated – “concordance”, (c) whether the experimental values of the similar compounds are close to the predictions for those compounds – “accuracy”, (d) whether there are molecular fragments in the target compound that are uncommon in the compounds within the training set, (e) whether the descriptor values for the target compound are within the range of the descriptor, and (f) whether changing the values of any of the descriptors by 10% results in a large modification of the predicted value. Furthermore, the software automatically checks for fragments that identify a chemical class for which the prediction is more uncertain and reports this as a warning. On the basis of all these checks, this “applicability domain index” calculates a “safety value” for the prediction.

3.4 Reliability for regulatory purposes

Uncertainty is present in all experimental and predictive values in the life sciences. In toxicology, *in vivo*, *in vitro*, and *in silico* methods all involve predicting complex environmental and human toxicity by using controlled laboratory experiments as “models.” All three sets of methods therefore involve recognizing and addressing issues of similarity, reliability, and uncertainty.

In evaluations for regulatory purposes, issues of reliability also are affected by the thresholds introduced by the legislation. The significance of a margin of error will depend on the proximity of the experimental or predicted value to the threshold. If the value is very close to the regulatory threshold, then we may need a high level of reliability (low uncertainty). Conversely, if the value is very far from the threshold, we may accept lower reliability (higher uncertainty).

The reliability of a toxicity evaluation, therefore, must be reviewed according to the purpose of the evaluation. It may be necessary to evaluate the compound in relation to legal thresholds, or it may be necessary to produce a continuous toxicity value, as in risk assessment. Under REACH, the regulatory demands on the evaluation of a compound depend on the annual tonnage on the market for each registrant. For these reasons, in the examples below we address reliability in terms of proximity to thresholds, as well as in terms of the reliability of the value itself.

3.5 Uncertainty levels in experimental *in vivo* results for BCF

In the evaluation of the BCF, the current uncertainty within experimental results from *in vivo* tests is between 0.40 and 0.75 log units (Zhao et al., 2008). A more recent study identified 0.6 as the typical uncertainty (Lombardo et al., 2010). (We use this latter figure below.) Participants in the review were informed of this experimental uncertainty.

This experimental uncertainty is regarded as acceptable and as the best achievable level of reliability. Therefore, it has to be kept in mind as a kind of benchmark when reviewing the reliability and acceptability of predictions from QSAR models. It is also important to remember that uncertainty exists within the experimental data on which the QSAR models are based, though QSAR models have the advantage of using a range of experimental results rather than relying on a single result.

3.6 The descriptions and discussion of the supporting information

For the three example compounds (the three case studies), we describe the supporting information that each platform provides to accompany the BCF prediction and discuss the use of this information in assessing the reliability of the BCF prediction. (To avoid repetition, some explanations are in more detail for Chemical 1.) This section, therefore, shows the information that was available to the participants, on which they based their judgment.

The discussions address the central practical challenges of how to determine whether a toxicity prediction for a compound is reliable and how to discuss the predictions and supporting information so that the evidence and reasoning can be understood, reviewed, and potentially accepted by others. They illustrate the kind of analytical decision-making processes that we (as model developers) envisage in the use of QSAR models. The comments are by Emilio Benfenati, as an author of the VEGA platform, with contributions from Todd Martin, as an author of the T.E.S.T. platform. (The authors of EPISuite were invited to contribute, but were unable to do so.)



These discussions cannot be viewed as an ideal, however, or as a recipe for regulatory acceptance. Every analysis and decision-making process draws on the user's toxicological expertise and must be oriented to particular compounds, particular toxicity endpoints, and particular regulatory demands. For these reasons the discussions here are intended to generate understanding and to stimulate discussion on the process of using predictions from QSAR models in regulatory toxicology. There is a clear need to support new users in their use of QSAR models and to counter fears about computer outputs somehow replacing toxicological expertise. We hope this illustration is useful in both ways and that it will prompt other case study illustrations. The priority is to develop and promote good scientific practice in the use of QSAR models.

3.7 The expert participants and their review

The purpose of the user review, stated to participants, was to obtain feedback from users (regulators, industry, consultants, researchers, and experts within academia) on the reliability and acceptability of results from QSAR models by considering specific and practical case studies.

51 individual experts in QSAR and/or BCF were contacted by e-mail, mainly in Europe. A total of 28 completed replies were received from 11 countries: Austria, Belgium, France, Germany, Italy, Portugal, Slovenia, Spain, Switzerland, UK, and USA. Three groups provided replies from several experts within the same group. This was an encouraging rate of response, reflecting the interest in the issue.

Participation in the exercise was anonymous, using the online form shown in supplementary file 5 at <http://www.altex-edition.org>. The software recorded those who replied, and allocated a number to each participant, but did not associate the responses to the name of the participant. It was made clear to participants that the exercise was NOT about the validation of the models; it was only about the acceptance of the outputs by the range of stakeholders. Equally, the purpose of the exercise was NOT to evaluate the model quality or predictive performance; this requires a much deeper study and is an on-going focus within the EC ANTARES project (<http://www.antes-life.eu>). However, we do report participant comments on the experience of using the three platforms when these were given as explanations for the users' judgments on reliability. When we consider that it may be useful for the reader, we also offer explanations in response to participant comments (below and supplementary file 6 at <http://www.altex-edition.org>).

3.8 The participants' review procedure

For each compound, participants were asked:

On the basis of all the pieces of information provided, do you consider the BCF value obtained from the model to be sufficiently reliable?

☐ Yes

☐ No

Comment (optional)

Participants were asked to review the predication and supporting information provided by each platform separately. This was to ensure that judgments of reliability were based on the

evidence provided by each platform, and not on a consensus of predictions across the platforms. The results indicate that the participants followed this request. Participants often drew different conclusions for a compound across the three platforms. This is most apparent in the results for Chemical 1 (see section 4.2), but even for Chemical 2 a full 8 of the 20 respondents gave different responses across the three platforms.

A judgment about the reliability of a prediction requires confidence that the compound is within the domain of applicability of the model, and confidence in making sense both of the prediction and of the range of supporting evidence and documentation. If there is doubt and/or a lack of evidence to confirm reliability, then it is both wise and good practice to judge that the prediction is not reliable. Faced with a yes/no decision on reliability, the participants' comments indicate that in cases where they could not be confident in the evidence of reliability they erred on the side of caution and rightly made the decision that the result was *not* reliable. As requested, they did so even when they had concluded that a similar result from another platform was reliable.

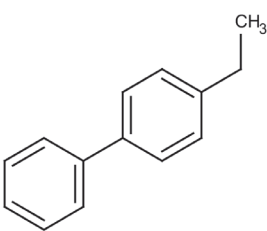
Participants were not required to produce the predictions and supporting information from the platforms. Instead, these were provided as online files. This made the review quicker and less onerous for the participants, so producing a larger response. More importantly, it also ensured that all participants reviewed exactly the same sets of predictions and supporting information, rather than perhaps missing a vital output. This reflects our interest in the use of the supporting information, and a desire that this not be confused with the technical process of using the software. Participants were given brief written guidance on the kinds of information produced by these QSAR platforms. This ensured that participants could recognize the significance of the various outputs, to make an informed review of the information provided. (The full output files are in supplementary files 1 (VEGA), 2 (EPISuite), and 3 (T.E.S.T.) at <http://www.altex-edition.org>. The exercise is available online for use by others in training at <http://www.orchestra-qsar.eu/webforms/231>.)

The platforms were presented to participants in the order: VEGA, EPISuite, T.E.S.T.. This was for two reasons. Firstly, for those toxicologists *experienced* in the use of QSAR models, the newer VEGA platform was potentially an attraction to take part in the review, while the other platforms were more familiar, especially EPISuite. Secondly, for those *less experienced* in the use of QSAR models, it was considered helpful to start with VEGA because the greater level of explanation, guidance, and supporting information would aid their participation. However, in the sections below we discuss the information provided by the platforms in the sequence EPISuite, T.E.S.T., VEGA, because this reflects the chronological order of their original development and the alphabetic order.

3.9 Our discussion of the participants' review

Our interest is in user recognition of reliability and uncertainty, and in the level of shared understanding and agreement in assessing reliability across experts from different institutions and sectors. As many stakeholders have argued in our interviews (Pardoe et al., 2010, 2011), cross-sector agreement on reliabil-

Tab. 1: Summary of predictions for Chemical 1

			<p>Chemical 1 was chosen to provide participant users with an example of high reliability, and to provide an important case study of recognising and documenting that reliability. The BCF models on all three platforms were capable of providing a reliable prediction and no critical issues were identified in the reports.</p>	
EPISuite	T.E.S.T.	VEGA		
Regression-based estimate: 2.83	Consensus 2.79	Hybrid model 3.13		
Arnot-Gobas estimate: 2.73	Hierarchical clustering 2.60	Model 1 3.03		
(BCF considering metabolism; arithmetic mean of the three trophic level values)	Single model 2.68	Model 2 3.05		
	Group contribution 2.85			
	FDA 2.86			
	Nearest neighbor 2.98			

ity is vital for QSAR predictions to be trusted and credible as a basis for investment and decision making. In EU regulatory toxicology, industry needs some confidence that their assessment of reliability in a particular case will be agreed to and accepted by regulators.

In the discussion of participants' comments, therefore, we focus on areas of disagreement and where conclusions are at odds with our understanding of the evidence. So for Chemical 1, we seek to understand the decision making by the minority of respondents who concluded the predictions were *unreliable*. For Chemical 2, we seek to understand the different reasoning of those who concluded that the prediction was reliable or unreliable. For Chemical 3 we focus on the decision making of the few who concluded that the prediction was actually reliable.

4 Results and discussion for Chemical 1: a case study in recognizing and documenting high reliability

4.1 The predictions and supporting information for Chemical 1

The BCF predictions for Chemical 1 from the three platforms are shown in Table 1. In this example of high reliability, the platforms also show a high level of agreement in their predictions. However, before looking for a potential consensus across the platforms, we first describe the information provided by each platform separately and the participants' review of this information.

EPISuite prediction for Chemical 1: highly reliable

Agreement between the EPISuite model results?

There is a very close agreement between the two EPISuite model results. The difference is less than the uncertainty range for experimental results of 0.6 (see 3.5).

Applicability domain check

The manual check of the applicability domain confirms that the compound is within the applicability domain for both models. The manual check involves verifying that the molecular weight and the logP values of each compound are within the range of values identified by the model developers.

Conclusion

For both of these reasons we can be confident in this prediction. (Given that the supporting information is limited, this conclusion inevitably relies on the established rigor and respected datasets of the EPISuite platform.)

T.E.S.T. prediction for Chemical 1: highly reliable

Agreement between the T.E.S.T. model results?

The T.E.S.T. platform offers a consensus model which integrates the results from the individual models; if one or more of the models have a poor performance for a compound then it is not included within the consensus model. For this compound, all the models are used for the consensus prediction, and so this supports the reliability of the result. In addition, there is a good agreement between the different T.E.S.T. models (with less difference than the uncertainty range for experimental results).

Applicability domain check

The automatic check of the applicability domain confirms that Chemical 1 is within the range.

Predictions and experimental values for similar compounds

T.E.S.T. shows the predictions alongside the actual experimental results for the most similar compounds in the external test set of the model (Fig. 1). For this compound, there are several compounds with good similarity, including the two compounds with CAS 13510-50-6 and 605-02-7. The similarity coefficient assigned to these two compounds is high. (The first has two ben-



zene rings, like the target compound, and has three aliphatic carbons, compared with two in the target compound, and so is quite similar. The second has a naphthalenic ring instead, so we would expect that compound to have a lower polarity, a higher logP value, and a higher BCF value.) It is therefore a sign of reliability that the prediction error for these two closely related compounds (i.e., the difference between the predicted and experimental values) is within the experimental error for BCF (see 3.5).

A similar analysis can be performed for the less closely related compounds in Figure 1. All these similar compounds seem to be a little bit more hydrophobic due to the presence of an additional aromatic ring, except for the first one, which has one more methyl group than the target compound and a different shape. Therefore, we would actually expect the BCF value for our target compound to be a little lower than the values for the similar compounds in Figure 1 (with the possible exception of the 3rd compound). At the same time, we draw confidence from the fact that the predicted values for the many similar compounds are close to the predicted value for the target compound.

Similar compounds present in the T.E.S.T. training set also can be seen by looking at the results from the nearest neighbor model (not shown here). Compounds with CAS 101-81-5 and 84-15-1 give confidence in the prediction in the same way as the first two similar compounds above. (Links to other T.E.S.T. models, that were not available to participants in the exercise, provide further information on model statistics. The FDA model also provides more structures for read-across.)

Conclusion

For all of these reasons, we can be confident in the T.E.S.T. prediction and conclude that it is reliable.

VEGA prediction for Chemical 1: highly reliable

Agreement between the VEGA model results?

There is a very close agreement between the predictions from the three VEGA models (and the difference is well within the range of experimental uncertainty).

Relation to regulatory thresholds

However, VEGA also indicates that the prediction is close to the REACH threshold for bioaccumulation of 3.3 log units, so it is important to confirm that reliability with further evidence from the supporting information.

Applicability domain check

The automatic applicability domain check, reported by a quantitative value of 1, confirms that Chemical 1 is within the range of the applicability domain (Fig. 2). (This AD check is carried out by an automated analysis of the most similar compounds, shown in Figure 3 below. “1” is the maximum value for the global index.)

Automated VEGA checks and warnings

VEGA also produces other automatic evaluations in order to identify potential reasons for concern (Fig. 2). In this case no reasons for concern are identified. The similarity index for similar molecules with known experimental values is considered good

when its value is higher than 0.7, as in this case. The average prediction error for similar molecules and the maximum prediction error are both low. The difference between the prediction and the experimental values of similar molecules is low.

In addition, the “Atom Centered Fragments similarity check” (ACF) reports no unusual fragments present in the molecule. This means there are no molecular fragments present that might cause the target chemical to have a different toxicity value from the related chemicals. The “Descriptors noise sensitivity analysis” reports that a 10% perturbation of model descriptors would not significantly affect the output value for this compound. This indicates that the model is stable for this prediction: the prediction is not vulnerable to small changes in the descriptors.

Predictions and experimental values for similar compounds

Rather than relying solely on these indices, it is intended that the user also manually compare the prediction with the experimental values and predictions for similar compounds in order to evaluate more directly and visually the similarity of the compounds. Like T.E.S.T., VEGA therefore enables the user to re-

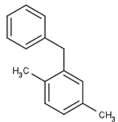
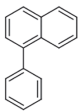
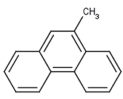
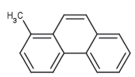
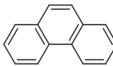
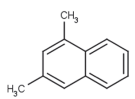
13540-50-6		0,90	3,32	2,69
605-02-7		0,89	3,51	2,74
883-20-5		0,70	2,56	2,77
832-69-9		0,70	3,18	2,73
85-01-8		0,69	3,31	2,72
575-41-7		0,66	3,37	2,75

Fig. 1: T.E.S.T.: a list of the most similar compounds in the test set of the model for Chemical 1

(The list includes all compounds with a similarity coefficient above 0.5; only the first six are reproduced here.)

view the most similar compounds considered by the model so that the user can evaluate the similarity between the compounds more directly and visually. This process of reviewing the experimental and predicted values, and the reasoning involved, is a kind of read-across evaluation.

The six compounds most similar to Chemical 1 within the training and the test set are reported (Fig. 3). We can analyze these compounds as follows:

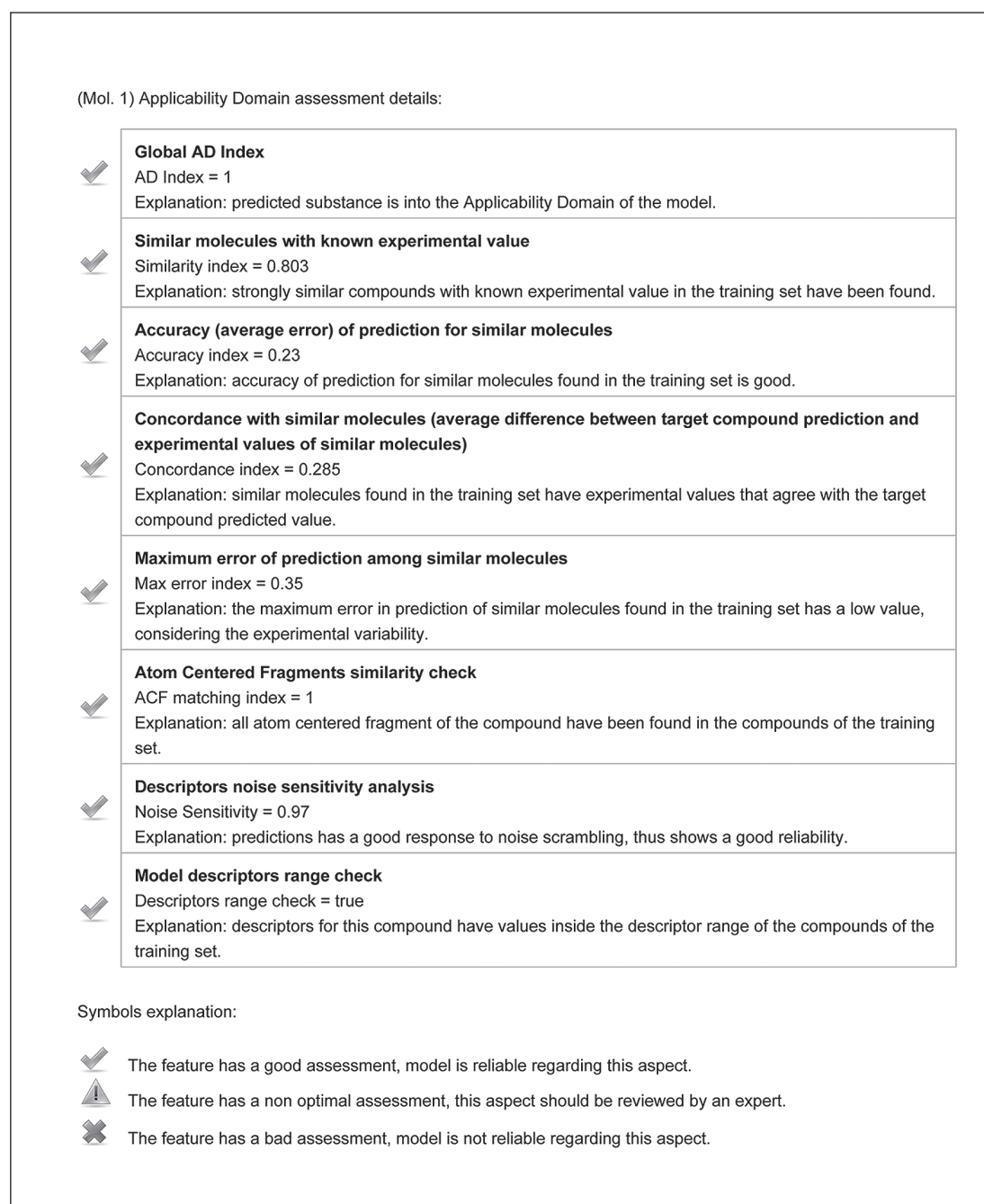
The 1st molecule, considered by VEGA to be the most similar to Chemical 1, has one more aromatic ring than our target compound, so it should be more hydrophobic. In view of the direct

correlation between logP (which takes into account hydrophobicity) and BCF, we would expect the BCF value for our target compound to be lower than the experimental value of the first compound, i.e., lower than 3.51.

The 2nd molecule displayed by VEGA can be considered quite similar to the target molecule, so we would expect the BCF value for our target compound to be similar to this experimental value of 2.94.

The 3rd molecule listed has only one benzene ring, so we would expect it to be less hydrophobic than our target compound. Its BCF experimental value of 2.68 should therefore be

Fig. 2:
Evaluation of the applicability domain within VEGA, and its components, for Chemical 1 (The indices and values are explained in supplementary file 4.)



lower than that of the target compound. A similar conclusion can be reached for the 6th molecule.

The 4th molecule has a hydroxyl group and thus it will be appreciably more hydrophilic than our target compound. Its BCF value of 1.59 should be significantly lower than that of the target compound.

The 5th molecule, which contains three tert-butyl groups, should be much more hydrophobic. Its BCF value is very high at 4.37.

The graph in Figure 4, provided by VEGA, shows the correlation between the experimental logBCF values and the calculated logP values for all compounds used within the VEGA BCF QSAR models. (Experimental logP values are not available for all of these compounds.) In the case of BCF, logP is one of the most important descriptors used in the literature, and it is mechanistically associated with the phenomenon observed: if the chemical substance has a lower logP value, it will prefer to stay in water, and vice versa. However, chemicals with a high logP value may not always obey this rule in case of reduced bioavailability, metabolism, or other phenomena. So while logP is an important descriptor within the VEGA/CAESAR model, it is

only one of several. As the points below the “cloud” of values in Figure 4 demonstrate, it may be that while the logP value is high the logBCF value is lower than expected. In such cases, where the target compound lies outside the “cloud” of highly correlated values, the expert should carefully evaluate the prediction. To guide this further evaluation, VEGA reports some “mitigating” fragments, i.e., molecular fragments that have been associated with BCF values lower than expected. VEGA also shows the compounds that are most similar to the target compound and have this “mitigating” fragment.

In this case, the large dark dot shows that our target compound (Chemical 1) lies within the typical behavior of most of these compounds. This provides confirmation that there are no potentially critical issues.

VEGA also offers a closer look at this correlation, as shown in Figure 5. Again, the large dark dot represents the predicted value for the target compound. It shows the same correlation of logBCF against MlogP values, but for only the three most similar compounds. The open circles represent the *experimental* logBCF values: the size of the circle is proportional to the similarity index and therefore shows that these three compounds are similar. The small black squares represent the *predicted* logBCF values. The vertical bar indicates the error between the predicted and experimental logBCF value for the three similar compounds and shows that these errors are within the experimental uncertainty for BCF of 0.6 (see 3.5).

In this way, Figure 5 enables the user to visualize the discussion we offered for Figure 3. It also illustrates the importance of molecular fragments as well as logP in the prediction. This further contributes to the mechanistic interpretation of the BCF

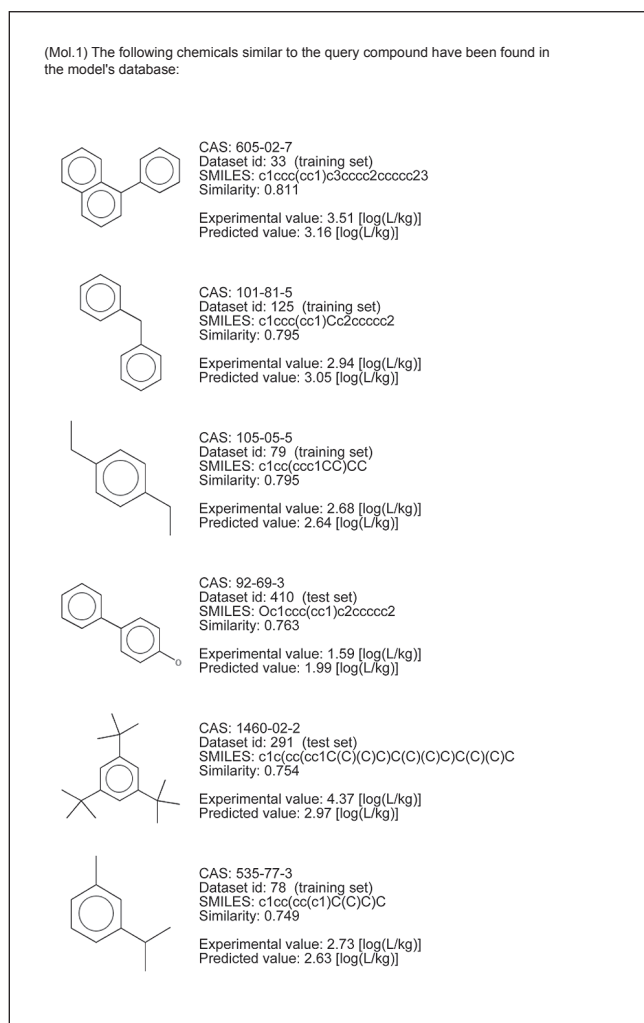


Fig. 3: VEGA: the six most similar compounds for Chemical 1 in the training and test set

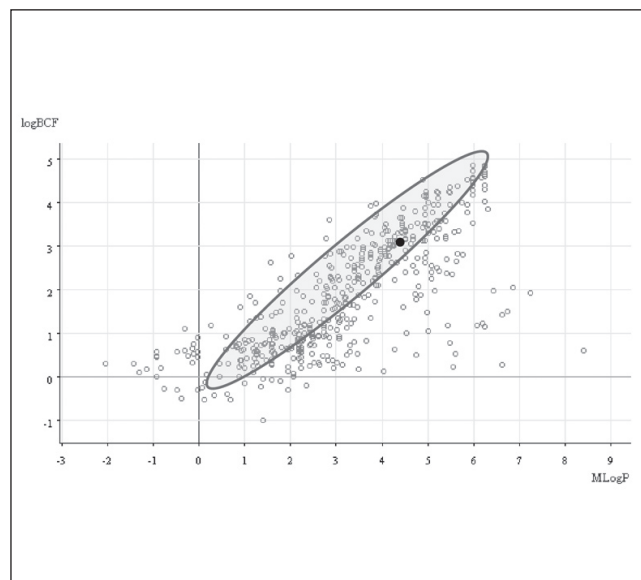


Fig. 4: The correlation between experimental log BCF and calculated MlogP

The target compound (Chemical 1) is shown in black. The ellipse circumscribes the plot area in which a linear relationship between BCF and MlogP exists.

value offered by VEGA. Referring to both Figure 3 and Figure 5, the compound shown on the right of Figure 5 is the first compound listed in Figure 3. This compound has *experimental* logBCF values higher than those for the target compound because it has three benzene rings. The third compound of Figure 3 is on the left in Figure 5. It has an *experimental* logBCF lower than that of the target compound, because it has one single benzene ring. The compound in the middle of Figure 5 is the second compound in Figure 3, and its experimental (and predicted) logBCF value is very close to that of the target compound, due to the similarity between the target compound and this compound.

(Note: Figure 5 is used mainly to evaluate the *experimental* values of the related chemicals (open circles). But it also allows the user to review whether there is a consistent error in the predicted values (over-prediction or under-prediction) for the related chemicals and, if so, to take this into account in judging the prediction for the target chemical. However, it is always important to recognize the typical uncertainty for experimental results; in this case errors lower than 0.6 log units can reasonably be viewed as not important, as is the case for the chemical on the right.)

In the lists and graphical representations of similar compounds, the user can vary the order of similarity on the basis of their expert judgment about the relative importance of the descriptors, the molecular fragments, and other molecular or physical-chemical properties in a particular case. This flexibility is an integral part of providing the user with a tool for evaluation rather than just a result. It enables the user to control and comment on the material given by the software and therefore to take responsibility for evaluating whether the predicted value is sufficiently reliable for the purpose.

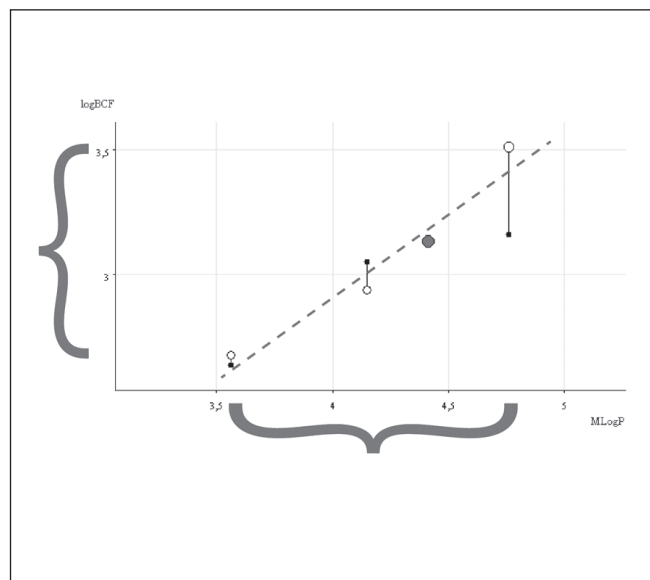


Fig. 5: The logBCF and MLogP values of the three compounds most similar to the target compound for Chemical 1

Target compound: large dot. Experimental values: white circles. Predicted values: black squares. Note the small range of both the MLogP values (x axis) and logBCF values (y axis).

The situation in Figure 5 is ideal for a prediction, because log BCF and MlogP values for the target compound lie within the range of experimental log BCF values and predicted MlogP values for the most similar compounds (shown by the brackets). It is also clear that there is a direct correlation between MlogP and logBCF (shown by the dashed line). Moreover, the predicted value for the target compound follows the linear relationship revealed by the experimental results.

Conclusion

By considering all of these complementary sources of information, we can be confident in this prediction.

4.2 Participant decisions on the reliability of the predictions for Chemical 1

Participant decisions on the reliability of the predictions for all three example chemicals (Chemicals 1, 2 and 3) are shown in Figure 6.

It is necessary to start the analysis for Chemical 1 with a caveat. While Chemical 1 requires a prediction from the VEGA models (and while both T.E.S.T. and EPISuite produce predictions) an experimental result for this compound was in fact used in the training sets for the development of the T.E.S.T. and EPISuite models. Given the transparency of the models, this was observed by some participants, so we should not attach significance to the level of *agreement* on the reliability for this compound. For this reason, the columns for Chemical 1 in Figure 6 are shown in outline only.

With this caveat in mind, the result is nevertheless extremely encouraging for VEGA. It was the first platform to be used

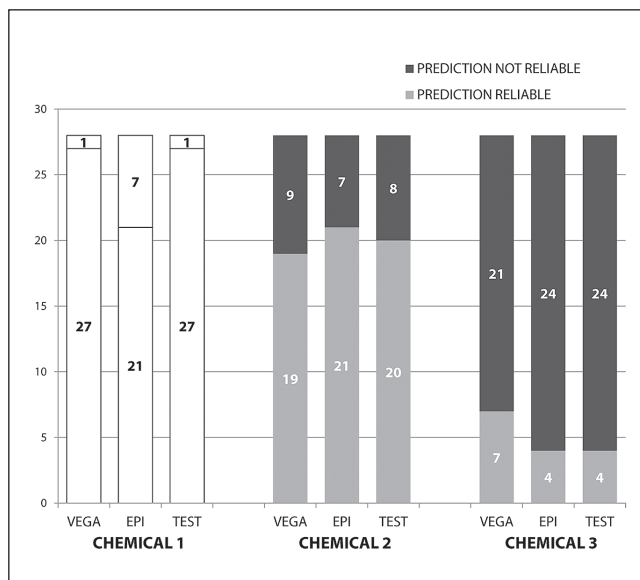


Fig. 6: Numerical results of expert judgments on the reliability of the predictions for three compounds across three QSAR platforms: VEGA, EPISuite, and T.E.S.T.

Expert judgments for Chemical 1 are in outline only because EPISuite and T.E.S.T. revealed an experimental value that was seen by some participants (see 4.2).



within the review exercise, so if indeed the participants used it first (and did not subsequently change their review after gaining confidence from seeing the experimental result in T.E.S.T.) then it would suggest that all but one of the 28 participants believed that VEGA provided the information necessary for them to be confident that the prediction was reliable.

The caveat also does not reduce the central insight from Chemical 1: it was chosen (i) to provide participant users with an example of what very high reliability looks like, and (ii) to provide an important case study of recognizing and documenting a prediction with very high reliability. It enables us to investigate whether the information provided by the QSAR models is sufficient to convince the wisely cautious toxicologist of that reliability, and, if not, what further information is needed.

The near unanimity on the reliability of the predictions from VEGA and T.E.S.T. is shown in Figure 6. In each case, 27 participants concluded that the prediction was reliable, and only one participant concluded that it was not. This is consistent with our analysis of the information above. When using each platform in turn, participants noted evidence for the reliability of the Chemical 1 prediction with similar reasoning to our own (discussion above). They noted that the compound is within the applicability domain, that there is a similarity of predictions from the different models within each platform, and that the compound is similar to those for which experimental results are available.

The most interesting result is that a full 7 of the 28 participants (25%) did not have sufficient confidence in the EPISuite prediction to conclude that it was reliable. This raises the question of whether this conclusion was (i) prompted by perceived evidence of uncertainty from EPISuite, or (ii) reached as a matter of wise caution from a lack of supporting information to confirm the reliability. In fact, the optional comments from participants all cite the latter reason. It seems there was insufficient supporting information from the EPISuite output to give them confidence in the reliability of the prediction, so they rightly erred on the side of caution and concluded it was not reliable. Their explanations can be summarized around three themes:

- *The EPISuite output contains fewer explanations than the other platforms and is harder to understand without previous experience in the use of the software.*
- *The definition of the applicability domain is based solely on Kow and MW.*
- *The lack of provision for read-across from experimental results and predictions for similar compounds make it difficult to judge the reliability for the target compound.*

This is a potentially significant finding to inform the further development of models. It would appear that even for such a highly reliable prediction, the EPISuite models did not provide the information needed to confirm that reliability for a significant number of users. Specifically, they needed: (i) fuller explanations when the platform was unfamiliar; (ii) a recognizably rigorous basis for the applicability domain, and (iii) an opportunity to judge the reliability from read-across.

While drawing those lessons for the development of QSAR platforms, it is also useful to respond to the critique in this case. Firstly, EPISuite is intended for use by those trained and expe-

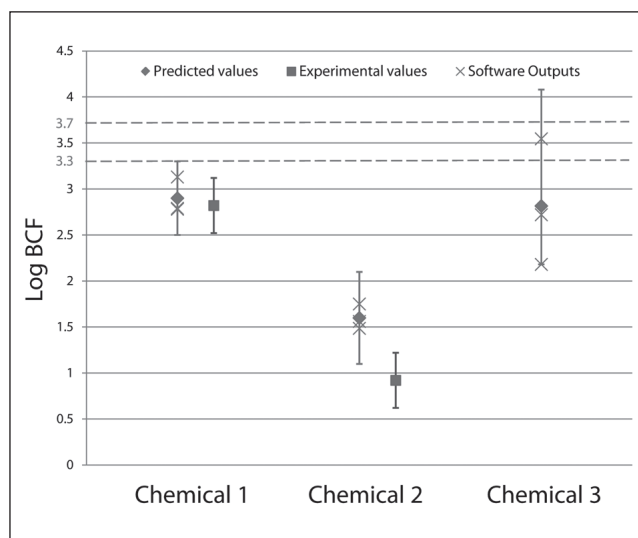


Fig. 7: A comparison between the individual QSAR platform outputs and also between the predicted and experimental values for Chemicals 1, 2, and 3

For the predicted values, we report our “weight of evidence”-based uncertainty. For the experimental values we add the experimental uncertainty.

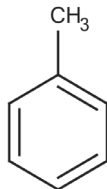
rienced in its use, as is the case with all QSAR platforms (and most computer software); it can seem inaccessible to others. As the earliest of the three platforms, it includes less explanatory information, so this can be an obstacle for those not already familiar with it. Secondly, EPISuite was not intended to enable read-across. In recent years, as more toxicologists have used QSAR models, it has become evident that read-across remains an important part of the decision-making process. It offers a way to visually review the evidence on which the prediction is based, to verify it, and therefore to feel confident about the basis for the QSAR prediction. These participant decisions and comments appear to confirm this.

(The guidance provided for the exercise can be accessed at: http://www.orchestra-qsar.eu/sites/default/files/Episuite_3compounds_exercise1.pdf. More details on the EPISuite BCF models are published: (Meylan et al., 1999; Arnot and Gobas, 2003).

4.3 A cross-platform consensus for Chemical 1?

After reviewing the reliability of outputs from a particular platform, it is recommended practice to look for potential consensus across platforms. (This was not requested from participants.) As Table 1 shows, all three platforms provide strong agreement on the BCF prediction. The maximum variation across all the outputs from the three platforms (between highest prediction of 3.13 from VEGA and lowest prediction of 2.60 from T.E.S.T.) is 0.53 log units. The deviation of these figures from the T.E.S.T. consensus figure of 2.79 is a maximum of 0.34. This is less than the uncertainty range for experimental results of between 0.40 and 0.75 log units. A review across the three platforms therefore further supports the view that there is a high level of reliability.

Tab. 2: Summary of predictions for Chemical 2

	<p>Chemical 2 was chosen to offer participants a more challenging and typical example of the use of QSAR models, where different elements of the supporting information suggest unreliability and reliability, so the user has to analyze the material using toxicological understanding.</p> <p>In evaluating Chemical 2 for BCF, we would initially be concerned about some indicators of reliability, but after consideration (below) would be confident with the predictions.</p>				
EPISuite	T.E.S.T.		VEGA		
Regression-based estimate:	1.47	Consensus	1.75	Hybrid model	1.56
Arnot-Gobas estimate:	1.50	Hierarchical clustering	1.56	Model 1	1.52
(BCF considering metabolism;		Single model	1.94	Model 2	1.60
arithmetic mean of the three		Group contribution	1.69		
trophic level values)		FDA	1.25		
		Nearest neighbor	2.33		

However, for an agreement across platforms to be significant, rather than merely offering duplication, we need to know that the platforms use different approaches, including different algorithms and experimental data. For BCF, the models used by these three platforms are different and independent, since they use different algorithms and they are based on partially different training sets. While it could be argued that they are not completely independent because they all typically use logP, the logP values are calculated in different ways on each platform, and other descriptors are also used. The three platforms therefore provide complementary evidence.

Based on the predictions, we could conclude that the logBCF value for Chemical 1 is likely to be between 2.50 and 3.30 (i.e., 2.90 ± 0.4). (The figure of 2.90 is the arithmetic mean of the EPISuite Meylan and Arnot-Gobas model predictions, the VEGA (CAESAR) hybrid model predictions, and the T.E.S.T. consensus model predictions. The potential error of 0.4 log units is a reasonable and realistic value, defined by the human expert on the basis of a weight of evidence approach, which takes into account the potential uncertainty within the models, referring to this single molecule.) Since the value of 3.3 log units is the threshold for whether substances are classified as bioaccumulative or not, there is a low probability that this compound is bioaccumulative. However, it is clear that the compound should not be classified as “very bioaccumulative” (vB), which has a higher threshold (3.7 log units). (As a matter of interest, the available experimental logBCF value for Chemical 1 is 2.82 and should be regarded as having an uncertainty of ± 0.3 .)

To provide a summary, Figure 7 shows the experimental values and the predicted values (individual QSAR platform outputs) with the potential margin of error, for Chemical 1 (and for Chemicals 2 and 3). It shows the similarity in the predictions, the similarity with the experimental value, and the proximity to the threshold.

5 Results and discussion for Chemical 2: a case study in analyzing and documenting complex reliability

5.1 The predictions and supporting information for Chemical 2

The logBCF predictions for Chemical 2 (toluene) from the three platforms are shown in Table 2. The predictions clearly provide some agreement but at a strikingly lower level than for Chemical 1. Before looking for a potential consensus across the platforms, we describe the information provided by each platform and the participants’ review of this information.

EPISuite prediction for Chemical 2: reliable

Agreement between the EPISuite model results?

As Table 2 shows, there is strong agreement between the results from the two EPISuite models.

Applicability domain check

A manual check of the applicability domain (see 4.1) confirms that Chemical 2 is within the range of both models.

Relation to experimental data

The logP value used in making the logBCF prediction is actually an experimental result, so this will improve the reliability of the prediction.

Relation to regulatory thresholds

The consistent predictions are also noticeably lower than the REACH BCF threshold of 3.3 log units, which suggests we can be confident that it is not bioaccumulative in terms of the REACH regulations.

Conclusion

For all these reasons we are fairly confident with this prediction



Fig. 8: Prediction and warnings of the VEGA output for Chemical 2

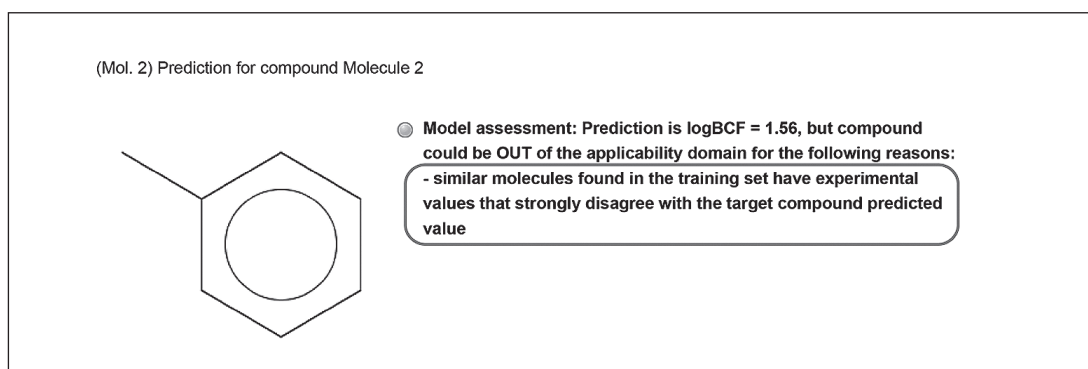
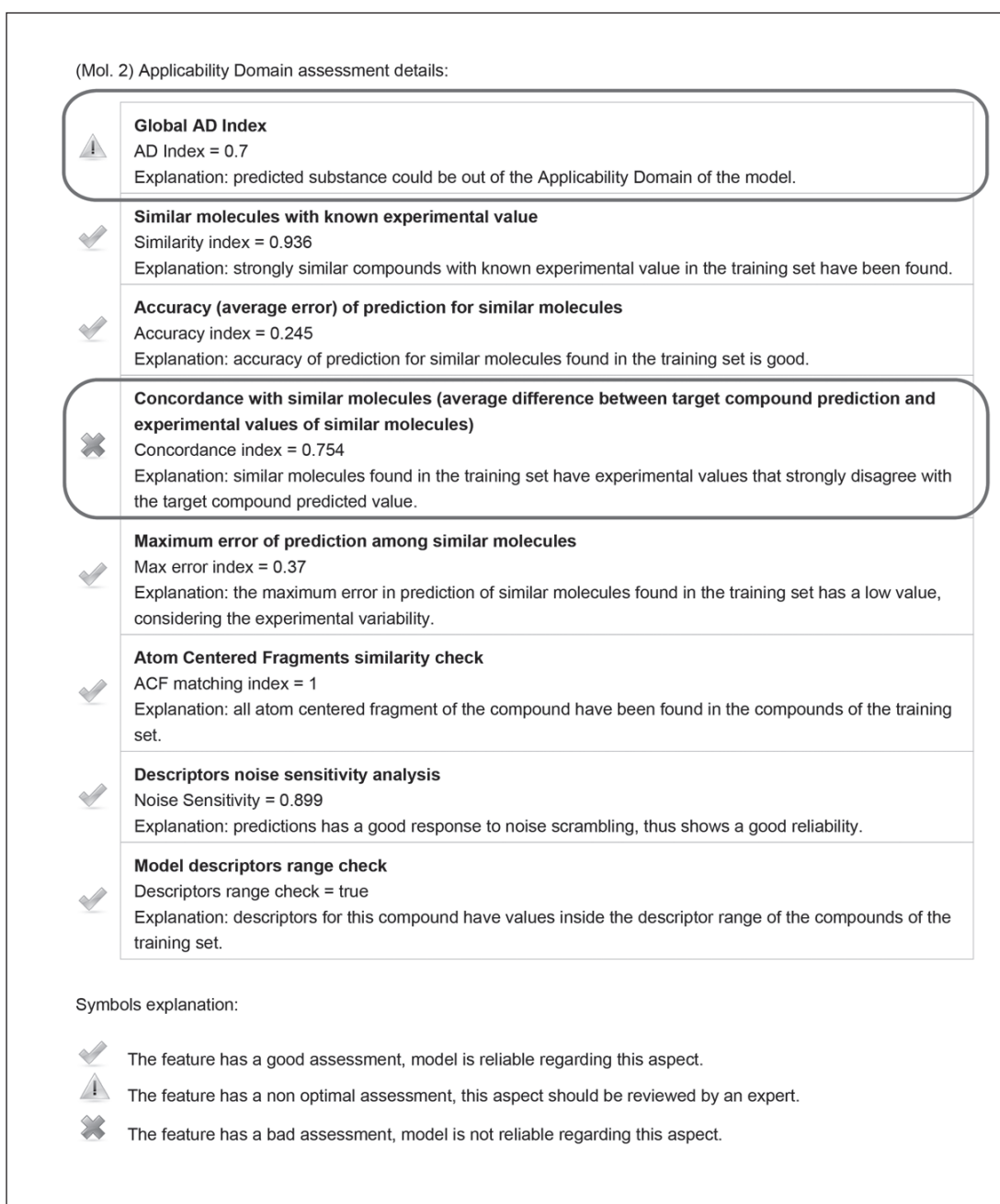


Fig. 9: The applicability domain index (ADI) report from VEGA for Chemical 2



but nevertheless would want to see confirmation and more detail from other platforms.

T.E.S.T. prediction for Chemical 2: reliable

Applicability domain check

The automatic applicability domain check confirms that Chemical 2 is within the applicability domain for the T.E.S.T. BCF models.

Agreement between the T.E.S.T. model results?

There is a fairly good agreement between the results of the different T.E.S.T. models. The exceptions, which are cause for concern, are the FDA and nearest neighbor results. While the maximum deviation from the T.E.S.T. consensus value is 0.58 and therefore within the upper uncertainty range for experimental results (0.75 log units; see 3.6), the spread between these results is more than 1 log unit, and such a divergence suggests some uncertainty in the prediction. Confidence, therefore, needs to come from closer investigation.

Relation to regulatory thresholds

If the user simply requires a BCF value in relation to the regulatory bioaccumulative threshold of 3.3 log units, then the most important observation is that all the predicted values for this compound are well below this threshold. Even with such a range of figures, we can state with some confidence that it is not bioaccumulative.

However, there is clearly a higher level of uncertainty about the BCF value compared with the previous case. So if the user wants to better analyze the evaluation and its reliability, it is necessary to look more closely at the model and the results, especially to understand the nearest neighbor and FDA results.

Predictions and experimental values for similar compounds

The nearest neighbor model has produced the much higher BCF prediction of 2.33. However, this higher result can be understood if we look at the compounds it relies on. “Nearest neighbor” is a kind of “local” model that takes into consideration the experimental BCF values of only a small number of the most similar compounds; it can be considered as a kind of small and automated read-across. For this compound, we can see that the three most similar compounds found by the program are all trimethylbenzenes, with experimental logBCF values of 2.32, 2.42, and 2.26. By referring to these compounds, the model consequently assigns a predicted value of 2.33 to the target compound. Yet if we reason from a chemical point of view, we know that the polarity of these three trimethylbenzenes is lower than that of toluene, owing to the higher number of methyls. Hence we can recognize that a prediction based on the nearest neighbor analysis will be an overestimate.

The analysis within the FDA model is similarly “local”; it is based on a smaller set of experimental results than the other predictions. For this compound, it is significant that the other four T.E.S.T. predictions are more consistent, and so for both reasons the expert can decide that these other predictions are potentially more reliable.

Conclusion

In summary, there is good agreement across the four T.E.S.T. predictions, and all four predictions are significantly below the regulatory threshold, so we can be confident that Chemical 2 is not bioaccumulative. At the same time, having recognized that it is less reliable than the prediction for the previous compound, the evaluation requires some further evidence and confirmation from different and independent platforms.

VEGA prediction for Chemical 2: reliable

Agreement between the VEGA model results?

As shown in the prediction values (Tab. 2) above, there is good agreement between the predictions from the three VEGA models. The difference between the values is within the experimental uncertainty for BCF of 0.6 (see 3.5).

Applicability domain check

However, VEGA’s automatic applicability domain check warns that “similar models in the training set have experimental values that strongly disagree with the target compound predicted value” (Fig. 8).

The global applicability domain index (ADI) is given as 0.7, with a warning that “the predicted substance could be out of the applicability domain of the model” (Fig. 9). Moreover, the concordance index is given as 0.754, with the same warning as in Figure 8. As with T.E.S.T., this indicates a potentially significant difference between the target chemical prediction and the experimental values of similar molecules.

Automated VEGA checks and warnings

VEGA automatically shows the safety margin associated with the prediction, as reported in Figure 10. This safety margin is dependent on the ADI as well as on the regulatory threshold. Its values have been calculated during the evaluation of the CAESAR model to avoid the presence of false negatives in the training set. This application of a safety margin to BCF predictions of new chemicals is intended as a precautionary way to reduce misclassification errors. Hence it is called a “conservative confidence interval” (details in supplementary file 4 at <http://www.altex-edition.org>).

By issuing these warnings, VEGA identifies the need for more detailed investigation by the expert to identify the significance of the results on these indices in terms of the reliability of the predictions.

Predictions and experimental values for similar compounds

From the list of similar compounds in the training set (not reproduced here), the user can observe that the compounds identified as most similar have 2 or 3 more methyl groups linked to the aromatic ring than our target compound. It would be expected, therefore, and highly reasonable, that their BCF values will be higher, exactly as the concordance index has warned.

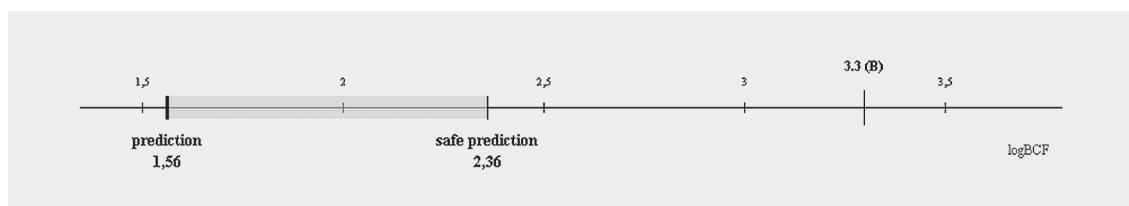
As for Chemical 1, this situation also is depicted in a zoomed scatter plot, Figure 11. The scatter plot suggests a clear linear relationship, where for every additional methyl we can anticipate a regular increase of both logP and BCF values.



Threshold 3.3 (bioaccumulative)

Following, a chart showing the predicted value together with its conservative confidence interval for safe classification. For the threshold $\log BCF = 3.3$, the current compound can be associated (due to its Applicability Domain index value) to a conservative interval of 0.8 log units.

On this basis, the compound can be safely classified as not bioaccumulative.



Threshold 3.7 (very bioaccumulative)

Following, a chart showing the predicted value together with its conservative confidence interval for safe classification. For the threshold $\log BCF = 3.7$, the current compound can be associated (due to its Applicability Domain index value) to a conservative interval of 0.9 log units.

On this basis, the compound can be safely classified as not very bioaccumulative.

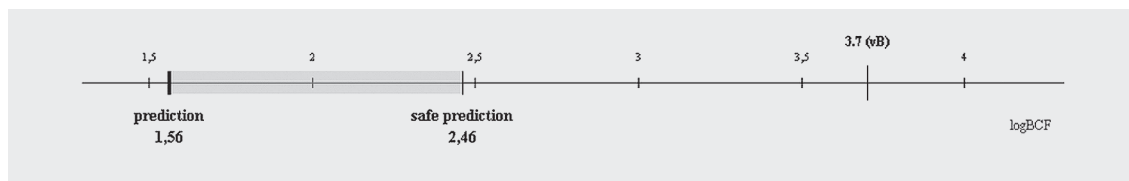


Fig. 10: The safety margin (or conservative confidence interval) for the B (upper) and vB (lower) classes for Chemical 2 (from VEGA)

However, in contrast to the situation for Chemical 1, we are not able to interpolate the predicted value *within* the range of experimental values. So, while we can understand and explain the warnings provided by VEGA, uncertainty about the actual BCF value remains. We would want to refer to the other platforms for further confidence in the prediction. Nevertheless, it is vital to observe that the compounds for which we have experimental data represent a kind of worst case, with higher BCF values. It is therefore safe to conclude that the BCF for Chemical 2 will be lower than these experimental results.

Conclusion

From this more detailed analysis of the information outputs, we can be confident in the prediction that Chemical 2 has a BCF value below the similar chemicals in Figure 11, and therefore clearly below the regulatory threshold of 3.3 log units. We can conclude, therefore, that it should be classed as not bioaccumulative.

5.2 Participant decisions on the reliability of the predictions for Chemical 2

Across the platforms, a fairly consistent number of participants concluded that the predictions for Chemical 2 were reliable or not reliable: 19-21 participants (approx. 70-75%) concluded that the predictions were reliable and 7-8 participants (approx. 25-30%) concluded that they were not reliable. The platforms seem to have been reviewed independently, with 8 of the 28 participants (25%) drawing a different conclusion for the different platforms.

The lack of unanimity across participants prompts the question of what these opposing decisions were based on, and the extent to which they represent a divergence of thinking by the experts.

This binary “yes” or “no” choice was requested in order to force a decision on whether the prediction is “reliable” or not. Yet it is important not to merely assume disagreement: demanding a yes/no response in any survey can potentially conceal degrees of agreement and shared understanding. In most real

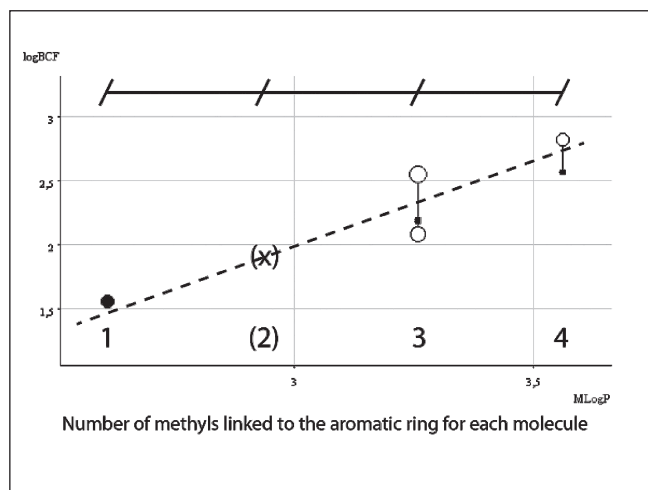


Fig. 11: The scatter plot provided by VEGA shows three experimental values (open circles) and the predicted value for the target compound (Chemical 2, dark dot)

(Note that there are two compounds with 3 methyls, with the same MlogP value, but with slightly different predicted logBCF values. The (x) is the likely position of a compound with two methyls.)

evaluations with material consequences the user will review a decision with colleagues, so in this study, the commonality or divergence of understanding across users is potentially more interesting than the mere yes/no decisions. The relevant question is whether the expert judgments on Chemical 2 reflect a divergence of thinking, as the numbers of yes/no responses might suggest, or whether there is nevertheless a common understanding and basis for informed discussion between these users.

For brevity, we will discuss the additional comments made on the use of one platform. Given the apparently high confidence generated by VEGA and T.E.S.T. for Chemical 1, VEGA provides the largest shift for Chemical 2, with a split of 19:9 on the decision. It is therefore the most interesting focus for understanding the split decision. Of the 28 respondents, 20 provided comments on using VEGA for Chemical 2: 14 of the 19 who concluded it was reliable, and 5 of the 9 who concluded it was not reliable.

5.2.1 Participants' grounds for deciding that the VEGA prediction is not reliable

Respondents who decided that the VEGA prediction is *not reliable* indicated that they based this decision on the information provided by VEGA, rather than on a lack of information. They cited the issues we discussed above: the weak concordance index, the explicit warnings given, and the apparent difference between Chemical 2 and the compounds in the training set:

- weak concordance index,...
- Model assessment: Prediction is $\log BCF = 1.56$, but compound could be OUT of the applicability domain for the following reasons: similar molecules found in the training set have experimental values that strongly disagree with the target compound predicted value; Experimental values of simi-

lar molecules in the training set are considerably higher by nearly one log unit than the predicted value of the target compound [echoed by two other participants]

Evidently they responded to this information by erring on the side of caution and concluded that the prediction was not reliable. In our own decision-making process (above), these issues were a cause for concern, but they were not decisive. They were overridden by a more detailed review that suggested (i) a good linear relationship, (ii) a reason why the experimental values would be higher, and (iii) a recognition that all predictions were well within the REACH BCF threshold.

5.2.2 Participants' grounds for deciding that the VEGA prediction is reliable

Participants who concluded that the results for Chemical 2 are reliable offer comments and justifications that contrast to the straightforward comments they made for Chemical 1. For Chemical 2, participants show recognition of the issues that could reduce the reliability and which should cause initial concern, and then they include explicit justifications for their “yes” response.

The following are some examples, grouped under three themes:

- (i) *Comments on the process of weighing different information:*

#2: *Despite failure to comply with one of the domain assessment criteria, the prediction seems robust, especially when considered that the experimental data for this compound varies from 31 to 1000.*

#13: *The majority of the indicators were met. The “concordance with similar molecules” statistic was not met – this doesn’t necessarily mean the prediction is bad if it doesn’t agree with a nearest neighbors – the neighbors could be different enough to yield a different tox value.*

#20: *Reasons for considering the prediction reliable: - Descriptors for this compound have values inside the descriptor range of the compounds of the training set. - LogP value is in the range of the compounds of the training set. - As expected, the log BCF predicted for target compound, which has a logP value=2.61, is lower than the experimental and predicted logBCF of the most similar compounds (characterized by higher logP values). This is also clearly shown in the plots MlogP vs. logBCF.*

- (ii) *Relation to experimental data for similar compounds:*

#4: *Even though VEGA found critical behavior, the prediction is anyhow reliable, on the basis of the information given by the analysis of the similar compounds.*

#18: *The compound is slightly out of the applicability domain for a bad concordance with the two top similar compounds. However, read across assessment as well as the scatter plot of MlogP against response values of the top-three most similar compounds would indicate that prediction is still reliable.*

#16: *Both of the [VEGA] prediction models get similar results. However, the substance is quite different compared to the most similar ones. But the similar ones are much more*



hydrophobic, which is in agreement with the lower predicted BCF of the test substance. I have compared the experimental value of benzene (BCF 24; log 1.3) and am convinced that the predicted value of the test substance is reliable.

#8: Even if the compound is at the limit of the AD, the model calculation for similar compounds is convincing and the value is in agreement with the expected effect of the structure variation.

#28: Yes [it is reliable] because it is a simple structure, the substance is in the logKow domain and the training set is large. However some restrictions exist: - similar compounds are not very similar - extrapolation of BCF with respect to most similar compounds instead of interpolation. If results would strongly influence regulatory decision, additional tests might be needed.

(iii) *Relation to the regulatory threshold:*

#6: Sufficiently reliable as part of weight of evidence approach since both predicted value and experimental results of similar compounds would not lead to very different conclusions for the purpose of classification, PBT and risk assessment.

5.2.3 A divergence of thinking or a basis for informed discussion?

In this way the comments from those who answered “no” and “yes” to reliability reveal a potentially greater level of agreement on the issues across the 28 participants than is evident simply from the yes/no responses. The comments suggest that there is a shared basis for discussion and for potential agreement between experts even in cases (like Chemical 2) where the decision is not self-evident and instead depends on weighing up potentially contrary evidence.

As an outcome for this review exercise, this finding is potentially more important than simply finding agreement in the yes/no decision. The comments suggest that the QSAR models offer a basis for common understanding, discussion, and a considered judgment across users, even when their initial conclusions may differ.

5.3 A cross-platform consensus for Chemical 2?

When predictions have potentially low (or even slightly questionable) reliability, it becomes even more important to adopt a consensus approach of looking at predictions across the platforms.

The BCF predictions for Chemical 2 from the three platforms (Tab. 2) clearly provide some agreement but at a significantly lower level than for Chemical 1. The T.E.S.T. consensus prediction (1.79) is in line with VEGA and EPISuite predictions of 1.56 and 1.47, respectively. (The FDA and nearest neighbor predictions are discussed in 5.1 above.)

For Chemical 2, the low potential to bioaccumulate identified by the T.E.S.T. consensus model is confirmed by the “safety margin” assigned in VEGA (see Fig. 10) and by the low predicted value of 1.76 log units provided by the Arnot-Gobas model in EPISuite (even when assuming the absence of metabolism). Given that the EPISuite, T.E.S.T., and VEGA programs use different and independent approaches, it is significant that they

provide a good level of agreement for this prediction. Moreover, in each case, the documentation produced by the models further supports the prediction.

In this way, the consensus prediction for Chemical 2 offers an interesting contrast with that for Chemical 1. Using all three platforms, we are able to observe a similarity across the key results and so conclude that the prediction is reliable. However, there is also a clear need to elaborate and qualify that judgment in terms of its function. If the function of our evaluation is to predict an exact BCF value, then there is a clear level of uncertainty across the key predictions for Chemical 2. For this reason we should assign a higher figure to the uncertainty for Chemical 2, compared with Chemical 1. However, if our goal is to classify the compound in relation to regulatory thresholds, then the distance of all of the predictions for Chemical 2 from the REACH threshold of 3.3 for bioaccumulation means that we can be confident in classifying the chemical as non bioaccumulative. By comparison, the more reliable predictions for Chemical 1 were close to the regulatory threshold, and so indicated a low probability that it is bioaccumulative (4.3 above).

The BCF value for Chemical 2 is most likely to be 1.60 ± 0.5 . (0.5 is the standard deviation of the results from the three platforms, and it gives a higher uncertainty value, than for Chemical 1.) The compound, therefore, should not be classified as bioaccumulative. (For interest, there is an experimental log BCF value for this compound of 0.92 provided by EPISuite, which should be read as 0.92 ± 0.3 . A few reviewers may have rated the predictions for Chemical 2 as being more reliable since the experimental value matched the predicted value, but we think that this information was not used in most of the cases.) Figure 7 summarizes the predicted and experimental values for this chemical.

6 Results and discussion for Chemical 3: a case study in recognizing and documenting uncertainty

6.1 The predictions and supporting information for Chemical 3

The LogBCF predictions for Chemical 3 from the three platforms are shown in Table 3. They show a low level of agreement, both within and across the three platforms. Before looking across the platforms, we describe the information provided by each platform and the participants’ review of this information.

EPISuite prediction for Chemical 3: unreliable

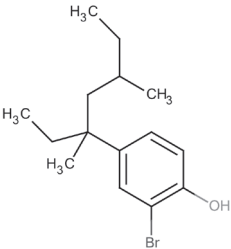
Agreement between the EPISuite model results?

There is more than 1 log unit difference between the results of the EPISuite models, and that is clearly outside the range of the experimental uncertainty.

Applicability domain check

A manual applicability domain check (see 4.1) confirms that Chemical 3 is within the range of applicability domain for both the models.

Tab. 3: Summary of predictions for Chemical 3

	<p>Chemical 3 was chosen because it has a more difficult molecular structure in terms of BCF prediction.</p> <p>All three platforms revealed difficulty in predicting consistent values. We would judge the prediction for this compound to be unreliable.</p>				
EPISuite	T.E.S.T.		VEGA		
Regression-based estimate:	4.08	Consensus	2.18	Hybrid model	2.72
Arnot-Gobas estimate:	3.01	Hierarchical clustering	2.77	Model 1	2.76
(BCF considering metabolism;		Single model	2.26	Model 2	2.53
arithmetic mean of the three		Group contribution	2.82		
trophic level values)		FDA	0.75		
		Nearest neighbor	2.31		

Relation to regulatory thresholds

One of the values exceeds the REACH BCF thresholds of 3.3 and 3.7 log units and so may define the compound as very bioaccumulative.

Conclusion

Given the difference between the predictions from the two EPISuite models and the level of uncertainty this demonstrates, we cannot be confident in this prediction. Moreover, both the region of uncertainty and the extent of the difference between the figures are critical in terms of bioaccumulative classification.

T.E.S.T. prediction for Chemical 3: **unreliable**

Agreement between the T.E.S.T. model results?

There is some degree of agreement between the results for the different models, with the significant exception of the FDA model prediction, which is more than 2 log units below the highest value. The FDA model calculates a BCF value of 0.75, which may be because (as discussed for Chemical 2) this model is more “local” compared to others. However, this very different value remains a warning of a possible uncertainty that requires analysis of the experimental data for similar compounds.

Predictions and experimental values for similar compounds

From observing the list of similar compounds (see supplementary file 3 at <http://www.altex-edition.org>), the user can see that the most similar compounds are actually very different from the target compound. The experimental values for the three most similar test set compounds range widely from 1.39-3.59. Looking at the wider range of similar chemicals in the nearest neighbor method, it is also evident that there are no really similar compounds in the training set. For example, none contains a bromo group attached to an aromatic ring.

Conclusion

There is high variability in the values predicted by the T.E.S.T. models, and there is also a lack of sufficient information from the similar compounds to differentiate between the models and so identify which model may be reliable. For these reasons we are not confident in this prediction and conclude that it is not reliable.

VEGA prediction for Chemical 3: **unreliable**

Agreement between the VEGA model results?

There is a good agreement between the results of the three VEGA models. In particular the difference between the values of the different models is within the range of the experimental uncertainty.

Applicability domain check and automated warnings

The automatic applicability domain check warns “compound could be OUT of applicability domain,” and cites the following reasons which, cumulatively, serve as warnings about the predictions:

- the accuracy of prediction for similar molecules found in the training set is not adequate,
- similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value,
- the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability.

Predictions and experimental values for similar compounds

As observed for T.E.S.T., when we check the similar compounds for read-across, it becomes clear that we cannot find any compound that is related to our target compound. Looking at this visually, we can see that the position of the target compound in

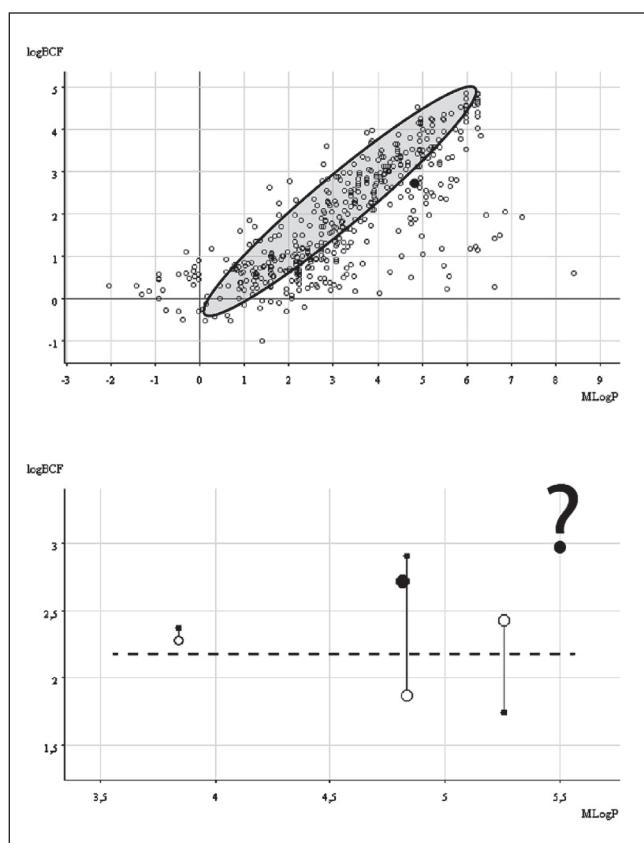


Fig. 12: VEGA scatter plot (top) and zoomed scatter plot (below) for Chemical 3

the VEGA scatter plot is not optimal, but it is not far from the central cloud. However, when we then look at the zoom, it is not possible to recognize any kind of trend associated with the three most similar compounds. This does not help in the evaluation. This situation is depicted in Figure 12.

Conclusion

For these reasons we would conclude that the prediction for Chemical 3 is not reliable.

6.2 Participant decisions on the (un)reliability of the predictions for Chemical 3

While decisions against reliability (e.g., for Chemical 1) can constitute wise caution, an unwarranted conclusion of reliability can be potentially dangerous. QSAR platforms aim to provide clear evidence that shows when an unreliable prediction is unre-

liable. This is important not only because it is vital to recognize uncertainty but also because it means that the same platforms will be trusted when they claim a reliable prediction. For this reason, Chemical 3 provides an important case study to complement the other two.

For Chemical 3, there was broad agreement that the prediction is not reliable, but a small number of participants (4) nevertheless appeared to consider the prediction to be reliable. When using VEGA, this minority was a potentially significant 7 participants (25%). This prompts the question of why they concluded the prediction was reliable.

Of the participants who ticked “yes” to reliability, 5 offered an explanation. However, for 3 of these, the affirmative answer “yes” appears to be a confirmation of the *adequacy of the information* rather than of the *reliability of the prediction*. The three comments are as follows (in full, unedited).

#27: As assumed by the predictor, there are problems with this chemical and a nice explanation is provided. It is difficult to assess similarity for such molecules, maybe they are estimated too high.

#10: The results are presented very transparent and user-friendly for assessors (regarding REACH), who are less familiar with the characteristics of BCF-determination. The description and reference to similar substance is illustrative and helpful.

#17: VEGA provides a good picture of applicability domain, showing the similar structures from training set. The analysis of applicability domain scores is very useful. MlogP descriptor should be explained.

Comments from these three participants on EPISuite and T.E.S.T. also appear to refer to the quality of the information. As a potential explanation, we have observed that for Chemicals 1 and 2, some participants rightly answered “no” to reliability on the basis of a lack of information. Equally, a “yes” for reliability involved judging the information to be adequate. So it is possible that, by Chemical 3, a few participants had shifted to using the “yes”/“no” as a review of the quality of the information, not as a decision on the reliability. (We have therefore checked that for Chemicals 1 and 2 the comments do support the judgments.)

If we conclude that these three respondents were evaluating the quality of the information rather than the reliability of the prediction, and so remove them from the results, then the responses suggest almost 90% recognition across the participants that the prediction for Chemical 3 was unreliable (Tab. 4).

The few participants who concluded positive reliability for Chemical 3 showed recognition of the uncertainty, and wanted either further evidence or to restrict the use of the result:

Tab. 4: Participant conclusions excluding participants #10, 17 & 27 for Chemical 3

VEGA		EPISuite		T.E.S.T.	
‘yes’	‘no’	‘yes’	‘no’	‘yes’	‘no’
3	22	3	22	2	23

#2: (for VEGA) Although 2 criteria for domain assessment have not been met, the analogue selection is reasonably good. I would want to check this result against other techniques, but have a good level of confidence.

#9: (for VEGA) Prediction could be reliable with restrictions (predictions for halogen compounds seem to be acceptable).

The responses for Chemical 3 can therefore be taken as somewhat reassuring: all three QSAR models provided evidence of a lack of reliability that was recognized by almost all of the expert users.

The comments from those concluding that the predictions were unreliable are similar to those we offer in our discussion above, namely that while the three predictive values from VEGA were similar, the indices, the experimental results, and the predictions for similar compounds indicated a clear lack of reliability.

6.3 A cross-platform consensus for Chemical 3?

The predictions and supporting information from all three QSAR platforms give the user clear signs of uncertainty and a consequent lack of reliability in the BCF predictions for Chemical 3:

- (i) there is a noticeable lack of consistency across the predictions from individual models within each platform, which is evidence of difficulty in making a prediction;
- (ii) the extent of that difference between the predictions, and their proximity to the regulatory thresholds, mean that (if used) they would lead to contradictory conclusions on the effect;
- (iii) the predictions for Chemical 3 are very different from the experimental results for the compounds identified as most similar;
- (iv) there is a noticeable difference also between the predictions and the experimental results for the compounds identified as most similar;
- (v) the compounds in the training sets and test sets used to create the model are not actually similar to the target compound in terms of their molecular and/or physical-chemical properties, and this applies even to those compounds identified as most similar; it means that even a manual read-across would be unreliable.

For all of these reasons, VEGA, EPISuite, and TEST all indicate that they cannot offer reliable predictions for Chemical 3; the uncertainty associated with the predictions is high. We therefore cannot give any estimation of the BCF value or classify the compound from the experimental data and QSAR models currently available. (We also can find no experimental data for Chemical 3.) Figure 7 therefore shows the range of predicted values only for Chemical 3.

7 Concluding comments

The rationale for focusing on the use of QSAR models *in practice*, and the discussion of the results of the case studies, are included in the previous sections. Further responses to some

criticisms of the platforms by participants are included in supplementary file 6 at <http://www.altex-edition.org>. Here we offer only a few concluding comments.

In this review exercise and article we set out to address the central practical challenge of how to determine whether a toxicity prediction for a compound is reliable, and how to discuss the predictions and supporting information explicitly so that the evidence and reasoning can be understood, reviewed, and potentially accepted by others. Specifically, we (i) described the supporting information provided by three leading QSAR models and discussed and illustrated its practical use in reviewing the reliability of predictions, and (ii) involved expert toxicologists across sectors in a case study review of using that information *in practice*. From both, it is clear that the use of QSAR models requires considerable toxicological understanding. The study affirms the important role of the human expert in producing decisions on toxicity and reliability. Specifically, it shows the need for toxicological expertise to use the evidence to draw conclusions about toxicity, and not just expertise in the use of QSAR software. We hope that the explicit elaboration of the information and its potential use will promote discussion and encourage further case studies.

In terms of user decisions from the QSAR models, the results of this small review exercise are encouraging. Users recognized the evidence of reliability and of uncertainty, and they applied their toxicological expertise to reviewing that evidence. Where participants concluded a lack of reliability for Chemicals 1 and 2, it was from a wisely cautious response to a perceived lack of evidence (or competing evidence in the case of Chemical 2), and seemed to be based on insufficient experience in using the outputs rather than any fundamental misunderstanding of the outputs.

Given the need for discussion between experts on toxicity evaluations that have real material consequences, and for common understanding across sectors, it was particularly encouraging to find a shared basis for discussion and potential agreement between experts, even for Chemical 2 where the decision is not self-evident and depends on weighing up potentially contrary evidence.

The review nevertheless highlights certain priorities for the supporting information, including providing the user with the opportunity to review the evidence in a read-across process. We hope that by describing the “state of the art” in the supporting information provided, and its use in practice, this article will contribute to raising expectations of what QSAR platforms can provide. The pursuit of rigor and reliability by users of QSAR models is not only a regulatory requirement, it is also essential for developing toxicological evidence that can reduce and replace costly *in vivo* tests.

QSAR models are rapidly evolving and improving. From our involvement and experience, we would want to emphasize that different models and systems are available, and that they together improve the capability of the human expert to evaluate compounds. In the same way that REACH encourages a shift from a single test to a “weight of evidence,” it is not useful to think of the different models in terms of a search for a sin-



gle “best model.” Professional advice is to use more than one model, and more than one platform, whenever possible. Being based on different sets of experimental data, and using different molecular descriptors, they together have the potential to improve the evaluation of a target compound and to strengthen the quality of the evidence.

Note: This review exercise can be used and adapted in training: the full output files are in supplementary files 1 (VEGA), 2 (EPISuite), and 3 (T.E.S.T.), the guidance information is in supplementary files 4 and 5 (all at <http://www.altex-edition.org>), and the exercise is online at <http://www.orchestra-qsar.eu/webforms/231>.

References

- Annot, J. A. and Gobas, F. A. P. C. (2003). A generic QSAR for assessing the bioaccumulation potential of organic chemicals in aquatic food webs. *QSAR Comb Sci* 22, 337-345.
- Benfenati, E., Gonella Diaza R., Cassano, A., et al. (2011). The acceptance of in silico models for REACH: Requirements, barriers, and perspectives. *Chem Cent J* 5, 58.
- ECHA (2008). Guidance on information requirements and chemical safety assessment: Chapter R.6: QSARs and grouping of chemicals (p.10). <http://echa.europa.eu/web/guest/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment>
- ECHA (2009). Practical guide 5: How to report (Q)SAR. 2009 (p.5). http://echa.europa.eu/documents/10162/13655/pg_report_qsars_en.pdf
- Lombardo, A., Roncaglioni, A., Boriani, E., et al. (2010). Assessment and validation of the CAESAR predictive model for bioconcentrating factor (BCF) in fish. *Chem Cent J* 4, Suppl 1, S1.
- Meylan, W. M., Howard, P. H., Boethling, R. S., et al. (1999). Improved method for estimating bioconcentration / bioaccumulation factor from octanol/water partition coefficient. *Environ Toxicol Chem* 18, 664-672.
- Pardoe, S., Benfenati, E., and Cazzato, L. (2010). Interview with Professor Wim de Coen, Head of Evaluation Unit 1, ECHA. (Video interview: 7 min). <http://www.orchestra-qsar.eu/resources/videos/interview-wim-de-coen>
- Pardoe, S., Cazzato, L., Golding, A. et al. (2011). QSARs in REACH? Uses, issues and priorities. Video documentary based on 20 interviews with regulators, industry representatives and expert QSAR developers. (30 min.). Lancaster: PublicSpace Ltd. <http://www.qsars-in-reach.researchdissemination.eu>
- Zhao, C., Boriani, E., Chana, A., et al. (2008). A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* 73, 1701-1707. <http://www.sciencedirect.com/science/article/pii/S0045653508011922>

Acknowledgements

We acknowledge the EC project ORCHESTRA for funding. We acknowledge all the participants of the exercise and Dr Ilias Koutinas, Patras University, for his assistance on the web version of the exercise.

Correspondence to

Emilio Benfenati, PhD
Mario Negri Institute for Pharmacological Research
Laboratory of Environmental Chemistry and Toxicology
Via Giuseppe La Masa, 19
20156 Milano
Italy