

## Results of a round-robin exercise on read-across

E. Benfenati, M. Belli, T. Borges, E. Casimiro, J. Cester, A. Fernandez, G. Gini, M. Honma, M. Kinzl, R. Knauf, A. Manganaro, E. Mombelli, M. I. Petoumenou, M. Paparella, P. Paris & G. Raitano

**To cite this article:** E. Benfenati, M. Belli, T. Borges, E. Casimiro, J. Cester, A. Fernandez, G. Gini, M. Honma, M. Kinzl, R. Knauf, A. Manganaro, E. Mombelli, M. I. Petoumenou, M. Paparella, P. Paris & G. Raitano (2016): Results of a round-robin exercise on read-across, SAR and QSAR in Environmental Research, DOI: [10.1080/1062936X.2016.1178171](https://doi.org/10.1080/1062936X.2016.1178171)

**To link to this article:** <http://dx.doi.org/10.1080/1062936X.2016.1178171>



View supplementary material [↗](#)



Published online: 11 May 2016.



Submit your article to this journal [↗](#)





View related articles [↗](#)



View Crossmark data [↗](#)

## Results of a round-robin exercise on read-across

E. Benfenati<sup>a</sup>, M. Belli<sup>a</sup>, T. Borges<sup>b</sup>, E. Casimiro<sup>c</sup>, J. Cester<sup>d</sup>, A. Fernandez<sup>d</sup> , G. Gini<sup>e</sup>   
M. Honma<sup>f</sup>, M. Kinzl<sup>g</sup>, R. Knauf<sup>h</sup>, A. Manganaro<sup>i</sup>, E. Mombelli<sup>j</sup>, M. I. Petoumenou<sup>a</sup>, M.  
Paparella<sup>g</sup>, P. Paris<sup>k</sup> and G. Raitano<sup>a</sup>

<sup>a</sup>IRCCS – Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy; <sup>b</sup>Direcção-Geral da Saúde, Lisboa, Portugal; <sup>c</sup>INFOTOX, Consultores de Riscos Ambientais e Tecnológicos, Lda, Lisboa, Portugal; <sup>d</sup>Universitat Rovira i Virgili, Tarragona, Spain; <sup>e</sup>Politecnico di Milano, Dipartimento di Elettronica e Informazione, Milan, Italy; <sup>f</sup>Division of Genetics and Mutagenesis, National Institute of Health Sciences, Tokyo, Japan; <sup>g</sup>Umweltbundesamt GmbH, Vienna, Austria; <sup>h</sup>Centro REACH S.r.l., Milan, Italy; <sup>i</sup>Kode S.r.l. Pisa, Italy; <sup>j</sup>Institut National de l'Environnement Industriel et des Risques, Verneuil-en-Halatte, France; <sup>k</sup>Istituto Superiore per la Protezione e la Ricerca Ambientale, Rome, Italy

### ABSTRACT

A round-robin exercise was conducted within the CALEIDOS LIFE project. The participants were invited to assess the hazard posed by a substance, applying *in silico* methods and read-across approaches. The exercise was based on three endpoints: mutagenicity, bioconcentration factor and fish acute toxicity. Nine chemicals were assigned for each endpoint and the participants were invited to complete a specific questionnaire communicating their conclusions. The interesting aspect of this exercise is the justification behind the answers more than the final prediction in itself. Which tools were used? How did the approach selected affect the final answer?

### ARTICLE HISTORY

Received 18 March 2016  
Accepted 11 April 2016

### KEYWORDS

Read-across; mutagenicity;  
bioconcentration factor

## Introduction

Read-across is the non-testing method mainly used by registrants to comply with REACH regulation [1,2]. The rationale behind this alternative approach is based on the perception that substances with similar physico-chemical structures will have similar (eco) toxicological properties [3]. For 75% of the chemicals, it has been used to fill data gaps for registered substances as an additional source of data within a weight-of-evidence approach [4]. It is likely to be used more in the future, because the substances to be registered under REACH by the next deadline, in 2018, have fewer experimental data than the higher-tonnage substances registered in the past. It is also expected that the REACH 2018 registration process will involve more small and medium enterprises (SMEs), which have fewer resources to generate testing data.

The European Chemicals Agency (ECHA) recently published a document containing a Read-Across Assessment Framework (RAAF), to harmonize the evaluation of read-across documents, and also to provide initial guidance for users [5]. There are many possible ways

**CONTACT** E. Benfenati  [emilio.benfenati@marionegri.it](mailto:emilio.benfenati@marionegri.it)

 Supplemental data for this article can be accessed [here](#).

© 2016 Informa UK Limited, trading as Taylor & Francis Group

to a solution on the basis of a read-across approach, and different results may well be obtained for the same substance. The guidance documents [6,7] contain the keystones of this method, while recent studies concerning its validity indicate that scientific justification has to be solid and thoroughly documented [8].

There have been several exercises on the results of quantitative structure–activity relationships (QSAR) for different endpoints [9–18]; also, some EC-funded projects – ANTARES ([www.antes-life.eu](http://www.antes-life.eu)), CALEIDOS ([www.caleidos-life.eu](http://www.caleidos-life.eu)), and PROSIL ([www.life-prosil.eu](http://www.life-prosil.eu)) – looked into the use of QSAR models for REACH compliance by comparing different models for a number of properties, such as carcinogenicity, mutagenicity, log *P*, fish acute toxicity, bioconcentration factor (BCF) and developmental toxicity. However, to the best of our knowledge there are still no articles assessing the property values obtained according to the read-across approach. Predictions using QSAR models may be more straightforward than robust read-across exercises depending, of course, on the endpoint. Read-across normally concerns a single substance assessment aiming to fulfil the information requirements of the REACH registration.

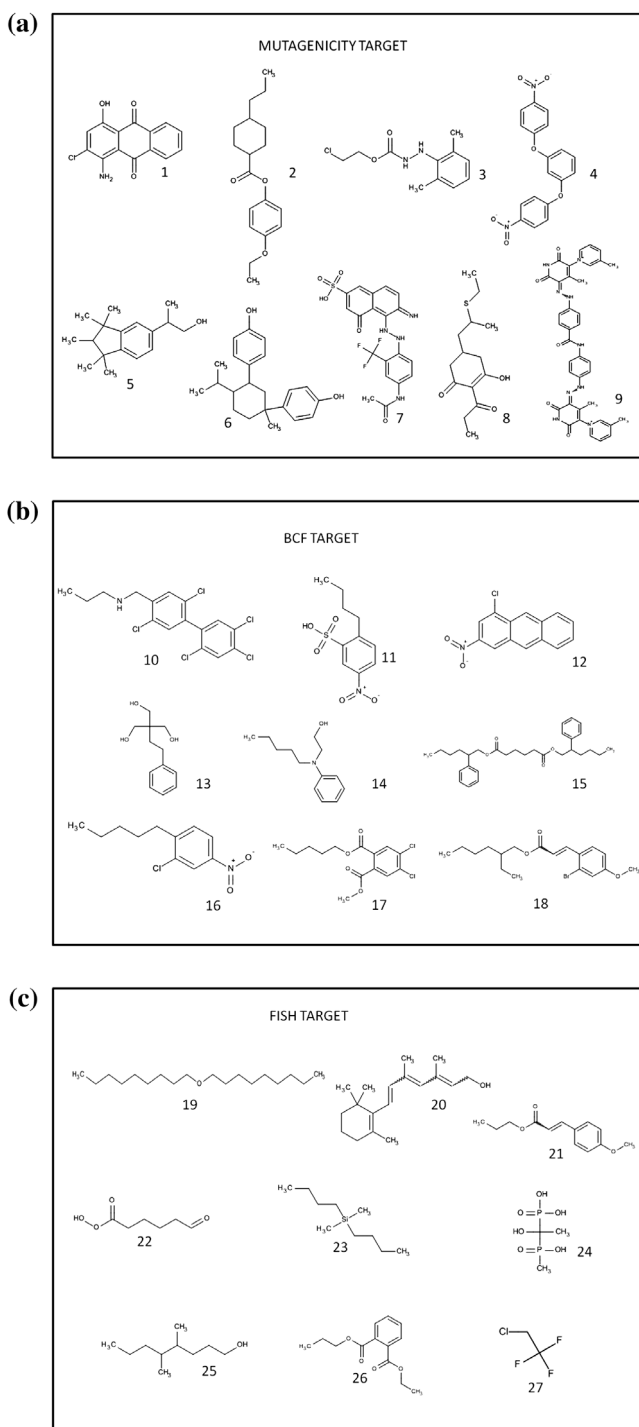
Within the CALEIDOS LIFE project a round-robin exercise was organized among 40 participants who were asked to evaluate 27 substances, nine for each of the following properties: mutagenicity (Ames test), BCF, and fish acute toxicity. The aim of the exercise was to evaluate the reproducibility of the results. We also took into account the approach used and the similar substances that were adopted as structural analogues. The results of this exercise are described here.

## Material and methods

### *Selected substances*

The general principle was that participants should not make use of the property value found in the literature or on the Internet to assess the target substances. Thus, we looked for substances without known values. For mutagenicity we relied on the exercise organized by Health Japan on QSAR predictions, which made 4000 substances available, mostly with unpublished property values. Nine substances were selected based on the consistency of the predicted results with different QSAR models (VEGA CAESAR, VEGA ISS, VEGA SARPy, T.E.S.T and Toxtree 2.6.6) ([www.vega-qsar.eu](http://www.vega-qsar.eu), [www.epa.gov](http://www.epa.gov), <http://toxtree.sourceforge.net>): three substances were predicted consistently, three gave conflicting results, and three gave results with uncertainty for some models, considering the applicability domain. Some of the selected substances were predicted as mutagenic and others as non-mutagenic. The experimental values were not known when we organized the exercise and the property values were revealed by Health Japan only at the end of the exercise.

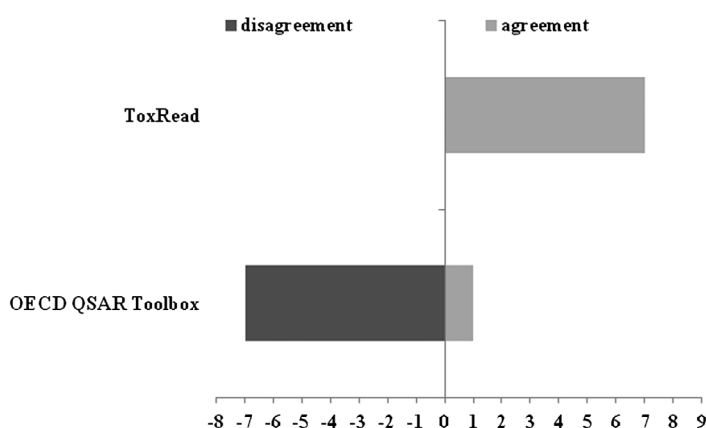
For BCF and fish acute toxicity, nine substances were selected for each property, without any experimental value. These substances were selected using the criteria explained above: three substances produced quite consistent results with QSAR models, apparently with reliable predictions, three were more “difficult”, and three were “borderline”, with some uncertainty concerning the predicted values. The QSAR models used for BCF assessment were the CAESAR, Meylan and kNN models within VEGA, and for fish acute toxicity SARPy, kNN also present in VEGA ([www.vega-qsar.eu](http://www.vega-qsar.eu)) and the model from T.E.S.T ([www.epa.gov](http://www.epa.gov)). Figure 1 (a–c) shows the 27 substances used.



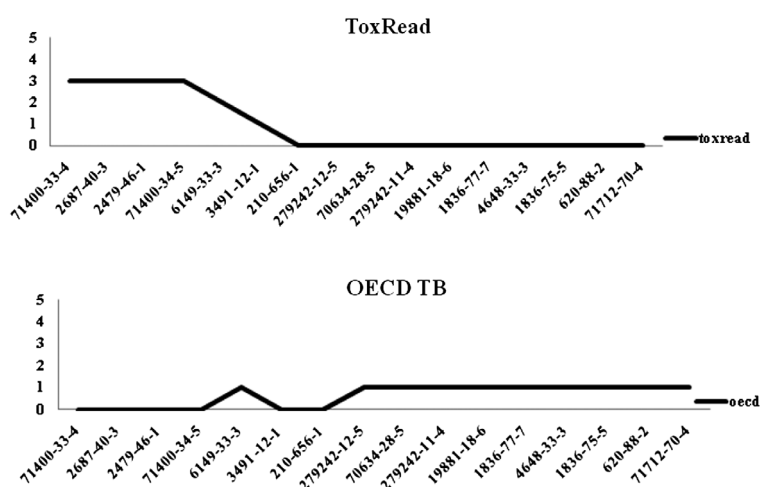
**Figure 1.** Chemicals used for the read-across exercise. (a) Chemicals 1 to 9 were used for the assessment of mutagenicity. (b) Chemicals 10 to 18 were used for BCF. (c) Chemicals 19 to 27 were used for fish acute toxicity.

## The online questionnaire

To facilitate collection of information from the participants doing the exercise, an online questionnaire was developed. The participant had to choose the property (one or more), and for each property a substance was shown. The participant had then to submit the property value, indicating the method adopted, the level of uncertainty in the evaluation, the software used, if any, its usefulness and simplicity, and the similar substances selected as the basis for the assessment. Further information was requested from the participants regarding the approximate time needed to complete the assessment, their level of experience and their occupational sector. Mutagenicity provided a categorical assessment (mutagenic, non-mutagenic, not sure), and a continuous value was needed for the other two properties, BCF and acute fish toxicity. Figures 1, 2, and 3 of the Supplementary material show the



**Figure 2.** Agreement among participants in the evaluation of the mutagenicity, depending on the software used.



**Figure 3.** Reproducibility in identifying the most similar substances in the mutagenicity assessment of the target chemical (target molecule number 2 in Figure 1. (a)).

questions (available via the Supplementary Content tab on the article's online page). The output file could be uploaded to support the assessment. Once the substance was evaluated, users were asked whether they wanted to proceed to another substance. Results were made anonymous.

## Results

In all, 181 questionnaires from 40 participants were submitted. The majority did the exercise for mutagenicity (93), while 47 questionnaires were for BCF and 41 for fish acute toxicity. The results are described below for each individual property.

### Mutagenicity

#### Method and results

Table 1 shows the replies for the nine substances. All (or most of) the participants considered substances 1, 2, 3, and 7 mutagenic. For three substances (4, 5, and 6) the replies mostly indicated them as non-mutagenic (30/31), while for the other two (8, 9) there was no clear assignment. Thus it can be concluded that there was a considerable agreement among the participants for the positive and negative cases.

Table 2 shows the assessment tools used (one user did the assessment manually, without any software), and the users' sectors. Some participants used only a single software package, others more than one.

ToxRead ([www.toxgate.eu](http://www.toxgate.eu)) was the software used most (58 participants), often in combination with other programs. ToxRead has been described elsewhere [19–21]. For the cases

**Table 1.** The replies provided by the participants for the mutagenicity assessment of the nine substances used for the mutagenicity exercise.

Chemical ID	Mutagenic	Non mutagenic	Not sure
1	9	2	
2	8		1
3	6		1
4		9	
5		11	
6		10	1
7	8	1	3
8	3	6	2
9	6	4	2

**Table 2.** The programs used by the participants for mutagenicity, and the stakeholder sector of the participants.

	OECD QSAR Toolbox	VEGA platform	Toxread software	T.E.S.T.	Toxtree	ChemID Plus + eChemportal	Leadscope
Tot	32	41	58	6	14	1	1
Unique software	20	2	18	2		1	
Academic	3	36	40	6	14		1
Consultancy	10	4	5				
Industry	1	1	9			1	
Regulatory	18		4				

where only one program was used, the OECD QSAR Toolbox ([www.qsartoolbox.org](http://www.qsartoolbox.org)) was the most used. Regulators and consultants used the OECD QSAR Toolbox as single tool in most cases, possibly because it has already been developed under the auspices of the OECD, and users have acquired some experience as a result of training events. There may be an additional reason for this choice: the OECD QSAR Toolbox takes quite a long time, more than in the other methods, according to the participants' replies, and this may leave less time available to explore other methods further. Most of the participants using ToxRead also applied a second program, usually VEGA ([www.vega-qsar.eu](http://www.vega-qsar.eu)), showing that they were looking for confirmation, or further evidence for the final assessment.

Figure 2 shows the agreement in the evaluation, in relation to the software used. We show the results for participants using only one program. ToxRead was applied for seven substances by more than one participant and in all the cases there was agreement in the outcome of the evaluation.

Does ToxRead improve the concordance of the answers given by different experts? Given the very limited number of completed questionnaires no full statistical analysis can be undertaken to confirm this. However, considering the mutagenicity example, where answers are categorical, we can do a simple statistical test. We used data about the concordance on the expert evaluations for each of the nine molecules. We computed the  $p$ -value with Student's  $t$  to reject the zero hypothesis that "the concordances between assessments done using only ToxRead and using anything else are equivalent". We set the  $p$ -value of 0.05 as the threshold to reject the zero hypothesis. We generated a first series of the ratio of concordance over answers for using only ToxRead and a second series for answers given using anything else (including ToxRead in a mix). The resulting  $p$ -value is 0.014.

We also tested the concordance in using only ToxRead against not using ToxRead at all. The resulting  $p$ -value is 0.031. Those results show that the zero hypothesis could be rejected.

The results for participants using the OECD QSAR Toolbox were the opposite. For eight substances with replies from more than one participant, all the cases except one showed disagreement among the assessments. This finding is surprising, and was analysed further.

We considered the different approaches adopted by the 20 participants who used the OECD QSAR Toolbox as single program. Eight participants used structural similarity as the driving strategy, and seven did not describe the method clearly or did not report it at all. Four participants used structural alerts in addition to similarity, and the last used structural similarity in addition to visual inspection.

We also checked whether the different participants based their assignment on the same set of similar chemicals. Figure 3 lists the similar chemicals used for the assignment of chemical number 2 in Figure 1; the other substances gave similar results. The participants using ToxRead repeatedly listed the same set of most similar compounds, whereas the participants using the OECD QSAR Toolbox did not agree on the chemicals chosen; thus, each participant based the evaluation using different sets of compounds. On the basis of these observations, the likely reason for the very different levels of agreement between the participants using ToxRead and the OECD QSAR Toolbox was that at an early stage of the assessment, ToxRead shows all the possible causes of effect (or lack of effect) in the same picture. Conversely, the OECD QSAR Toolbox provides a complementary approach. The user has to decide which profiler should be applied, and this leaves many possibilities open which may result in a different selection of structural analogues, possibly leading to contradictory conclusions.

**Table 3.** The correctness of the results, for the different sectors of the participants.

		Sector			
		Academic	Consultancy	Industry	Regulatory
Assessment	Correct	26	7	8	7
	Incorrect	23	3	3	6
	Not sure	3	0	0	7
	Correct rate	0.5	0.7	0.73	0.35

### *The correctness of the evaluations*

As already mentioned, for mutagenicity we also had the experimental values, given by Health Japan at the end of the exercise. Only chemical number 3 in Table 1 was mutagenic. It was predicted correctly by all participants using ToxRead alone or together with other tools (the one participant using the OECD QSAR Toolbox wrote (s)he was not sure about the effect). However, obviously, we had several false positives (though no false negatives). The proportion of wrong assignment was about 42%, due to false positives. This indicates that the evaluation was too conservative.

Table 3 shows the correctness of the results according to the participant's sector. Participants from industry and consultancy gave the highest rate of correct replies. On the other hand, regulators had the highest rate of "not sure" answers. In this case, seven out of 10 were using the OECD QSAR Toolbox alone, or combined with other methods in two cases. The three other participants who replied "not sure" were from academia. This indicates that regulators are more cautious in their assessment. This may reflect their institutional duty that positions them as the ultimate gatekeeper against false negatives.

### *Predicting genotoxicity*

We examined whether the participant's sector influenced the assignment of the mutagenicity score. In six cases out of nine regulators had a tendency to assign the chemicals as mutagenic, while in the consultants sector eight out of nine were identified as non-mutagenic.

### *Software simplicity*

Participants were asked to evaluate the ease of use of the software and its simplicity (Figure 4). Programs such as ToxRead and VEGA were judged very user-friendly by most of the participants, while the OECD QSAR Toolbox was considered more difficult to handle. The answer may be biased by the evaluator's experience with the different tools.

### *Confidence in the assessment results*

We asked the participants to indicate their level of confidence in the assessments (Table 4). There are clear differences depending on the sector. Again, regulators were particularly prudent in their conclusion on mutagenicity, whereas the other participants tended to declare a high level of certainty, particularly high for participants from industry.

### *Correctness of the assessment in relation to the declared experience*

We asked participants to declare their experience in read-across. Table 5 compares their statements with the correctness of the assessment. Self-declared experts did not say they were not sure, but they also made mistakes. We also asked participants to declare how certain they were in their assessment (Table 6). Self-evaluation on the level of certainty was not



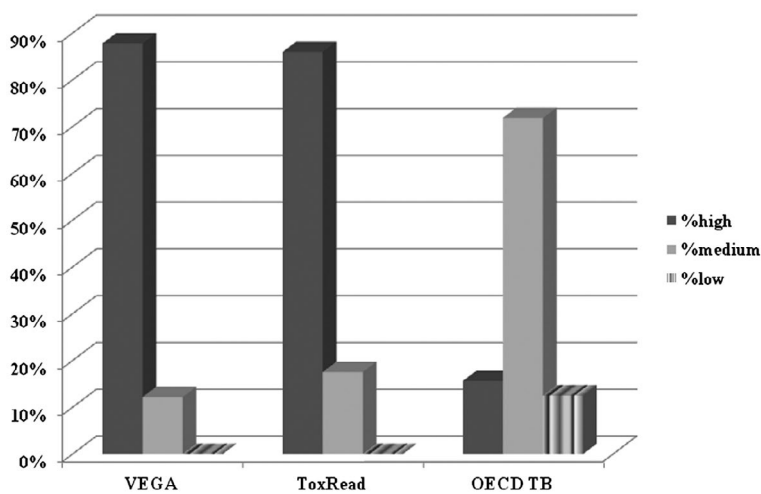


Figure 4. The simplicity of the software.

Table 4. The level of confidence for the different sectors of the participants.

		Sector			
		Academic	Consultancy	Industry	Regulatory
Certainty	High	34	4	8	4
	Medium	10	2	2	5
	Low	8	4	1	11

Table 5. The correctness of the results, for the declared experience on read-across.

		Experience		
		Expert	Familiar	Unfamiliar
Assessment	Correct	10	15	23
	Incorrect	6	10	19
	Not Sure	0	7	3

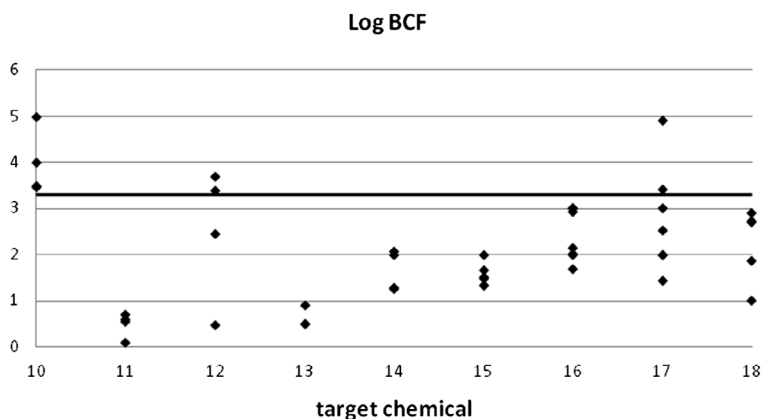
Table 6. The correctness of the results, for the declared certainty of the assessment.

		Certainty		
		High	Medium	Low
Assessment	Correct	31	9	8
	Incorrect	18	9	8
	Not Sure	1	1	8

sufficient to avoid errors, even though certainty was to some extent related to more correct answers: 31 correct answers versus 18 incorrect ones.

BCF

The substances selected for BCF were not associated with experimental values so could not be verified. Therefore the evaluation was based on the level of agreement among the



**Figure 5.** Agreement among participants in the evaluation of the BCF (Y axis) for each target chemical.

different participants. In the case of BCF each molecule had a minimum of three assessments (molecule 13) and maximum of eight (molecule 16). Most of the participants also provided information on the similar compounds they used for the assessment of the target compound.

For this endpoint the results showed a much higher level of agreement than for mutagenicity (see Figure 5). This probably reflects the nature of the endpoint, which allows a scale of continuous values, whereas for genotoxicity the assessment is strictly binary, so conflicts are more visible. Disagreement was apparent only in three cases. Molecule 12 was assigned as non-bioaccumulative using the OECD QSAR Toolbox by one participant, while a second participant, using the same program, assigned the chemical as very bioaccumulative.

One participant for chemical 14 declared it was impossible to reach a conclusion using the OECD QSAR Toolbox. However, another participant, using the same program, concluded that the chemical was non-bioaccumulative.

For chemical 17 two participants produced conflicting results assigning logBCF values of 4.9 and 2 using ToxRead. This remained with any clear explanation, suggesting erroneous assessments, because the first participant identified substances with much lower values as chemicals used for read-across.

Still, BCF assessment participants stated that the ToxRead and VEGA programs were easier to use than the OECD QSAR Toolbox. Regulators used only the OECD QSAR Toolbox for the assessment, while academics and industries preferred to use more than one software package. However, due to the small number of participants on the BCF exercise, it is hard to assess the implications of these parameters in the outcome of the replies.

### **Fish acute toxicity**

In the fish acute toxicity assessment, the substances selected did not have experimental data for comparison so the evaluation could only be based on the consistency of the replies.

In the case of this endpoint, no module is currently available within ToxRead. For fish acute toxicity the participants mostly used the models from OECD QSAR Toolbox and VEGA platforms.

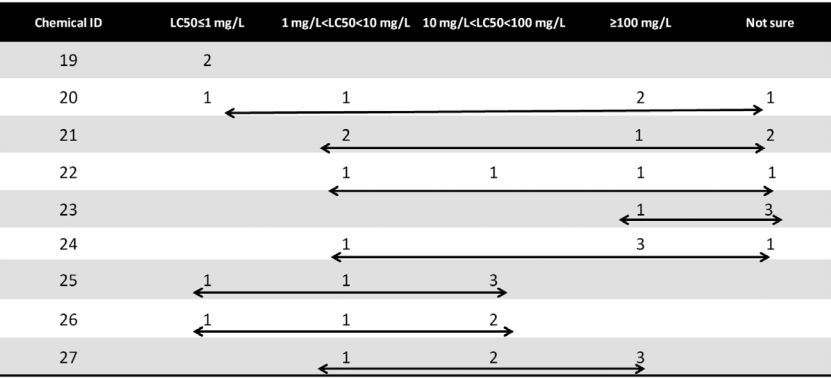


Figure 6. The spread of fish acute toxicity values assessed by the different participants.

In contrast to the BCF predictions, here there were more disagreements even when the same software was used, the values often being spread over several orders of magnitude (Figure 6). This can be explained at least partially by the absence of similar compounds related to the target compounds. For molecule 20 there were various assignments, from non-toxic to toxic. In this case those who were not familiar with the software had more conservative answers. A peculiar case was molecule 21, where there were three different assessments using the same software (OECD QSAR Toolbox). The assessments were the most consistent for molecule 24.

We highlight the fact that some participants reported continuous values in mg/L and others stated only whether the substance was toxic or not.

Discussion

Read-across approaches will be more and more widespread in the future, to evaluate industrial chemicals, cosmetics, pesticides, etc. But this is *ad hoc* modelling that is, by its nature, not as well formalized as other evaluation approaches such as prediction models for *in vitro* methods, and QSARs, which are developed on a large database with careful data-pruning and formal statistical approaches. Therefore read-across methods are more prone to a subjective bias and poor reproducibility. *Ad hoc* data inclusion or exclusion, and *ad hoc* weights assigned to the data contribute to this. This is particularly the case for complex endpoints where expertise matters considerably for the reliability of the assessment [20].

We organized the exercise described to ascertain the level of agreement among different assessors and the basis for their assessment. We are grateful to all participants who dedicated their time to this exercise. The purpose was not to select a winner, which is why the replies were anonymous. The majority of participants used programs for read-across like ToxRead, OECD QSAR Toolbox and VEGA. Other tools were also used, but only to obtain additional data in most cases.

This exercise confirmed that there is a certain disagreement among assessors, varying on the basis of the endpoint and the program used. Some endpoints are more consistently evaluated among experts, and BCF seems to be one of them, perhaps because it is easier to assess. In the well-known data collections most substances have values under 3.3 [11]. This

makes common agreement on “not-bioaccumulative” more likely. The availability of a recognized descriptor such as the partition coefficient between octanol and water ( $\log K_{ow}$ ) for BCF also improves the chance of more uniform predictions between experts.

In the case of mutagenicity we noticed wider disagreement, particularly when participants used the OECD QSAR Toolbox. This may be due to the fact that this evaluation is not based on a quantitative value, like BCF. For BCF the only grey area is close to the threshold of 3.3, while assessments far from the threshold may be more likely to be judged as belonging to one or the other category. This does not apply to mutagenicity, because the evaluation is binary (either positive or negative). However, for mutagenicity there are quite large collections of chemicals (many thousands) with experimental values, which may contribute to closer concordance among the QSAR models and greater accuracy in the predictions, the latter being in the range of experimental variability [13,16]. Thus, *ad hoc* modelling by read-across may be less reliable than QSARs.

A robust comparison cannot be made on the basis of this read-across exercise due to the limited number of chemicals and the larger number of non-genotoxic substances, but the data generated are consistent with this assumption.

In the case of fish acute toxicity there was ample discrepancy among the assessments, larger than for mutagenicity and BCF. This endpoint was difficult to predict also using QSAR models [9]. Although there are collections of data for this endpoint gathering hundreds of chemicals, the results are spread among different fish species and protocol variants, resulting in high experimental data variability and the need to decide on building *ad hoc* read-across approaches or QSAR models for more homogeneous protocol variants with fewer data or less homogeneous variants with more data. In terms of environmental protection goals it is hard to say which approach is more favourable, since it cannot be decided *a priori* which specific protocol variant may be more protective than another for a given substance. Furthermore, the toxicity endpoint is more complex than the BCF endpoint, since a wide range of toxic modes of action may lead to lethality in fish, which complicates the development of *ad hoc* read-across approaches or QSARs. These facts may explain that wider deviations of QSAR results and *ad hoc* read-across approaches should be acceptable, since experimental variability is also higher, while also at least partially explaining the poor reproducibility of the read-across exercise.

Besides the specificities related to the endpoint assessment, the use of different read-across software packages with different levels of difficulty is another reason for disagreement in the results. The OECD QSAR Toolbox is a powerful, versatile program, covering a large set of possible toxicity pathways; however, this probably produces a complex palette of alternative strategies and consequently variable read-across results, even from participants apparently familiar with the tool, such as regulators.

The ToxRead software was developed more recently, using a graphic display of the prioritized, similar chemicals, together with the reasons for concern expressed as structural alerts and rules based on continuous parameters [19,20]. Participants using this program found it user-friendly and were able to achieve reproducible results.

VEGA is another program that provides QSAR predictions while showing the closest structural analogues together with their experimental and modelled results ([www.vega-qsar.eu](http://www.vega-qsar.eu)). Thus, this feature can be used for read-across, and was applied by some participants, though it is already a combination of QSAR and read-across approaches for local validation of the result. Other programs for read-across exist, but they were not used by these participants.

The number of chemicals and participants in the exercise was not large enough to permit any general conclusions. Nevertheless, there was a large proportion of false positive predictions for mutagenicity. This bias is probably favoured by the large number of negatives among the selected chemicals, but might be symptomatic of an underlying cognitive behaviour that tends to avoid the risk of false negative predictions. QSAR models may be tuned towards higher sensitivity or specificity than read-across.

This exercise on read-across highlighted the tendency for some evaluators to be over-conservative in their assessment, and the risk of over-optimistic self-evaluation of certainty about a prediction. This risk is hard for the assessor to acknowledge, since one's own ability is of course evaluated individually. This may indeed compromise the outcome of the evaluation and should further promote the development of tools based on objective, reproducible representation of the different critical parameters underlying the assessment.

This exercise also showed that there are differences in the assessment depending on the occupational sector. Regulators have the inherent tendency to require more convincing proof of the effect before concluding. Conversely, other sectors tended to reach faster conclusions. Clearly, different behaviours, experience, and institutional duties have a role in the outcome. Regulators are probably trained to proceed according to the robustness of the evidence before taking a decision, and this increases their requirement for convincing data at the basis of the read-across, while other sectors, such as academia, may have faced the exercise without regulatory restraints.

## Conclusions

This exercise showed that there are large areas of uncertainty in read-across evaluations. BCF was assessed more consistently than acute fish toxicity, indicating that reproducibility from read-across approaches is likely to be endpoint dependent. It is probably influenced by the availability and heterogeneity of experimental data, the availability of recognized molecular descriptors, such as  $\log K_{ow}$  for BCF, and the balance of positive versus negative substances in the available databases. Most participants used programs to assist their analysis. However, this did not automatically generate more reproducible results, reproducibility being related to the program used.

Programs differ in their simplicity and their features to support harmonized decisions. Regulators tend to consider the conclusions from read-across evaluation less certain than industry and academic experts.

There is clearly a need to proceed towards more reproducible assessments to obtain robust read-across arguments in the future, in support of REACH compliance. This requires tools that facilitate the overall judgement of the assessor, presenting the evidence underlying the assessment in a clear, reproducible way. More documentation, reporting and examples of the read-across technique making use of software tools are also recommended in order to identify performance and limitations. The further information present in the read-across data needs to be processed and shown by future programs, in order to support experts in their overall conclusions and reduce the differences in their judgement.

## Acknowledgements

We are grateful to all participants of the questionnaire. We acknowledge the financial support of the EC, LIFE project CALEIDOS.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

A. Fernandez  <http://orcid.org/0000-0002-1241-1646>

G. Gini  <http://orcid.org/0000-0002-3876-8360>

## References

- [1] European Union, *Regulation (EC) No 1907/2006 of the European Parliament and of the Council, of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission.*
- [2] ECHA, *Technical Guidance Documents for the Implementation of REACH. Guidance on Information Requirements and Chemical Safety Assessment*, European Chemical Agency, Helsinki, Finland, 2008.
- [3] T.W. Schultz, P. Amcoff, E. Berggen, F. Gautier, M. Klaric, D.J. Knight, C. Mahony, M. Schwarz, A. White, and M.T.D. Cronin, *A strategy for constructing and reporting a read-across prediction of toxicity*, Regul. Toxicol. Pharmacol. 72 (2015), pp. 586–601.
- [4] ECHA, *The use of alternatives to testing on animals for the REACH Regulation - Second report under Article 117(3) of the REACH Regulation*, ECHA-14-A-07-EN. European Chemical Agency, Helsinki, Finland, 2014. Available at [http://echa.europa.eu/documents/10162/13639/alternatives\\_test\\_animals\\_2014\\_en.pdf](http://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2014_en.pdf)
- [5] ECHA, *Read-Across Assessment Framework (RAAF)*, ECHA/PR/15/07, European Chemical Agency, Helsinki, Finland, 2015.
- [6] OECD, *Guidance on grouping of chemicals. OECD Series on Testing and Assessment No. 80*, Organisation for Economic Co-operation and Development, Paris, France, 2007.
- [7] OECD, *Guidance on grouping of chemicals. OECD Series on Testing and Assessment No. 194*, Organisation for Economic Co-operation and Development, Paris, France, 2014.
- [8] N. Ball, M. Bartels, R. Budinsky, J. Klapacz, S. Hays, and C. Kirman, *The challenge of using read-across within the EU REACH regulatory framework; how much uncertainty is too much? Dipropylene glycol methyl ether acetate, an exemplary case study*, Regul. Toxicol. Pharmacol. 68 (2014), pp. 212–221.
- [9] C.I. Cappelli, A. Cassano, A. Golbamaki, Y. Moggio, A. Lombardo, M. Colafranceschi, and E. Benfenati, *Assessment of in silico models for acute aquatic toxicity towards fish under REACH legislation*, SAR QSAR Environ. Res. 26 (2015), pp. 977–999.
- [10] C.I. Cappelli, E. Benfenati, and J. Cester, *Evaluation of QSAR models for predicting the partition coefficient (log P) of chemicals under the REACH regulation*, Environ. Res. 143 (2015), pp. 26–32.
- [11] C.I. Cappelli, S. Manganelli, A. Lombardo, A. Gissi, and E. Benfenati, *Validation of quantitative structure-activity relationships models to predict water-solubility of organic compounds*, Sci. Total Environ. 463–464 (2013), pp. 781–789.
- [12] M.I. Petoumenou, F. Pizzo, J. Cester, A. Fernández, and E. Benfenati, *Comparison between bioconcentration factor (BCF) data provided by industry to the European Chemicals Agency (ECHA) and data derived from QSAR models*, Environ. Res. 142 (2015), pp. 529–534.
- [13] A. Cassano, G. Raitano, E. Mombelli, A. Fernández, J. Cester, A. Roncaglioni, and E. Benfenati, *Evaluation of QSAR models for the prediction of Ames genotoxicity: A retrospective exercise on the chemical substances registered under the EU REACH regulation*, J. Environ. Sci. Health C, Environ. Carcinog. Ecotoxicol. Rev. 32 (2014), pp. 273–298.
- [14] A. Gissi, A. Lombardo, A. Roncaglioni, D. Gadaleta, G.F. Mangiatordi, O. Nicolotti, and E. Benfenati, *Evaluation and comparison of benchmark QSAR models to predict a relevant REACH endpoint: The bioconcentration factor (BCF)*, Environ. Res. 137 (2015), pp. 398–409.

- [15] R. Gonella Diaza, S. Manganelli, A. Esposito, A. Roncaglioni, A. Manganaro, and E. Benfenati, *Comparison of in silico tools for evaluating rat oral acute toxicity*, SAR QSAR Environ. Res. 26 (2015), pp. 1–27.
- [16] N. Golbamaki Bakhtyari, G. Raitano, E. Benfenati, and T.M. Martin, and D. Young, *Comparison of in silico models for prediction of mutagenicity*, J. Environ. Sci. Health. C, Environ. Carcinog. Ecotoxicol. Rev. 31 (2013), pp. 45–66.
- [17] C. Milan, O. Schifanella, A. Roncaglioni, and E. Benfenati, *Comparison and possible use of in silico tools for carcinogenicity within REACH legislation*, J. Environ. Sci. Health. C. 29 (2011), pp. 300–323.
- [18] F. Pizzo, A. Lombardo, A. Manganaro, and E. Benfenati, *In silico models for predicting ready biodegradability under REACH: A comparative study*, Sci. Total Environ. 463–464 (2013), pp. 161–168.
- [19] G. Gini, A.M. Franchi., A. Manganaro, A. Golbamaki, and E. Benfenati, *ToxRead: A tool to assist in read-across and its use to assess mutagenicity of chemicals*, SAR QSAR Environ. Res. 25 (2014), pp. 999–1011.
- [20] E. Benfenati, A. Roncaglioni, M.I. Petoumenou, C.I. Cappelli, and G. Gini, *Integrating QSAR and read-across for environmental assessment*, SAR QSAR Environ. Res. 26 (2015), pp. 605–618.
- [21] E. Benfenati, S. Manganelli, S. Giordano, G. Raitano, and A. Manganaro, *Hierarchical rules for read-across and in silico models of mutagenicity*, J. Environ. Sci. Health C, Environ. Carcinog. Ecotoxicol. Rev. 33 (2015), pp. 385–403.