Chemistry Central Journal

COMMENTARY Open Access

The acceptance of *in silico* models for REACH: Requirements, barriers, and perspectives

Emilio Benfenati[1*], Rodolfo Gonella Diaza[1], Antonio Cassano[1], Simon Pardoe[2], Giuseppina Gini[3], Claire Mays[4], Ralf Knauf[5] and Ludger Benighaus[6]

Abstract

In silico models have prompted considerable interest and debate because of their potential value in predicting the properties of chemical substances for regulatory purposes. The European REACH legislation promotes innovation and encourages the use of alternative methods, but in practice the use of *in silico* models is still very limited. There are many stakeholders influencing the regulatory trajectory of quantitative structure-activity relationships (QSAR) models, including regulators, industry, model developers and consultants. Here we outline some of the issues and challenges involved in the acceptance of these methods for regulatory purposes.

Introduction

The concern to protect human health and the environment has prompted significant change in EU regulation on chemical substances. The REACH legislation requires industry to evaluate the toxicity not just of new chemicals, but of the tens of thousands of existing chemical substances that are in use but have never been subject to regulatory testing. Many argue that to achieve this by traditional testing would take decades, cost billions and consume many millions of vertebrates. It is therefore significant that the legislation explicitly encourages innovation in toxicity evaluation, demands the use of existing data where possible, and states that further animal testing can only be used 'as a last resort'.

In this context, *in silico* models are potentially invaluable because they use computer technology to connect, use and extend existing experimental data, and can be used to assess thousands of chemicals quickly. The REACH legislation sets out conditions specifically for the use of QSARs, and the European Chemicals Agency (ECHA), in its role in implementing REACH, offers detailed guidance. Even the basic 'Guidance in a nutshell' on substance registration advises industry to 'collect QSAR estimated results for the substance if suitable models are available' as an initial step [1].

However, in practice, the use of *in silico* models within REACH by European industry is still very limited. In moving forward, the current focus is therefore on the highly inter-connected issues of

(i) acceptance by regulators (ECHA and the national competent authorities),
(ii) uptake by industry and consultants, and
(iii) the progress of model developers in designing and developing models specifically for regulatory use.

It is now clearer to developers and users what is expected of *in silico* models for regulatory use. The regulatory framework of REACH, the OECD principles and the ECHA guidelines effectively work together to increase the demands on models in terms of rigour, reliability and transparency. Yet despite progress, there are still inevitable limitations in terms of

(a) the models currently available,
(b) the experience of using models for regulatory purposes, and
(c) the experimental data available for developing new models for particular endpoints.

Current policy is therefore that the use of *in silico* models will be accepted or rejected by EU regulators strictly on a *case-by-case* basis, following regulatory review that examines the details of the models used and

* Correspondence: emilio.benfenati@marionegri.it
[1]Istituto di Ricerche Farmacologiche "Mario Negri", Via La Masa 19, 20156, Milano, Italy
Full list of author information is available at the end of the article

Chemistry Central © 2011 Benfenati et al

of the ways they have been used within a submission. The intention seems to be to avoid a situation in which individual successful and unsuccessful examples of using *in silico* models could be over-generalised as indicative 'good' and 'bad' case studies.

An important way forward in developing a wider understanding of QSARs for regulatory purposes is therefore to step back from citing individual examples, and instead to elaborate and increase understanding of the often-complex issues around model development, use and acceptance. That is our focus in this paper.

The three primary stakeholder groups in this process (regulators, industry and consultants, and model developers) each has particular priorities and concerns, and particular expertise to bring to the debate around the use of *in silico* models. At the same time, their institutional requirements, and existing practices, may inevitably create important barriers to the further use of *in silico* models. Dialogue and shared understanding are therefore vital. Moreover, these primary stakeholders operate within a wider stakeholder context in which the interest (or lack of interest) from toxicology research communities, industry managers, shareholders, non-governmental organizations (NGOs) and citizens may be highly influential. A range of issues is intrinsically involved. Before describing some of those issues, it is useful to foreground the basic issue that expertise, functionality and judgement are central to regulatory toxicology, including the use of *in silico* models.

In silico models are usually developed by scientists, potentially across the fields of toxicology, chemistry, engineering and computer modelling. They are also developed for particular purposes, including academic research and industrial product development, and are produced both from government research funding and commercially. In the US, Canada and Denmark they have been widely developed by regulators.

The challenge for developers is to produce models that are not just 'scientific', but also functional and 'fit for purpose' within the regulatory context. The regulation inevitably makes demands on the scientific evidence that must be provided by industry to support claims that the safety and toxicity of a chemical have been measured and are known. Decisions in the design and acceptance of *in silico* models for *regulatory* purposes therefore have to be oriented to an understanding of regulatory as well as scientific aspects.

As with the historical case of defining geographical coordinates, the development of rules proceeds through a combination of scientific and political judgements. So thresholds may vary according to different regulations, and in different countries.

Regulatory thresholds and criteria also involve technical judgements, not just in terms of the numerical value, but in the way that numerical value is measured or expressed. For instance, for classes of aquatic toxicity defined for regulatory purposes the units are typically expressed as concentrations measured as weight per volume. Yet it is more usual for scientists to refer to the toxicity process in terms of moles per unit of volume, because the toxicity is assumed to be related to the molecule itself; thus, it seems more appropriate to use mmol/l instead than mg/l. This example simply illustrates that many *technical* judgements can also be involved in the setting of a threshold or criterion, and in the evaluation of an *in silico* model.

Complex technical expertise, evidence and judgement similarly form the basis for an *in silico* model. In observing and modelling the relationships between the molecular structure of carbon-based chemicals and their biological activity, both sides of this relationship are complex and multi-dimensional: there are thousands of ways in which a substance and its molecular structure can potentially be described, and many endpoints involve complex and sequential biological processes. There is also a range of technical factors that form the basis of an *in silico* model and determine its rigour and reliability. They include: the number and nature of the chemicals which are used to build the model; the nature of the observed property for this series of chemicals; the quality of the available experimental data; the way the chemical structures are described (e.g. SMILES or sdf file); the range of chemical information used for modelling purposes (such as chemical descriptors, or fragments); the mathematical algorithm at the basis of the model; the ways in which the developer checks the statistical characteristics of the model; and the ways in which the model expresses predictions and enables the user to identify associated uncertainties. The factors involved in the development of *in silico* models for regulatory purposes are discussed by Benfenati [2].

Given the range of endpoints, the vast numbers of chemicals, the large palette of mathematical algorithms available, and the potential to use tens of thousands of chemical fragments and thousands of chemical descriptors to build a predictive model, it is clearly possible for future scientists to generate huge numbers of *in silico* models. It is easy to imagine an explosion in the number of useable models, many with quite similar performance.

Discussion

The regulatory conditions for the use of QSAR models

In developing, using and discussing *in silico* models for regulatory purposes, it is necessary to start by recognising the requirements which have been defined by law. It is necessary to address a specific regulation, since different principles may apply to different regulations. Here we will refer to REACH, the European regulation for the

Registration, Evaluation, Authorisation and Restriction of Chemicals [3]. This means that certain assumptions may not be valid for other regulations.

REACH explicitly refers to *in silico* models, including quantitative structure-activity relationships (QSAR) and structure-activity relationships (SAR) models. (Other *in silico* models exist, such as those used for virtual screening, in which a ligand molecule is bound to a natural macromolecule through specific programs [4].) Below we will use the term QSAR to refer to both QSAR and SAR models.

According to REACH regulation (Annex XI) a '(Q) SAR is valid if:

- the model is recognised as scientifically valid;
- the substance is included in the applicability domain of the model;
- results are adequate for classification and labelling and for risk assessment;
- adequate documentation of the methods is provided.'

These four 'conditions' address important scientific and practical aspects of the use of *in silico* models, including those specific to the application. We discuss each of these below (in a different sequence).

The four conditions are also specific to the regulatory context. For example, from a scientific point of view a model can be a good model even if full documentation is not provided. Similarly, industry develops models for its own purposes, using algorithms and data which are restricted and confidential, intended for endpoints which differ from those of interest within REACH; these models might be scientifically correct but are inappropriate for REACH purposes.

Adequacy for classification and labelling and for risk assessment

The condition of adequacy (the third listed above) states that to be acceptable for use within REACH, the results from a QSAR model must be 'adequate' for the particular regulatory function. This condition has also been described as requiring that 'the prediction should be fit for the regulatory purpose' [5]. Classification and labelling (C&L) and/or risk assessment (RA) are the two uses identified, though a third use is the prioritisation of substances, (addressed by REACH in Article 44). Thus QSAR models can be used for three different uses or purposes within REACH.

These three regulatory uses can each require the use of different QSAR models, because different endpoints are addressed, and because different levels/kinds of certainty are required. This fact has not been discussed in detail so far, and needs wider attention.

Most attention has naturally been given to the use of QSAR models to provide values for the registration of substances. This is one of the most important aspects, compared to the other possible uses, and in general, risk assessment (RA) evaluation is requested for higher tonnage chemicals. RA can require continuous values, in order to calculate the mathematical ratio between the toxic dose and the exposure level. We therefore have to remember that some models, especially for human toxicity, do not provide continuous values, so in such cases of substance registration where continuous values are required, the use of QSAR models can be limited.

The demands for C&L contrast with those for RA in several important respects. Firstly, different endpoints have to be addressed for C&L, secondly, a yes/no classification requires less accuracy in terms of the relationship between toxicity and dose levels, and thirdly, within REACH C&L is carried out on the basis of the information available. There is therefore potential for a significantly wider use of QSAR models in the evaluation of chemicals where there is a lack of existing data for C&L. Arguably, all chemicals could be assessed using the available models, and this would probably increase the safety of those chemicals.

For C&L and RA, the nature of the QSAR model required may be different in many cases. C&L involves creating a classification of whether a chemical is toxic or not for a particular endpoint, so it requires classifier models. Yet this distinction in terms of what is required is more complex than it appears. Even for QSAR models which are typically classifiers, such as those used to predict carcinogenicity, it may be more appropriate and useful to have a quantitative evaluation of the potency, because if the chemical is carcinogenic then it is important to be able to evaluate the risk at a given dose level. Moreover, for certain endpoints, whether a continuous value or category value is required by REACH, depends on the tonnage of use on the market. For example for the bioconcentration factor endpoint, only the simple classification (of bioaccumulative or not) is required for lower tonnage chemicals, while a continuous value is required for higher tonnage chemicals (above 100 tonnes/year).

In the third regulatory use of QSAR models, for prioritisation, the situation appears to be more straightforward. Neither the lower number of endpoints of interest, nor the potential uncertainty of the results, is critical, because the goal is to determine which chemicals require higher scrutiny.

The implications in terms of evaluating the adequacy of a QSAR model is that we must always be very clear about the regulatory purpose for which it is being used and the relevant regulatory requirements. A given QSAR model may be adequate and very suitable for a specific use within

REACH, while being clearly inadequate and therefore unsuitable for a different use.

Providing documentation of the methods

REACH requires detailed documentation to be submitted by the registrant of a substance. If QSARs are used, this documentation needs to include details of (i) the QSAR models and the data sets they are based on, and (ii) the ways in which the models have been used to assess the toxicity/safety of the target chemical. The intention is that the documentation should enable the regulator to conduct an independent and informed evaluation of the chemical safety evidence provided.

Hypothetically, we might imagine a model which erroneously predicts certain compounds to be safe, because it was designed to do that (in order to allow the use of a certain compound on the market). If the model details are not made available (perhaps for confidentiality issues) the regulator would be unable to identify the false predictions. A clear priority for regulation is to avoid the possibility of such a scenario.

In a more realistic scenario, it is possible that two models could give opposite predictions for the same chemical. If the two models were confidential, it would not be possible for either the user, or the regulator (or even a peer reviewer) to determine why the two models had produced opposite results. However, if the appropriate documentation is available it may then be possible to identify, for example, that while one model was built on a training set of compounds that are similar to the target chemical, the second model was not.

In this way, the reliability of a model, and the authority of its predictions, depend on its transparency. The issue of providing 'adequate documentation of the models' and their use, is central to demanding and ensuring the quality of models and the quality of their use, now and in the future. It is ultimately about enabling the regulator to evaluate chemical safety assessments in the interest of human health and the environment.

In our experience it is actually preferable for more than one model to be used, and for the user to analyse how the models work in order to understand the potential significance of any difference in their predictions for a target chemical. (The example of two models with different predictions, above, illustrated the issue of whether the target chemical was within the applicability domain of the model, explained below.) With large numbers of QSAR models now available within and across the main platforms, and with the fast processing of chemical data made possible by QSAR technology, the use of several models is feasible in practice as well as advisable.

In the case of commercial software, the full documentation of each model is typically not available. Indeed, the algorithm and the training set used to build the model are often confidential. Despite this, the use of commercial programs has not been explicitly criticised by regulatory authorities. In reality, commercial programs have been used in the USA and Europe, because of their ease of use, their availability and because they represent a major source of QSAR models. We do not foresee restrictions in their use, and in our opinion it would be a pity to renounce these models. However, it is likely that in the case of two similar models being available, one commercial and one freely available, the user will prefer the second one. Within REACH, it also seems likely that the regulators will demand more transparency from the commercial models, and/or prefer the use of fully documented models.

Inclusion within the applicability domain

The REACH legislation does not list a series of 'approved' QSAR models, and ECHA does not intend to. The purpose of REACH is *not* to define QSAR models, but to register, evaluate, authorise and restrict chemicals. REACH states in broader terms that the validity of QSAR models must be characterised and documented. The ECHA guidance suggests this be done by referring to the five agreed OECD principles. One of these principles is that 'The (Q)SAR model should be associated with a defined domain of applicability' [5]. The term 'domain of applicability' refers to a defined set or range of chemicals for which the model is intended to be used.

The related condition for using a QSAR model within REACH (that 'the substance is included in the applicability domain of the model') therefore builds on this principle in practice. In effect it requires that

> (i) each model being used has a clearly defined domain of applicability, and
> (ii) this has been followed in practice, so that the chemical being evaluated actually lies within the range of chemicals for which the model was designed and intended.

Every user of QSARs has to be aware that QSAR models are only appropriate and reliable for specific sets of chemicals. A highly reliable model will not produce reliable results for chemicals that lie outside the domain of applicability.

This basic issue is true to varying degrees for all methods of toxicity evaluation. Within REACH the concern is to protect human health and the environment, so even traditional animal tests are actually only surrogates or *models*; their reliability depends on testing relevant effects in relevant animals and anticipating relevant biological and environmental processes. As history has shown, such tests do not provide certainty about the ultimate effects of a chemical on human beings or the

ecosystem. The priority when using any method is to acknowledge that all results bring uncertainty, and to recognise and understand the strengths and potential weaknesses of the particular approaches being used. In the case of QSAR models, a clear strength is the ability to make use of literally thousands of experimental results to both build and test the model, and to be able to select from the thousands of potential descriptions of the chemical substances those descriptors which most accurately correlate with the observed toxicity. But the inevitable limitations come from the reliability of those initial animal tests, from the rigour within the model development and review process, and from the similarity of the target chemical to those tested.

Hence the use of experimental models, and QSAR models, requires an initial evaluation of their correct application for the situation of interest. A question that is frequently asked by potential users of QSAR models is whether they can be used for a particular compound. This vital evaluation is addressed by the applicability domain (AD). Some tools even have been developed to advise the user. However, most QSAR methods use the 'chemical space' of the chemicals in the training set, and evaluate whether the target chemical is similar to these in relevant ways (the chemometric approach) [6]. A more complex approach has been developed for the CAESAR models [7]. In this case, the evaluation of whether the model is appropriate for a particular target chemical is based on the accuracy of the predictions produced by the model for compounds that are similar to the target compound, and for which there are experimental results.

(It should be noted that not all QSAR models include an explicit applicability domain. In some cases the AD cannot be addressed for intrinsic reasons. This may be the case for models which search for molecular fragments which have been identified as producing toxicity, from current biological understanding. Such models typically do not refer to a training set of compounds, even though the list of toxic fragments is inevitably incomplete. Thus, if models like Derek or Toxtree produce the result that no toxic fragment is identified, we cannot then infer that the target chemical has no toxicity; the negative result may be due to lack of current knowledge on the toxic fragment. Such models have inevitably been criticised for this limitation.)

The scientific validity of the model

We started this paper by stating that many technical factors form the basis of an *in silico* model and determine its rigour and reliability, such as the number of chemicals in the training set, the quality of the data, the chemical information, and the algorithm. We have also observed that requirements from REACH impact on this evaluation, and different criteria apply for different uses, and different levels of reliability may be demanded for different chemicals for several important reasons.

In its guidance for industry and regulators, ECHA states that the validity of QSAR models for regulatory purposes is 'characterised and documented according to the five agreed OECD principles' [5]. These principles were adopted by the OECD in 2004, after exploring the complexity of factors involved in the regulatory evaluation of QSAR models [8]. The principles were intended to anticipate a range of potential regulations, so it is important to note that they were not related to REACH and the legislation itself does not refer to them. However, in the implementation of REACH, ECHA and others usefully refer to them. As ECHA's guidance reports [9] the OECD stated that 'to facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

1. a defined endpoint;
2. an unambiguous algorithm;
3. a defined domain of applicability;
4. appropriate measures of goodness-of-fit, robustness and predictivity;
5. a mechanistic interpretation, if possible.'

While these OECD principles address the qualities of QSAR models, ECHA focuses on the quality of their use in practice. ECHA emphasises that:

> The principles constitute the basis of a conceptual framework, but they do not in themselves provide criteria for the regulatory acceptance of (Q)SARs. Fixed criteria will be difficult, if not impossible, to define in a pragmatic way, given the highly context-dependent framework in which non-testing data will be used. Instead, experience and common understanding should be gained by a learning-by-doing approach, and by documenting the learnings...[9]

This caveat illustrates the policy of ECHA that the use of QSAR models must be evaluated by industry and regulators 'on a *case-by-case* basis' [9]. As we have said, the validity of predictions depends not only on the scientific quality of the models used, but also on the regulatory function and the target chemicals that they have been used for, and on how they have been used.

In this way, REACH does not refer to *validated* models, but instead to the use of a *valid* model (Annex XI). Acceptance is not of the model, as something universally valid, but of its use for a certain regulatory function and chemical substance. (Moreover, with potentially thousands of QSAR models, an official validation process would take years and be constantly out-of-date in the evolving technology.)

Other examples of this REACH / ECHA focus on the quality of use in practice (already described above) include whether the particular *target chemical* is within the applicability domain of the model, whether the endpoint, algorithm, data sets and applicability domain are actually *documented* to support the prediction, and whether the model predictions are adequate in practice for the *specific regulatory functions* of risk assessment and classification and labelling.

The OECD emphasis on the statistical evaluation of QSAR models is worth noting. (ECHA guidance also articulates principle 4 as the 'appropriate performance of the model (the statistical "goodness" of the model, robustness and predictivity)'[5].) This statistical focus addresses the functional shift from using QSARs for scientific exploration to using them for *prediction* in the regulatory protection of human health and the environment. Documents from the OECD [8] and ECHA [10] highlight the need to use models which have a sound statistical basis both in terms of the model development, and (for particular sets of chemicals and a particular endpoint) in terms of the proven quality of their predictions.

Most early QSAR models were interested in the description of a particular phenomenon, and in the explanation of the physico-chemical factors involved. At that stage, the aim was not to predict the biological activity of compounds outside the set of the chemicals used to build the model. Instead, the aim was to understand whether certain parameters were related to the observed phenomenon within that set of chemicals. Typical examples are the models for aquatic toxicity, in which toxicity was correlated with logP (the logarithm of the partition coefficient between octanol and water).

The central issue is that a scientifically plausible model is not necessarily a predictive model. The current interest in developing and using QSAR models to *predict* the property value of chemicals, and in using them where possible to replace traditional experimental models, rightly brings new demands for evidence of the model's predictive reliability and of its relevance to the target chemical. In other words we take Galileo's lesson to provide experimental evidence of a theory; including of each claimed relationship between molecular structure and biological activity.

Interestingly, if a model developer uses a range of descriptors to create a perfect fit between structure and observed activity for the initial group of chemicals with known experimental results (the 'training set'), this can result in the model producing poorer predictions for other related chemicals. So developing a model for reliable prediction requires a slightly different approach. Then to evaluate the model, it is necessary to test whether it can make accurate predictions for chemicals that were not used to develop the model, but for which experimental measures of toxicity are available (a 'test set'). The prevalence of active (toxic) and non-active compounds in the training set and test set is of course important. (The methodologies for performing internal and external validation are available in the literature [2].)

When evaluating a model, it is clearly important to address the regulatory significance of errors in predictions, rather than simply their size. For example, while the under-prediction (false negatives) and over-prediction (false positives) of toxicity may appear equivalent from a purely statistical point of view, these two types of errors are not equivalent for the regulator, industry or citizen. Regulators want to avoid evaluating a toxic chemical as safe, yet the opposite error is less critical. They therefore generally prefer a conservative model. Conversely for industry, an error which wrongly predicts a chemical to be toxic (false positive) represents a potential economic loss.

Typically, QSARs for risk assessment are regression models which use statistical evaluation based on the errors squared. This does not distinguish between over- and under-estimations. Only a few models have so far taken this into account: those developed within the projects DEMETRA [2] and CAESAR [7,11-15] gave different weights to under-estimations when building the models. However, for QSAR models designed as classifiers, it is common to evaluate separately the specificity and sensitivity, or the false positives and false negatives, so the evaluation is useful when selecting a model for a particular regulatory purpose.

The evaluation of whether or not a model is capable of correctly predicting the toxicity of chemicals is of course also addressed within the issue of the applicability domain. But once again we see the complexity, and how different aspects are linked. In case of the predictivity, the evaluation is on the basis of the quality of predictions for a population of chemicals. In case of the AD analysis, REACH asks whether the model is reasonable for a specific substance. Other aspects can therefore be included in that AD analysis, which are not statistical. An AD can be formulated also on the basis of certain chemical moieties present in the molecule, and/or from an interpretation of the mechanism of toxicity for the chemical of interest.

The term 'mechanistic interpretation' refers to a current scientific interpretation of the mechanism by which the chemical produces the biological activity. The OECD recognised that it was not always possible, so added 'if possible' to the fifth principle. It seems worth noting that REACH does not mention it. ECHA merely requests that any mechanistic associations between the descriptors used in a model and the relevant endpoint be

documented, because 'it can add strength to the confidence in the model' [9].

The importance of a mechanistic interpretation is a point of debate. Some stakeholders involved in the use of *in silico* methods see it as vital in providing a scientific rationale both for grouping a set of chemicals in terms of the particular endpoint, and for justifying why the target chemical is related to them. Others, including ourselves, argue that the highly complex nature of molecular structures, and the highly complex and sequential nature of biological and environmental toxicity, inevitably mean that any mechanistic interpretation is only a current hypothesis. We argue that it is important to base QSAR models and predictions on demonstrable correlations between measured toxicity and molecular qualities, rather than limiting these models to the few cases where there is scientific knowledge of the toxicity mechanism.

Clearly, stakeholders can feel more confident in a prediction of toxicity if, for example, we can see that a number of chemicals have the same toxic fragment that is present in the target chemical. Similarly, we may recognise that a scientifically plausible toxic mechanism of action is likely, due to the presence of a certain chemical moiety in the molecule. Yet it is vital to realise that these assumptions do not represent certainty for a given mechanism, or for toxic activity in the target chemical; they are merely useful indications that this is likely to be the case. The mere presence of a certain toxic fragment is not conclusive. There may be chemicals which contain the fragment but do not show toxicity: for example, the toxic fragment may not be accessible due to steric factors, and/or other components in the molecule may prevail and result in a detoxification process.

While certainty is not achievable from any model (including experimental models), we argue that in order to provide a better appreciation of the probability of a particular toxicity, it is vital to generate a sound statistical analysis of exactly how many chemicals with a certain toxic fragment (for example) are really toxic, and how many are not toxic. A statistical analysis can provide a probability of toxicity for a particular chemical, and so provide a quantitative basis for assessing risk. It also provides a basis for further investigation of the complex factors producing the toxicity.

The needs of different stakeholders and barriers to the use of QSAR models

The preceding discussion has indicated some of the reasons for the current limited use and acceptance of QSAR models in practice within REACH, and the challenges. Clearly, different stakeholders have different needs in relation to increasing the use of QSAR models. In a separate paper (forthcoming) we report an initial investigation and analysis of these barriers and needs. Here, we focus on some core issues from experience and discussions with stakeholders, in order to help all stakeholders to recognise the commonalities, differences and challenges.

There are at least seven reasons to use QSAR models and other *in silico* methods within REACH:

1. **Innovation** in the evaluation of toxicity is encouraged by the REACH legislation; the development of alternative methods is given as one of the purposes of REACH.
2. The **time** necessary to produce the data requested under REACH for tens of thousands of existing chemicals would be very long if *in vivo* approaches were used, probably decades; however, where sufficient experimental data sets are already available from *in vivo* or *in vitro* tests of related substances, then *in silico* methods can be used to assess thousands of chemicals in a day.
3. There is a **lack of laboratories** in Europe capable of performing *in vivo* tests for such large numbers of substances; conversely *in vitro* tests require less investment, and *in silico* methods can be employed on an office computer.
4. The **costs** of carrying out these evaluations with *in vivo* methods would be billions of Euros; alternative methods, and especially *in silico* methods, are cheaper by orders of magnitude.
5. If *in vivo* methods were used, the evaluations would consume many millions of vertebrates; alternative methods can **reduce or replace the use of animal testing**; this is not only a stated priority of REACH for regulators and industry, but it is also increasingly desired by industry shareholders, managers and consumers, and by citizens and their representatives.
6. *In silico* methods offer tools for the **prioritisation** of chemicals according to their predicted toxicity. This means that the time delays, economic costs and use of vertebrates can all be reduced by simply targeting the use of traditional *in vivo* tests on those substances which have the highest probability of toxicity and higher risk.
7. The predictive ability of *in silico* methods enables a **proactive approach** within product development and regulatory toxicology. Toxicity evaluation can be brought 'upstream' in the product development and decision making processes, so that chemicals are selected and products developed to be non-toxic from the beginning, rather than the adverse properties being revealed only late in the process, with economic and other consequences.

Regulator interest in the use of QSAR models is prompted and encouraged by the explicit promotion of

innovation and alternative methods for the evaluation of toxicity within the REACH legislation. This is made clear in the first article of the legislation, and comes from a regulatory awareness both of the limitations of traditional methods and of the new possibilities created by toxicology research in recent years. Given that the priority of REACH, and the focus of ECHA and national regulatory authorities, is to protect human health and the environment, then the issue for all is how to combine and use methods of toxicity evaluation in order to achieve it.

REACH and ECHA actively encourage industry registrants to use the range of different sources of information available, and to integrate these within a 'weight of evidence' approach. In this way, QSAR models can be used both as an alternative and a supplementary source of information. Their use as an alternative source and only source of information is possible when the reliability of the results is sufficient. Their use as a supplementary source is possible when further information or confirmation is needed from other sources.

It is often valuable to use more than one model, based on different descriptors, algorithms or approaches, so that the results can give different perspectives and so contribute to the assessment. Several articles have shown that consensus approaches yield better predictions [2,16-18]. We suggest that regulators should encourage industry to exploit as many appropriate QSAR models as possible, in order to comply with the REACH demand to generate the most information for each chemical while avoiding the unnecessary use of animals. Depending on the purpose, different kinds of models may be suitable, and different levels of uncertainty may be acceptable (as discussed above) in achieving the over-riding priority of protecting human health and the environment.

For each use of a QSAR model, regulators need, and require, the information we have outlined and discussed in the previous sections. This includes transparency in the model, clarity in its intended endpoint, applicability domain, predictions and associated uncertainty, and above all, information that directly addresses the regulatory demands.

Looking ahead, in this period of rapid change, national regulators also need to be kept abreast of developments within and beyond Europe. For instance, the Tox21 [19] and ToxCast [20] initiatives in the USA are involved in screening thousands of chemicals for toxicity and are expected to reshape the toxicological procedures for the evaluation of chemicals. By providing up-to-date data for thousands of chemicals, they will increase the need for *in silico* models, in order to analyse and make full use of that data for future evaluations.

Industry in Europe now faces an urgent need to provide toxicological data on their existing and new substances in order to maintain a market for their products and sustain their economic interest. The ultimate threat from REACH is simply 'no data, no market'. The potential to use cheaper and faster methods of toxicity evaluation is therefore extremely attractive to many in industry. Yet they also need to be sure firstly that the evaluation will be accepted by regulators, and secondly that their evaluation is reliable so that they maintain the quality and reputation of their products as safe. If it is uncertain whether regulators will accept a QSAR prediction, then industry will prefer to use a traditional experimental method in order to avoid uncertainty and delay in regulatory approval. Similarly, if those in industry themselves lack confidence in the prediction, they may again prefer to stay with a traditional experimental method. Thus, the position of industry on using QSAR models can be both different to that of regulators, and closely related to the acceptance by regulators.

It is important for all to recognise that QSAR models compete with the more standardised *in vivo* methods, where the procedures are very clearly defined in official protocols. Even though the *in vivo* tests contain uncertainty, there is evident confidence from years of experience that they will receive regulatory acceptance if it is shown that the official protocol has been followed. The level of uncertainty for industry in achieving regulatory approval is therefore currently lower if traditional methods are chosen rather than QSAR models.

There is also currently far less knowledge in industry about QSAR models than about traditional experimental methods. Much more dialogue and training is necessary to build understanding and confidence. This needs to explain what is possible with QSAR models, but also address the limitations. Honesty is vital if the technology is to be understood and trusted, and if potential users are to be able to judge how to use it wisely.

Consultants are in a potentially central role in terms of whether innovation and the use of alternative methods actually happen in practice. While large companies have their experts and laboratories in-house, small and medium enterprises usually rely on consultants for the evaluation of the toxic properties and environmental issues. Consultants usually have a background in toxicology and provide small companies with vital advice and information. However, in many cases they have access or links to laboratories which do animal experiments, and as a result, there may be a tendency to be sceptical about alternative methods, and a potential conflict of interest may result in a preference for using animal models. This needs be taken into account by regulators and others when reviewing the positions, needs and interests of different stakeholders.

Model developers (scientist and engineers) are in the professional role of exploring and developing new

methods. However, in contrast to engaging in general scientific research, those developing QSAR models for regulatory use need to recognise that their focus is the specific regulatory application of the more general QSAR methodology. Developing scientific models without first developing a detailed knowledge of the specific regulatory requirements will lead only to frustration.

In practical terms, it is useful for a developer to turn the REACH and ECHA demands into a check-list of actions, so that the model specifically addresses the regulatory demands. In most cases this will help in developing the explicit scientific rigour and reliability of the model. Any lack of documented information about a model is simply a guaranteed barrier to its acceptance. Models should be as transparent as possible, in terms of the explanation of the model, its subsequent use, and the clarity of the result.

Developers of QSAR models clearly have their own interests, needs and desire for recognition. But the shared goal among developers has to be to produce a broad palette of different, but focussed, applications. Given the value of using more than one model in an evaluation (discussed above) developers should not see their models as being in competition with those created by others. Different models, especially if they are based on different assumptions and approaches (such as on the basis of toxic fragments chemical descriptors or statistical methods) can increase the confidence in the QSAR prediction if the predictions agree. If they disagree, an evaluation of the reasons can provide the user with insights for the assessment of the chemical, and help to define the uncertainty of the results. In both cases, the value and contribution of QSARs is strengthened by the complementary use of different models. It is therefore important for developers not only to engage in constructive peer review, and to draw understanding from potentially different predictions, but also to avoid merely criticising the methods of other developers in order to promote their own method or seek recognition. This will bring confusion and reduce general acceptance. We are in a positive scenario where models are available, cheap and fast to use, and can be used in combination, so the potential for difference and complementarity between models becomes a strength and an asset for all.

Citizens and NGOs who are active in the protection of animal rights will favour QSAR models for reducing animal testing, albeit also partly relying on animal testing. Citizens associations more widely can benefit from the accessibility and facilitated use of QSAR models because these models may offer ways to investigate the adverse effects of substances. Clearly, the use of predictions from QSAR models by any stakeholder requires some understanding of what makes them reliable.

Future directions

QSAR models now and in the future may prove to be an increasingly valuable technology, with a potentially important function for protecting human health and the environment. It is therefore vital to avoid the familiar trajectory of a developing technology being initially overstated in terms of its capabilities, and then discredited when it is over-applied. It is in the interests of all stakeholders that QSAR models are explained openly and used appropriately with care.

Our intention in this paper has been to elaborate and increase understanding of the seemingly complex issues around model development, use and acceptance within REACH. Our aim is to generate debate and understanding among and between the different stakeholders, as a way of exploring how stakeholders can benefit from QSAR models and can together increase the regulatory use and acceptance of QSAR models in practice.

As the previous sections have shown, ongoing debate and understanding need to look beyond the technical issues in using QSAR models, to the other significant challenges raised by their regulatory use. Discussions can usefully include the following:

1. *Inclusion of QSARs*: The fundamental point is that REACH requires the use of all available data, including predictions generated by QSAR models, and ECHA includes the use of QSARs in its basic guidance.

2. *Specificity of purpose*: REACH identifies different assessments to be carried out in order to register, classify or prioritise chemical substances, and those different regulatory tasks make different demands for evidence. Stakeholder discussions of the use of QSAR models should therefore take care to be specific about the regulatory purpose and requirements being addressed.

3. *Specificity of the model used*: As ECHA makes explicit in their case-by-case policy, the acceptance of the use of a QSAR model for one chemical and one regulatory function does not imply likely acceptance for another. Models have to be understood as being specific to particular endpoints and particular domains of applicability, and so care has to be taken when selecting models to use in a particular evaluation.

4. *Weight of evidence*: The focus on innovation and the use of a range of sources requires industry (or their consultants) to shift from using a single *in vivo* test towards using a range of complementary sources of evidence within a 'weight of evidence' approach.

5. *Explicit uncertainty*: All evaluations of chemical toxicity inevitably contain uncertainty, whether produced by *in vivo*, *in vitro* or *in silico* models. The recognition of this uncertainty, and the use of complementary models and other sources of information to address it,

is a vital step in any evaluation, whether for risk assessment or for other purposes.

6. *Transparency*: The regulatory acceptance of QSAR models requires clear documentation both of the model and of the way it has been used, sufficient to enable the regulator to conduct an independent review of the evidence.

As specific case studies become available, it will be possible to learn from them - while of course recognising that regulatory acceptance is explicitly *case-by-case*, so that the acceptance or rejection of these individual uses of QSAR models should not be over-generalised.

Future debate needs to clarify the acceptable uncertainty in data from QSAR models across the different regulatory functions. (Studies have been done on this [21].) Decisions involve both scientific and political judgements, so the debate should involve all stakeholders. Acceptable uncertainty will be specific to particular evaluations and chemicals, perhaps because of their level of toxicity or exposure levels or because of the proximity to the regulatory decision point or simply the risk of being wrong; it will be specific to particular endpoints, given their different levels of impact. Discussions of acceptable uncertainty need to acknowledge and approximate to the levels of uncertainty that are viewed as acceptable or inevitable within experimental values for that endpoint.

In QSAR modelling, the complexity of both the molecular structures and the biological activities under investigation is huge, and different approximations and assumptions are inevitably necessary. Clearly no single model or platform can provide all the predictions needed. Even for particular chemicals and endpoints, our preference is towards the use of multiple models within a strategy that is capable of utilising what each model can offer.

This requires a reshaping of the common procedure to evaluate chemical data, but it is exactly in line with the 'weight of evidence' approach advocated by ECHA and within REACH, where uncertainty is reduced by the complementary use of different models or methods.

In the near future a huge amount of data will become available, including the data produced by the Tox21 [19] and ToxCast [20] initiatives. To make use of this unprecedented amount of data, new tools are necessary. It is wise now to develop the debate on the suitable integration of QSAR models within a broader process of evaluation of chemical properties, especially within a 'weight of evidence' approach.

Acknowledgements and Funding
We acknowledge the financial contribution of the EC project ORCHESTRA (number 226521).

Author details
[1]Istituto di Ricerche Farmacologiche "Mario Negri", Via La Masa 19, 20156, Milano, Italy. [2]PublicSpace Ltd, Bletherbeck House, Ulverston, LA12 8DB, UK. [3]Department of Electronics and Information, Politecnico di Milano, Piazza L. da Vinci 32, 20133, Milano, Italy. [4]Symlog, 262 rue St Jacques, 75005, Paris, France. [5]CentroReach, Via G. da Procida, 11, 20149, Milano, Italy. [6]Interdisciplinary Research Unit on Risk Governance and Sustainable Technology Development, University of Stuttgart, Seidenstraße 36, 70174, Stuttgart, Germany.

Authors' contributions
EB prepared a first draft, which was discussed with all authors. SP and EB further developed the article in the light of external peer reviews.

Competing interests
The authors declare that they have no competing interests.

Received: 20 July 2011 Accepted: 7 October 2011
Published: 7 October 2011

References

1. ECHA: **Guidance in a nutshell: Registration data and dossier handling.** 2009, 14 & 19 [http://guidance.echa.europa.eu/docs/guidance_document/nutshell_guidance.pdf].
2. Benfenati E, Ed: *Quantitative structure-activity relationships (QSAR) for pesticides regulatory purposes* Amsterdam: Elsevier; 2007.
3. Regulation (EC) No 1907/2006: **REACH - Registration, Evaluation, Authorisation and Restriction of Chemicals.**[http://ec.europa.eu/enterprise/sectors/chemicals/reach/index_en.htm].
4. Roncaglioni A, Benfenati E: ***In silico*-aided prediction of biological properties of chemicals: oestrogen receptor-mediated effects.** *Chem Soc Rev* 2008, **37**:441-450.
5. ECHA: **Practical guide 5: How to report (Q)SAR.** 2009, 2-4[http://echa.europa.eu/doc/publications/practical_guides/pg_report_qsars.pdf].
6. AMBIT software:[http://ambit.sourceforge.net/].
7. Cassano A, Manganaro A, Martin T, Young Y, Piclin N, Pintore M, Bigoni D, Benfenati E: **The CAESAR models for developmental toxicity.** *Chemistry Central Journal* 2010, **4(Suppl 1)**:S4.
8. OECD: **Environment Directorate. Validation of (Q)SAR models.**[http://www.oecd.org/dataoecd/33/37/37849783.pdf].
9. ECHA: **Guidance on information requirements and chemical safety assessment: Chapter R.6: QSARs and grouping of chemicals.** 2008, 12-13 [http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r6_en.pdf].
10. ECHA: **Practical Guide 10. How to avoid unnecessary testing on animals.** 2010 [http://echa.europa.eu/doc/publications/practical_guides/pg_10_avoid_animal_testing_en.pdf].
11. Benfenati E: **The CAESAR project for *in silico* models for the REACH legislation.** *Chemistry Central Journal* 2010, **4(Suppl 1)**:11.
12. Lombardo A, Roncaglioni A, Boriani E, Milan C, Benfenati E: **Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish.** *Chemistry Central Journal* 2010, **4(Suppl 1)**:S1.
13. Ferrari T, Gini G: **An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts.** *Chemistry Central Journal* 2010, **4(Suppl 1)**:S2.
14. Fjodorova N, Vracko M, Novic M, Roncaglioni A, Benfenati E: **New public QSAR model for carcinogenicity.** *Chemistry Central Journal* 2010, **4(Suppl 1)**:S3.
15. Chaudhry Q, Piclin N, Cotterill J, Pintore M, Price NR, Chrétien JR, Roncaglioni A: **Global QSAR models of skin sensitisers for regulatory purposes.** *Chemistry Central Journal* 2010, **4(Suppl 1)**:S5.
16. Zhu H, Martin TM, Young DM, Tropsha A: **Combinatorial QSAR Modeling of Rat Acute Toxicity by Oral Exposure.** *Chem Res Toxicol* 2009, **22**:1913-1921.
17. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV: **Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis.** *J Chem Inf Model* 2008, **48**:766-784.
18. Sushko I, Novotarskyi S, Körner R, Pandey AK, Cherkasov A, Li J, Gramatica P, Hansen K, Schroeter T, Müller KR, Xi L, Liu H, Yao X, Öberg T, Hormozdiari F, Dao P, Sahinalp C, Todeschini R, Polishchuk P, Artemenko A, Kuz'min V, Martin TM, Young DM, Fourches D, Muratov E, Tropsha A, Baskin I, Horvath D, Marcou G, Muller C, Varnek A, Prokopenko VV, Tetko IV: **Applicability domains for classification problems: benchmarking of**

distance to models for AMES mutagenicity set. *J Chem Inf Model* 2010, **50**:2094-2111.
19. **Tox21 Research Program.** [http://www.epa.gov/ncct/Tox21/].
20. **ToxCast™: Screening Chemicals to Predict Toxicity Faster and Better.** [http://www.epa.gov/ncct/toxcast/].
21. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P: **Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs.** *Environ Health Perspect* 2003, **111**:1361-1375.

doi:10.1186/1752-153X-5-58
Cite this article as: Benfenati *et al.*: The acceptance of *in silico* models for REACH: Requirements, barriers, and perspectives. *Chemistry Central Journal* 2011 **5**:58.

Publish with ChemistryCentral and every scientist can read your work free of charge

"Open access provides opportunities to our colleagues in other parts of the globe, by allowing anyone to view the content free of charge."

W. Jeffery Hurst, The Hershey Company.

- available free of charge to the entire scientific community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.chemistrycentral.com/manuscript/

ChemistryCentral