

Implementation of *in silico* toxicology protocols within a visual and interactive hazard assessment platform

Glenn J. Myatt^{a,*}, Arianna Bassan^b, Dave Bower^a, Candice Johnson^a, Scott Miller^a,
Manuela Pavan^b, Kevin P. Cross^a

^a Instem, 1393 Dublin Road, Columbus, OH 43215, USA

^b Innovatune, Via Giulio Zanon 130/D, 35129 Padova, Italy

ARTICLE INFO

Edited by Dr. Mark Cronin.

Keywords:

In silico toxicology
Visual framework
ICH M7
Pharmaceutical impurities
Genetic toxicology
Skin sensitization

ABSTRACT

Mechanistically-driven alternative approaches to hazard assessment invariably require a battery of tests, including both *in silico* models and experimental data. The decision-making process, from selection of the methods to combining the information based on the weight-of-evidence, is ideally described in published guidelines or protocols. This ensures that the application of such approaches is defensible to reviewers within regulatory agencies and across the industry. Examples include the ICH M7 pharmaceutical impurities guideline and the published *in silico* toxicology protocols. To support an efficient, transparent, consistent and fully documented implementation of these protocols, a new and novel interactive software solution is described to perform such an integrated hazard assessment based on public and proprietary information.

Introduction

In silico toxicology (or computational toxicology) is being used directly or as part of the weight-of-evidence (WoE) for an increasing number of regulatory and industrial applications. This is driven by the need to (1) fill data gaps for chemicals in commerce with limited information, (2) improve the efficiency of the discovery process for chemical products, (3) support the replacement, reduction, and refinement of animal use (3Rs), and (4) support regulatory guidelines where *in silico* approaches are defined as acceptable approaches [1]. One such regulatory guideline is the International Committee for Harmonization (ICH) M7 guideline “Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk” [2]. This guideline includes a computational toxicology option as a regulatory accepted test to predict the bacterial reverse mutation assay (often referred to as the Ames test) [3]. This fast computational test is included for several reasons. Firstly, for many of these impurities there may be insufficient amounts of the test material available for performing an actual Ames test. This may require synthesizing the chemical (including actual or potentially present impurities) which would substantially add to the time and cost of performing such an assessment. In addition, such models have been shown to be sufficiently accurate, especially when coupled with an expert review, and they support the desired high-throughput assessment of the impurities [4–6].

The ICH M7 guideline describes how both experimental data alongside computational toxicology results are used to assess the potential for DNA-reactive mutagenicity, as shown in Fig. 1. The guideline uses this information to assign an impurity to one of five classes (shown in Table 1), which in turn supports whether an impurity needs to be controlled further or if additional testing is required. To support the assessment of classes 1, 2 and 5, it is important to identify any bacterial mutagenicity and carcinogenicity data available for any of the impurities. The guideline also identifies chemical classes representing high potency mutagenic carcinogens (termed “cohorts of concern”) which need to be handled separately as part of any risk assessment. These cohorts of concern include aflatoxin-like-, N-nitroso-, and alkyl-azoxy compounds. In the absence of any adequate experimental data, a computational assessment based on two complementary methodologies is recommended. One methodology should be an expert rule-based technology and the second should be a statistical-based technology. An expert review of all the information is prudent to assess the relevance and reliability of the both the experimental data as well as the computational results [4,5,7,8]. In addition, an expert review can support the class assignment for inconclusive computational results and even refute the results given sufficient evidence, such as proprietary results for chemicals analogs. The principles and procedures for performing and documenting this process have been published by a working group including both regulators and industry [5].

* Corresponding author.

<https://doi.org/10.1016/j.comtox.2021.100201>

Received 16 September 2021; Received in revised form 23 October 2021; Accepted 26 October 2021

Available online 28 October 2021

2468-1113/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

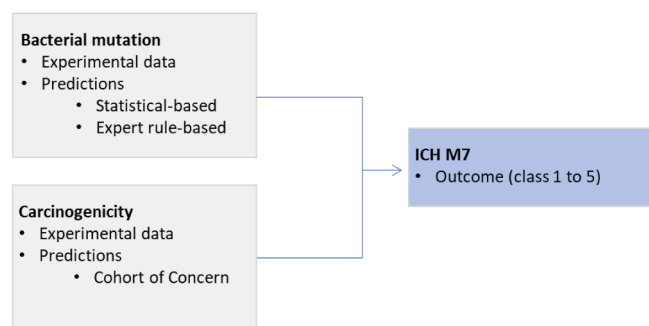


Fig. 1. Combining information on experimental data and computational toxicology results to support the ICH M7 class assignment. Bacterial mutagenicity and carcinogenicity data available for the target impurity are identified and combined with predictions. Statistical- and expert rule-based methods are applied for a computational toxicology assessment of mutagenicity. Predictions can identify high potency mutagenic carcinogens (cohorts of concern).

Table 1
ICH M7 Hazard Classification.

Class	Definition
1	Known mutagenic carcinogen
2	Known mutagen with unknown carcinogenic potential
3	Alerting structure, unrelated to the structure of the drug substance; no mutagenicity data
4	Alerting structure, same alert in related compounds which have been tested and are non-mutagenic
5	No structural alerts, or alerting structure with sufficient data to demonstrate lack of mutagenicity or carcinogenicity

The ICH M7 guideline is a widely adopted example of an approach to hazard assessment based on the integration of a battery of both experimental *in vitro* and *in vivo* data alongside *in silico* results coupled with an expert review. This type of integrated assessment is becoming increasingly common in approaches that support a more mechanistically-driven and animal-free assessment. Initiatives such as the Adverse Outcome Pathways (AOPs), Integrated Approaches to Testing and Assessment (IATA), New Approach Methodologies (NAMs), and Defined Approaches (DAs) are advancing and documenting the state of the science to enable these future alternative and integrated approaches [9–13].

Experimental data generated using accepted protocols, such as the OECD test guidelines [14], supports the use of this data across different regulatory authorities and industry. The development of equivalent protocols for the use of *in silico* methods would similarly support adoption of these methods, whether as a standalone alternative method or in combination with experimental results. The *in silico* protocols would build on work documenting best practices in computational toxicology, such as the OECD validation principles [15], and the described approaches to defining the battery of mechanisms and associated tests to support an integrated assessment.

A working group of over 70 organizations is currently generating such *in silico* toxicology protocols. This includes a framework outlining the components for any protocol [1] along with protocols for specific toxicology endpoints. To date, protocols for genetic toxicology [16] and skin sensitization [17] have been published with many protocols and position papers currently progressing. These protocols outline a series of defined toxicological effects or mechanisms that ideally should be assessed based upon available experimental data and/or *in silico* results. The protocols discuss the selection of such approaches, how to assess the reliability of the information provided, and how to combine the available information to establish an overall hazard assessment and associated level of confidence based on the WoE. The rules and principles underpinning this WoE process are provided within the protocol. This is illustrated conceptually in Fig. 2, showing how a series of toxicological

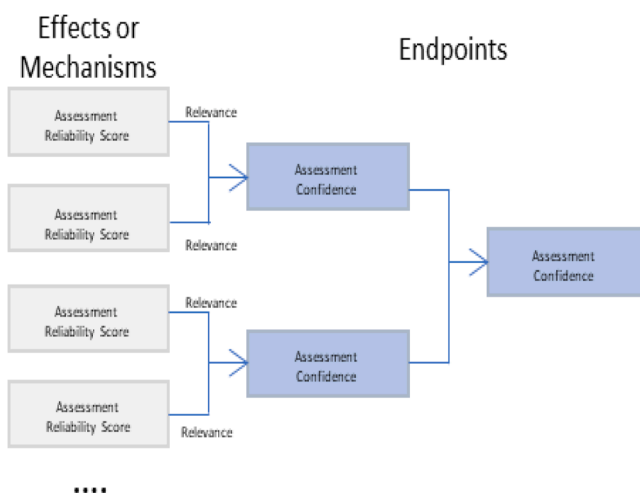


Fig. 2. A hazard assessment framework for *in silico* toxicology protocols.

effects/mechanisms are used to support the assessment of one or more toxicological endpoints; for example, this construct can be applied to assess the activation of the Nrf2-ARE pathway (the mechanism) within the prediction of skin sensitization in human (the endpoint). Guidelines for an expert review of the experimental and *in silico* results along with how the information may be combined are described within the protocols. The procedure for documenting the entire decision-making process, along with any expert review, is also described in the protocols. Hence, these protocols support the adoption of *in silico* approaches within a well-defined hazard assessment framework by ensuring such methods are performed in a consistent, transparent, and reproducible quality-driven manner.

Due to the complexity of these novel assessments described in such protocols, an interactive and visual software application for performing a hazard assessment is essential. This type of solution should support both the integration of the relevant experimental data and *in silico* predictions as well as the assessment of the reliability of the combined information. It should also steer the integration of all the available information based on the rules and principles described in the protocols. The tool should also provide the ability to perform an expert review of the experimental data and/or *in silico* results at the same time as allowing any reviewer to assess the overall process of combining the information. All expert review and any resulting changes should be documented along with the entire decision-making process.

The following paper outlines a proposal for an interactive and visual solution to this problem and discusses its implementation within the Leadscope computational toxicology solution. This includes the development of a visual and interactive hazard assessment platform in relation to the ICH M7 framework [4,5], the genetic toxicology *in silico* protocol [16], and the skin sensitization *in silico* protocol [17]. The paper covers how the content, including databases containing historical toxicity information and computation models, are developed. It explains how the results from such database searches and *in silico* model applications are integrated within a visual platform and how such a platform may be interrogated, and expert review performed and documented. The paper also presents information on the validation of the models and includes four case studies illustrating applications of such a platform.

Methods

Overview

The implementation of an integrated hazard assessment platform supporting the application of *in silico* toxicology protocols [1] is summarized in Fig. 3. The visual hazard assessment platform ideally queries

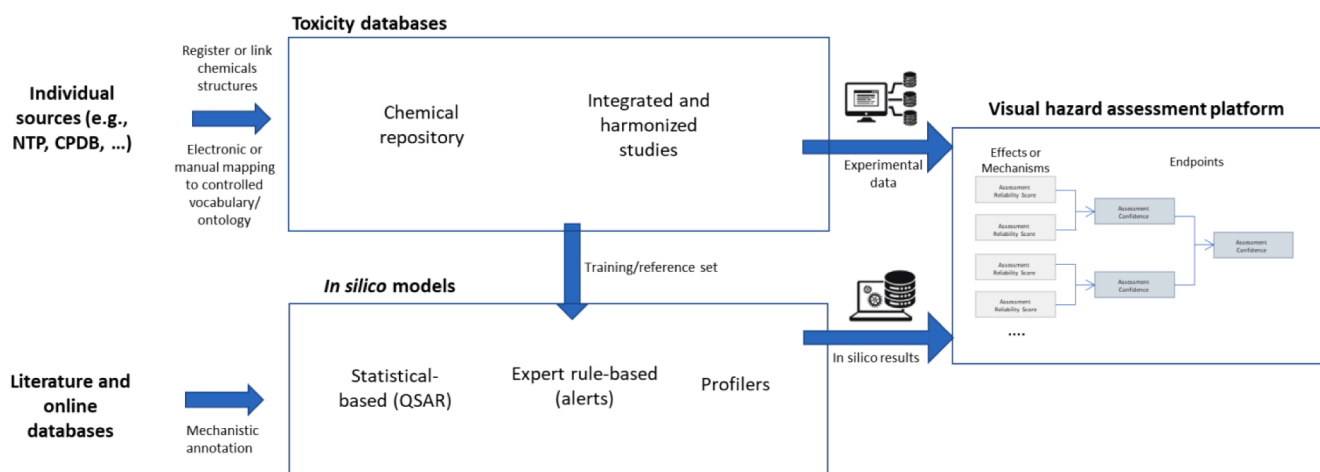


Fig. 3. Overview of the implementation of the visual hazard assessment platform. NTP refers to the National Toxicology Program online databases and CPDB refers to the Carcinogenicity Potency Database, QSAR refers to Quantitative Structure-Activity Relationships.

both the toxicity databases as well as applies *in silico* models to support the assessment of individual effects or mechanisms. Indeed, the platform uses both experimental and *in silico* results for each effect/mechanism defined in the protocol or guideline. To support access to the experimental results, the platform searches a database of historical toxicological studies linked to chemical structures. Public sources of toxicology data are used to populate this database. The database also supports the generation of *in silico* models based on different methodologies. In addition, these models are refined and annotated through access to the literature and other online databases which enrich the models with mechanistic interpretations. Once experimental and/or *in silico* results for individual effects/mechanisms have been identified and reviewed, endpoints are then calculated based on the input along with the rules and principles documented in the protocol. The visual platform interactively supports interrogation of the results and performing an expert review. The following sections outline how such toxicity databases and *in silico* models are developed within Leadscope computational toxicology solutions, how these resources are integrated within the platform, and how this platform can be interrogated. How such a platform has been developed to support the ICH M7 guideline as well as the two published protocols [16,17] is specifically discussed.

Toxicity database

As illustrated in Fig. 3, there are many public sources of toxicity study information. These include online databases (such as the National Toxicology Program [NTP] [18]), secondary sources of compiled information (such as the Carcinogenicity Potency Database [CPDB] [19]) as well as individual study records contained in publications or regulatory submissions. Information on both the tested chemicals and the toxicity study design and results are converted into an electronic database with information integrated for each compound. The following reviews the process of creating this content.

It is important that all studies for the same chemical are linked to the same electronic depiction of the chemical structure. This is achieved by comparing each chemical (test article) to the existing database. Based on this comparison, the test article is either registered as a new chemical and given a new Leadscope ID or it is linked to a previously registered chemical. It can be challenging when only a chemical name has been reported, especially when the chemical has been referred to by different names. When a chemical structure is displayed within the source material, the depiction of its stereochemistry as well as aromaticity and tautomerism are considered as part of this matching. To support the computational modelling, mixtures and salt forms are often linked to the modifiable forms of the chemical, referred to as the SAR form.

Studies can vary significantly in the level of detail provided in describing the methodology used in identifying, verifying, and representing the chemical substances of primary interest being reported on. In the best-case scenarios, an author will report three types of identification for substances: typed identification numbers, tradenames or systematic names, and a structural representation. In the worst-case scenarios, an author may only provide a synonym or codename for a substance, which, in some cases makes it impossible to determine any chemical structure representation. In each case encountered the information regarding the substance identification is vetted and cross-compared to ensure agreement. If a conflict arises in the cross-comparison efforts, the context of the study is taken into consideration to provide guidance in correctly identifying substances. For example, an examination of the totality of the information supports any resolution where different or incomplete stereochemistry is provided.

As part of the content building, information on both the study design and results needs to be included in the database to support transparency and expert review. The underlying information is not always in an electronic form that is suitable for processing automatically. In certain cases, it is necessary to enter the information by hand into an electronic representation. Where it is in an electronic form, it is possible to develop customized applications to read the content directly into the database. An essential process, irrespective of whether the step is performed manually or automatically, is to map the data elements described in the source material onto standardized terms. The Toxicity Markup Language or ToxML is a standardized organization of toxicity study design and results supported by controlled vocabularies that ensures the creation of a harmonized database [20].

A process of grading (i.e., creating an overall call for a specific toxicological effect or mechanism) is possible once the chemical structure registration process and the content processing is completed and the harmonized study records are linked to these chemicals. As an example, an overall assessment for bacterial mutagenicity would include an examination of the test and study calls for all entries matching each registered chemical. This process uses a series of rules to assess the different study results, such as whether an individual study source is trusted or authoritative and if the study is compliant with accepted guidelines. In cases where the results are conflicting across the different studies, the WoE needs to be considered to derive an overall assessment.

Table 2 summarizes the Leadscope database content used in the current platform and how such content maps onto the effects/mechanisms within the three implemented frameworks (ICH M7, the genetic toxicology *in silico* protocol, and the skin sensitization *in silico* protocol).

Table 2
Databases used to support the hazard assessment platform

Database	Sources	Mapped to effects/ mechanisms	Hazard assessment framework
Carcinogenicity	CCRIS, CDER, CFSAN-PAFA, CPDB, DSSTox, DBCAN, IARC, ISSCAN, NTP, RTECS	Carcinogenicity	ICH M7
Genetic toxicology	CCRIS, CDER, CFSAN-OFAS, CFSAN-PAFA, CPDB, EPA-Genetox, Submissions from organizations, Japanese NIHS, NTP, Tokyo Eiken, Publications, RTECS	Bacterial mutation	ICH M7, genetic toxicology
		Mouse Lymphoma	Genetic toxicology
		Chromosome aberration <i>in vitro</i>	Genetic toxicology
		Micronucleus <i>in vitro</i>	Genetic toxicology
		Chromosome aberration <i>in vivo</i>	Genetic toxicology
Skin sensitization	Publications, ICCVAM, NICEATM, ECHA	Micronucleus <i>in vivo</i>	Genetic toxicology
		Protein reactivity	Skin sensitization
		Activation of Nrf2-ARE	Skin sensitization
		Expression of co-stimulatory and adhesion molecules	Skin sensitization
		Reaction domain	Skin sensitization
		Rodent local lymph node proliferation	Skin sensitization
		Rodent maximization	Skin sensitization
Human skin sensitization	Skin sensitization		

Legend: CCRIS - Chemical Carcinogenesis Research Information System; CDER - US FDA CDER (Center for Drug Evaluation and Research) product approval reviews; CFSAN-PAFA - US FDA CFSAN (Center for Food Safety and Applied Nutrition) PAFA (Priority-based Assessment of Food Additives) database; CFSAN-OFAS - Genetic toxicity database from the US FDA CFSAN (Center for Food Safety and Applied Nutrition) reviews; CPDB - Carcinogenicity Potency Data Base; DSSTox DBCAN - Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network; DBCAN: EPA Water Disinfection By-Products with Carcinogenicity Estimates; EPA-Genetox - Mutagenicity test data from the US EPA; IARC - International Agency for Research on Cancer and Carcinogenicity classification; ISSCAN - Chemical carcinogens: structures and experimental data from Istituto Superiore di Sanita; Japanese NIHS - National Institute of Health Sciences of Japan (Publicly release class A chemicals); NTP - National Toxicology Program; RTECS - Registry of Toxic Effects of Chemical Substances; ICCVAM - Interagency Coordinating Committee on the Validation of Alternative Methods; NICEATM - The NTP Interagency Center for the Evaluation of Alternative Toxicological Methods; ECHA - European Chemicals Agency

In silico models

Both the ICH M7 guideline and the *in silico* toxicology protocols recommend using multiple computational methodologies, including statistical-based and expert rule-based, since multiple concurring complementary methods increase the reliability of the prediction results [1]. Methodologies to profile chemicals into different toxicologically relevant categories and read-across approaches also provide key information in any hazard assessment framework. The following section outlines the computational models used within the hazard assessment platform.

Statistical-based or Quantitative Structure-Activity Relationship (QSAR) models are developed within the Leadscope predictive data miner [21]. A number of these models predict a binary outcome, for example, the bacterial mutation statistical QSAR model predicts whether a chemical is mutagenic or non-mutagenic based on predictions made using Ames test data. These models use a training set of chemicals and toxicological data (response) extracted from the toxicological

database previously described. The models are based on a number of calculated descriptors: (1) pre-defined structural features [22], (2) calculated physico-chemical properties, (3) chemical scaffolds automatically identified to map onto a disproportionate numbers positive or negative examples [23], and (4) significant active structural features extracted from the literature. Having selected an appropriate subset of these descriptors, a computational model based on the partial logistic regression algorithm [24] is applied to encode the relationship between the descriptors and the toxicological response. The models are further refined and then validated based on cross-validation and external validation wherever possible. The models generate a probability of a positive outcome, and a final prediction is made using defined cut-off values. For example, when the bacterial mutation model calculates a probability greater than 0.6 a positive assignment is made, a probability less than 0.4 is assigned to be negative, and those predictions between 0.4 and 0.6 are assigned as indeterminates. The implementation of the models performs an additional key step to assess whether the test chemical is within the applicability domain of the model, i.e., whether there is an increased reliability because of the overlap with similar training set examples as well as features used in the model.

There are three types of QSAR models used within this platform: (1) “single statistical” models (using the methodology discussed as in the case of the bacterial mutation QSAR model), (2) “balanced statistical” models (used in cases where the toxicity response is skewed as in the case of the *in vivo* micronucleus QSAR model), (3) “categorical statistical” models (used when the response is ordinal as in the case of models related to skin sensitization).

The balanced statistical approach uses a series of models that are based on subsets from the training set, where each set is over or under-sampled to create more even distribution of positive and negative examples. Training set examples from the underrepresented positive or negative class will be present in more than one subset. When making a prediction, the test chemical will be run through all models and an overall prediction calculated based on the combined results.

When the toxicological response outcome is a categorical value based on either the severity of the outcome or the toxic dose, a series of models are built and incorporated within a decision tree. For example, the toxicological outcome for a Local Lymph Node Assay (LLNA) model is strong/extreme (where the effect concentration ($EC3^1$) is less than 1, moderate ($1 \leq EC3 < 10$), weak ($10 \leq EC3 \leq 100$), or non-sensitizer ($EC3 > 100$). A series of individual models are built based on these cut-off values. In this case, three binary models are built to predict each of these categories; for example, an individual model predicts whether a chemical has an EC less than 1 (strong/extreme sensitizers), and two other models predict the moderate and weak categories. The results from each of the models are then combined within a decision tree (as illustrated in Fig. 4) to calculate the final category.

Besides QSAR, a second methodology referred to as expert rule-based is developed in the Leadscope computational toxicology solution [21]. This is based on a series of structural alerts that encode features that activate and deactivate the toxicity. Such alerts are derived from expert knowledge embedded in the literature and/or extracted from toxicity databases. Structural alerts are ideally accompanied by a monograph describing the relevance of the moiety in the context of the endpoint of interest, such as any mechanistic rationale, as well as all examples from the database to support a contextual assessment. The identification of the series of expert alerts encoded within the Leadscope computational toxicology solution [21] can be assisted by specific informatics capabilities, such as clustering [25] and identification of significant chemical scaffolds [23].

The application of expert alerts to any test chemical will result in a prediction (such as positive, negative, or indeterminate) alongside a

¹ $EC3$ value: the amount of a chemical that is required to elicit a three-fold increase in lymph node cell proliferative activity ($SI \geq 3$)

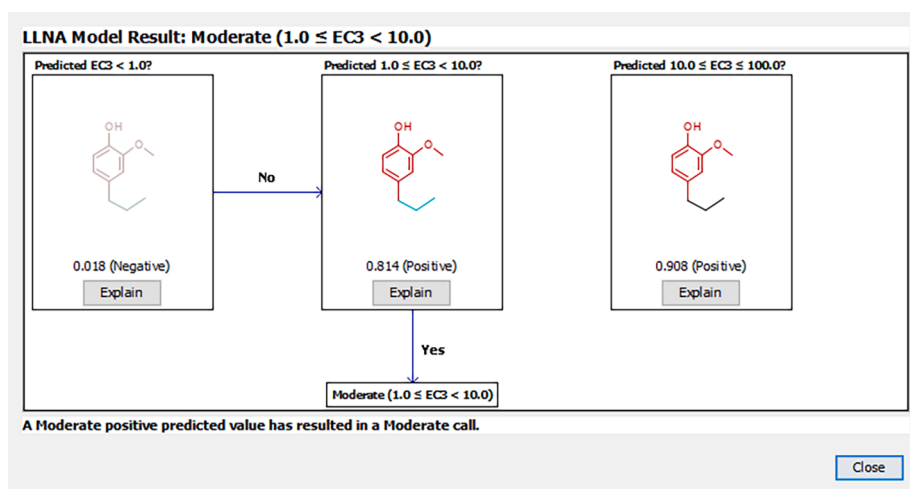


Fig. 4. Decision tree for calculation of LLNA severity as implemented in the Leadscope computation toxicology solution Leadscope Model Applier. The platform provides a means to explore such decision tree that combines the underlying “categorical statistical” models.

confidence score based on the toxicological response value’s precision derived from the matching examples. Since a prediction is being made, it is essential to understand the applicability domain through a comparison of the test chemical to the underlying reference set of compounds supporting the alerts.

In silico profilers also make use of mechanism-based structural alerts [26]; however, they do not directly predict a toxicological outcome but place a chemical into a category to support either an assessment or an expert review. Several profilers have been incorporated within the Leadscope platform, including carcinogenicity cohorts of concern [2] and reaction domains [27] to support the assessment of skin sensitization.

Finally, read-across is used to predict a toxicological outcome for a given chemical (target) based on the toxicological response from one or more sufficiently similar analogs (source). A read-across tool has been incorporated within the Leadscope computational toxicology solution to provide the opportunity to include such a prediction for the different effects or mechanisms. The tool identifies similar chemicals based on a series of different approaches, including structural similarity or a pre-defined chemical category. The tool supports the interactive exploration and refinement of the source chemicals, including the addition of proprietary examples, which can be documented in the tool. It also helps formulate how the toxicity data on the source chemicals is read-across onto the target. Frameworks, such as the Read-Across Assessment Framework or RAAF [28], are incorporated within the platform to support the complete expert review and documentation of the read-across study.

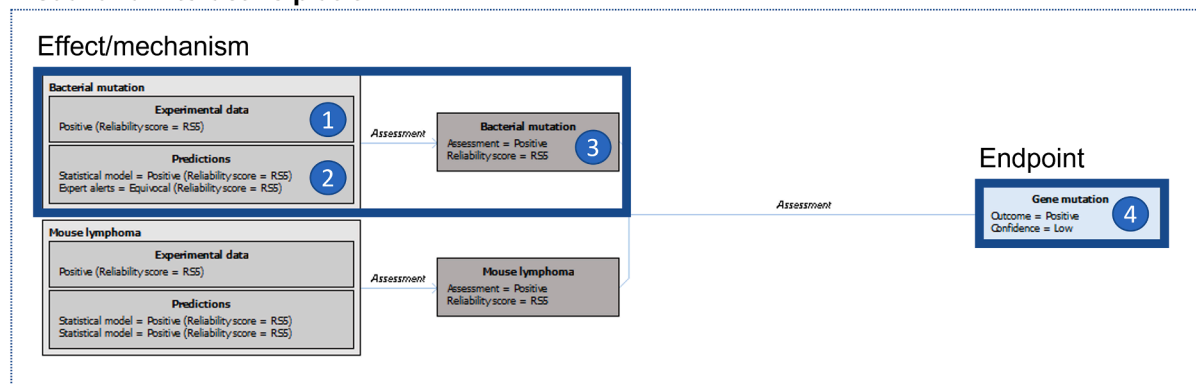
Table 3

Summary of models incorporated into the current hazard assessment platform.

Hazard assessment framework	Effect/mechanism	Computational models	Type of model	References
ICH M7	Carcinogenicity	Cohort of concern v1	Profiler	
ICH M7, genetic toxicology	Bacterial mutation	Bacterial mutation v2	Statistical-based	[33]
ICH M7, genetic toxicology	Bacterial mutation	Bacterial mutation v7	Expert rule-based	
Genetic toxicology	Mouse Lymphoma	MLA Activated v2; MLA unactivated v2	Statistical-based	
Genetic toxicology	Chromosome aberration <i>in vitro</i>	CA CHL v2	Statistical-based	
Genetic toxicology	Chromosome aberration <i>in vivo</i>	In vivo CA v2	Statistical-based	
Genetic toxicology	Micronucleus <i>in vivo</i>	In vivo micronucleus mouse v2	Statistical-based	[34]
Skin sensitization	Protein reactivity	DPRA v2	Statistical-based	
Skin sensitization	Activation of Nrf2-ARE	KeratiSense v2	Statistical-based	
Skin sensitization	Expression of co-stimulatory and adhesion molecules	h-CLAT v2	Statistical-based	
Skin sensitization	Reaction domain	Reaction domain v2	Profiler	
Skin sensitization	Rodent local lymph node proliferation	LLNA	Statistical-based	
Skin sensitization	Rodent local lymph node proliferation	LLNA	Expert rule-based	

Legend: ICH M7 - International Committee for Harmonization (ICH) M7 guideline; CA – Chromosome aberration; CHL - Chinese Hamster Lung cells; DPRA - Direct Peptide Reactivity Assay; h-CLAT – Human Cell Line Activation Test; LLNA - Local Lymph Node Assay

Visual and interactive platform



Corresponding on-demand details

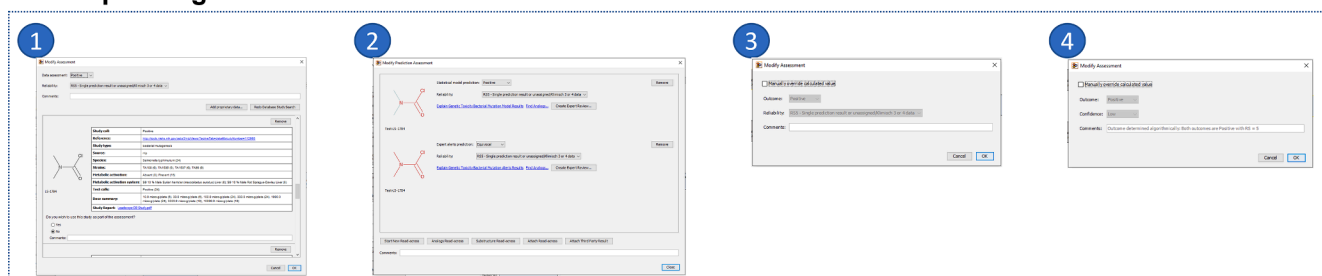


Fig. 5. On demand details available for nodes representing the effects/mechanisms and the derived endpoints.

Table 4
Reliability Scores

Reliability Score	Klimish Score	Description	Summary
RS1	1	Data reliable without restriction	Well documented and accepted study or data from the literature Performed according to valid and/or accepted test guidelines (e.g., OECD) Preferably performed according to good laboratory practices (GLP)
RS2	2	Data with restriction	Well documented and sufficient Primarily not performed according to GLP Partially complies with test guideline
RS3	–	Expert review	Read-across Expert review of <i>in silico</i> result (s) and/or Klimisch 3 or 4 data
RS4	–	Multiple concurring prediction results	
RS5	–	Single acceptable <i>in silico</i> result	
RS5	3	Data not reliable	Inferences between the measuring system and test substance Test system not relevant to exposureMethod not acceptable for the endpoint Not sufficiently documented for an expert review
RS5	4	Data no assignable	Lack of experimental details Referenced from short abstract or secondary literature

an assessment for the Gene Mutation endpoint (positive) along with a confidence score (low). The rules/principles for combining these types of information and generating the final assessment together with corresponding confidence are detailed in the protocols. It should be noted that the overall reliability score is associated with a single effect/mechanism based on available experimental and/or *in silico* data for that effect/mechanism. The confidence score is associated with a single endpoint; however, it is based on the propagation of information from all related effects/mechanisms as well as other endpoints.

Each node in the Leadscape platform is interactive: by clicking on any node further information is shown, as illustrated in Fig. 5. For example, by clicking on the box annotated with a “1”, information on the individual studies from the toxicology database search is displayed. This includes a summary of the results and a link to the full study report. It is possible to select which of the studies, based on a review of the study adequacy, to include in the current assessment. A default overall assessment and reliability score based on the studies identified is shown; however, it is possible to update both values following an expert review of the data. It is also possible that a proprietary study has been run on the chemical, and this study may be included in the expert review; it can thus be integrated and documented in the assessment by summarizing the results and uploading the full study report into the platform.

Further details on the predictions are also available by clicking on the box annotated with a “2” (Fig. 5). This includes an explanation of the model results and access to structurally similar analogs. The protocols provide guidelines for elements to consider as part of an expert review of the *in silico* results. These guidelines are also incorporated into the platform, as shown in Fig. 6. An inspection of any of these guideline elements may: (1) increase in the prediction’s reliability, (2) result in no increase in the prediction’s reliability, (3) refute the prediction, or (4) provide no additional supporting information. Fig. 6 shows how, for each of the elements of an expert review, it is possible to view contextual information to support and document such an examination.

The platform allows for the integration of a read-across study in an assessment for any of the effects/mechanism. A node is linked to the read-across tool that will both perform and document the read-across

Checklist of 7 review options

Contextual review of one option – applicability domain considerations

1. Applicability domain considerations

For the test compound to be within the applicability domain of the model, there must be at least one structural analog and at least one structural feature in both the training set and the model and at least one common structural feature in both the test compounds and the model. Analysis of this applicability domain information...

has not been concluded
Comments:

2. Calculation of probabilities. The probability is 0.992 and the prediction is Positive. This probability is based on the weight of the contributing features. An analysis of the feature weightings...

has not been concluded
Comments:

3. Relevancy of model descriptors. The model uses a series of descriptors. An examination of the relevancy of the descriptors...

has not been concluded
Comments:

4. Sufficiency of training set data. The descriptors used in the model are included based on the data from the training set. An examination of the sufficiency of this data...

has not been concluded
Comments:

5. Potentially reactive features. An examination of any other potentially reactive groups...

has not been concluded
Comments:

6. Comparison with drug substance or related compound. An examination of the drug substance or related compound...

has not been concluded
Comments:

7. Other considerations. Other considerations, such as the performance of the model on chemical analogs, may be considered. Analysis of these other considerations...

has not been concluded
Comments:

Overall recommendation: In your expert opinion, the review of the underlying QSAR information...

increases the prediction reliability
 does not increase the prediction reliability
 refutes the prediction
 has not been concluded

Comments:

Cancel OK

Applicability domain considerations

1. Applicability domain considerations

Test:LS-1784

For the test compound to be within the applicability domain of the model, there must be at least one structural analog and at least one structural feature in both the training set and the model and at least one common structural feature in both the test compounds and the model. Analysis of this applicability domain information...

increases the prediction reliability
 does not increase the prediction reliability
 refutes the prediction
 has not been concluded

Comments:

Analogs

The chemical analogs were identified from the training set of the model. The analogs are ordered according to their degree of similarity (S), shown as the first value. The second value shown is the data (1.0 is positive and 0.0 is negative). Double click on the chemical structure to access the data.

S	BM	Chemical Structure
S=1.0	BM=1.0	LS-1784
S=0.57	BM=1.0	LS-439387
S=0.55	BM=1.0	LS-159983

Cancel OK

Fig. 6. *In silico* expert review checklist and accompanying contextual information.

study. It is also possible to add the results from models not directly incorporated within the platform or a read-across assessment performed outside the platform. The details of these external results can be added, including meta information, modelling approach or any other parameters. The full *in silico* report can also be uploaded into the platform to provide full transparency.

The platform collects all the information tied to both experimental and *in silico* results, and any expert review that modified the individual assessments or reliability scores; an overall assessment for the effect/mechanisms is then automatically derived as shown in Fig. 5 (annotated with “3”). Additional details on the rules and principles that were used to derive these values can be inspected and potentially modified based on a documented expert review. Fig. 7 shows how the reliability score for the bacterial mutation experimental data was changed from RS5 to RS1, after an expert review concluded the results warrant the highest reliability score.

Fig. 5 also shows how this information associated with effects/mechanisms is, in turn, used to make an assessment for one or more derived endpoints alongside a confidence level. The rule/principles for deriving this call as documented in the protocol are available for inspection, i.e., clicking on the node annotated with “4” in Fig. 5. In a similar manner, it is also possible to revise the assessment based on a documented expert review.

For the specific case of a complete ICH M7 hazard assessment, the

corresponding platform is shown in Fig. 8 and includes experimental data and *in silico* prediction models/profilers for bacterial mutation and carcinogenicity. This information is combined, based on a series of rules (shown in Fig. 9), to generate an overall ICH M7 class designation (as shown in Table 1) along with supporting information on the reliability and confidence of the information. It should be noted, there is no intermediate assessment of individual effects/mechanisms as the ICH M7 class designation is based on the outcome of the available two nodes, i.e., bacterial mutation and carcinogenicity. Fig. 10 shows the complete genetic toxicology hazard assessment framework, and Fig. 11 the complete skin sensitization hazard assessment framework as they are implemented within the platform.

Results

Tables 5 and 6 summarize both the database content and the *in silico* models' performance that are used within the hazard assessment platform. The platform currently comprises three finalized frameworks: the ICH M7 hazard assessment framework, complete genetic toxicology hazard assessment framework [16], and the complete skin sensitization hazard assessment framework [17].

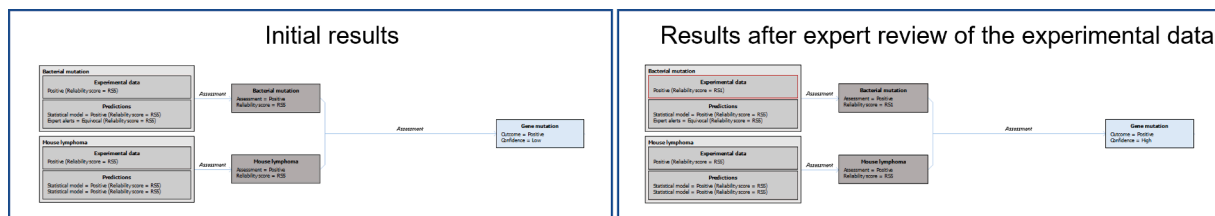


Fig. 7. Interactively modify the results.

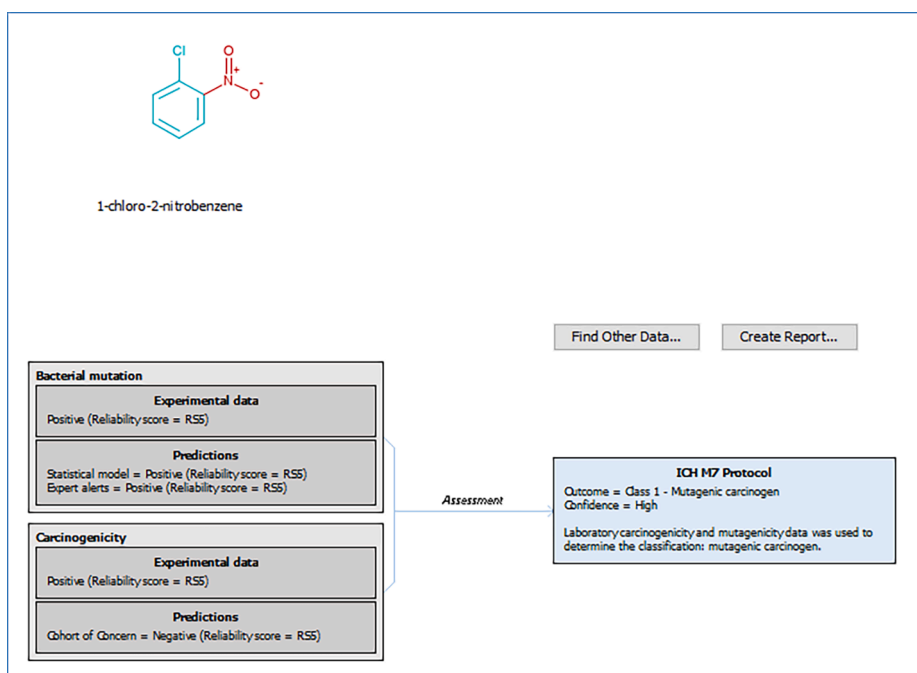
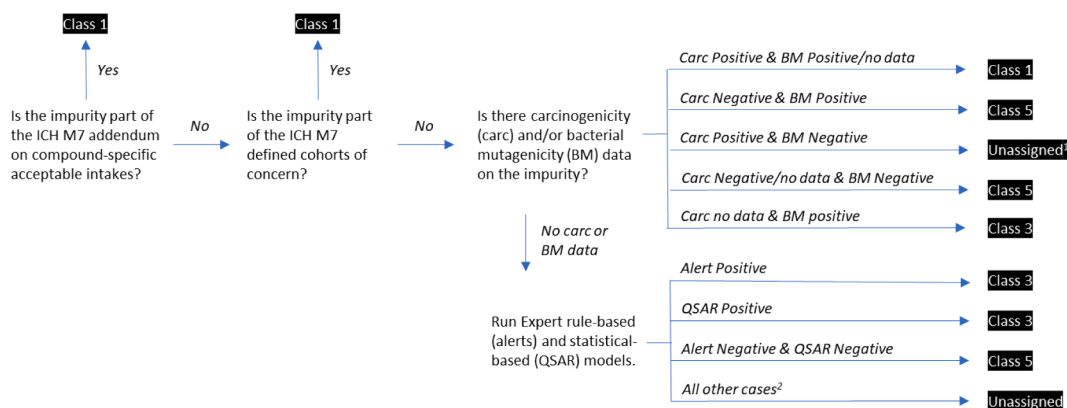


Fig. 8. Implementation of the ICH M7 hazard assessment framework within the platform.



Notes

1. Likely non-genotoxic carcinogen to be controlled according to ICH Q3C (Permissible Daily Exposure [PDE]). Carcinogens that are negative in the bacterial reverse mutation assay do not have a DNA reactive mechanism of carcinogenicity and therefore are not in scope of the ICH M7 guidance (e.g., acetamide and hydroxylamine), (Q&A 2 July 2020)
2. Single negative prediction (other prediction inconclusive or out-of-domain) or both predictions are inconclusive or out-of-domain

Fig. 9. Rules for deriving the ICH M7 classification.

Case studies

Overview

The only required information to initiate a hazard assessment is the electronic record of the chemical structure. This can be either uploaded from a file, such as a MOL file or SMILES string, identified through a database search or drawn within Leadscope's structure drawing package or elsewhere. The following case studies illustrate how the platform (implemented in the Leadscope model applicator v3.1) described in this paper can be used to assess four chemicals.

Case study 1: ICH M7 assessment of 2-bromo-5-acetamidobenzoic acid

Upon running the ICH M7 protocol for the target impurity (2-bromo-5-acetamidobenzoic acid), available experimental data and the outcome

of the individual models (i.e., expert rule- and statistical-based systems) are summarized in a table alongside the ICH M7 class assignment, corresponding confidence and additional supportive evidence and comments (Fig. 12). More specifically, no experimental bacterial mutagenicity data nor carcinogenicity data are available from the Leadscope databases, and the two complementary (Q)SAR methodologies provide negative predictions for bacterial mutagenicity. This information is automatically combined into a Class 5 assignment with a default medium level confidence. Such confidence level is justified by the absence of any expert review at this initial stage of the protocol workflow.

To increase the confidence level, an expert review can be conducted, and it is guided by the workflow encoded in the ICH M7 protocol in the Leadscope Model Applicator. Fig. 13 illustrates the different steps of this workflow. The expert review of the statistical-based model confirms the negative prediction of the target impurity based on the following

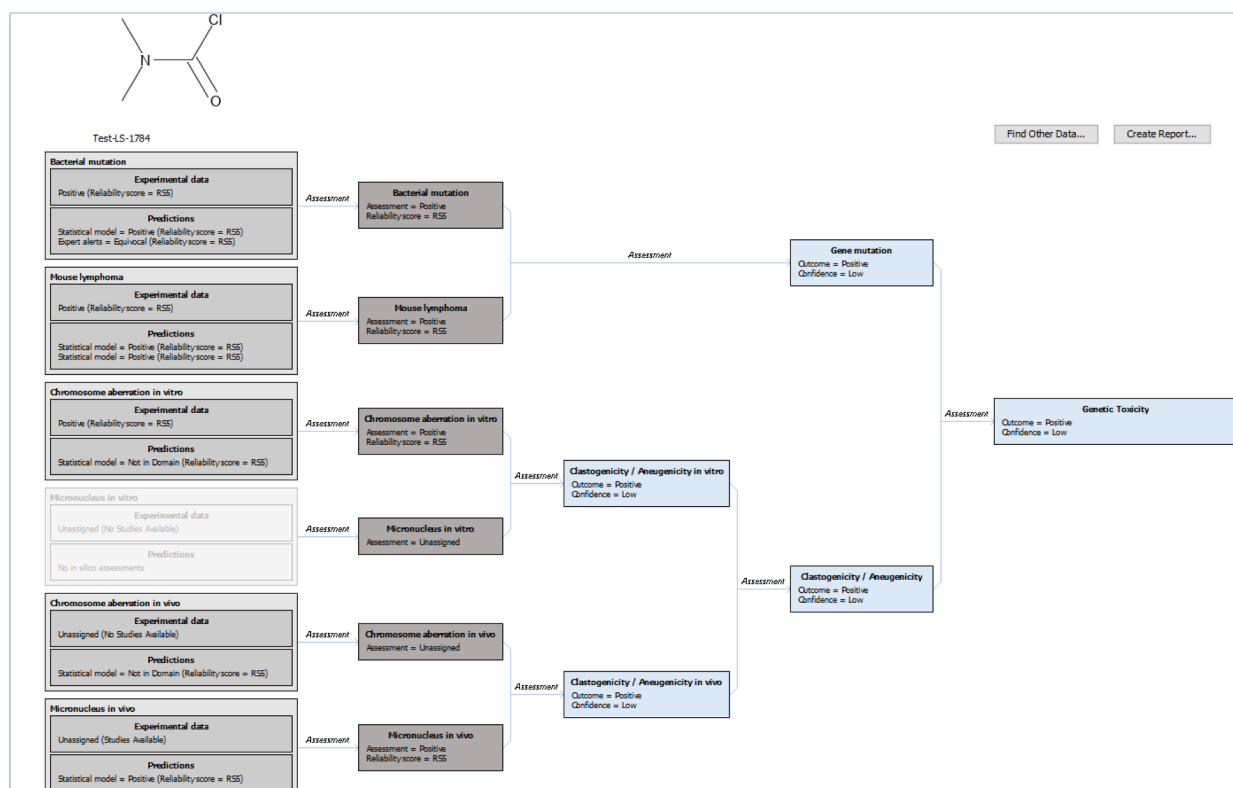


Fig. 10. Implementation of the genetic toxicology *in silico* protocol's hazard assessment framework within the platform.

elements: a) a low positive prediction probability ($PPP = 0.173$); b) a good coverage of the structure of the target impurity; c) negative features providing higher contributions to the prediction leading to a clear negative call; d) no concern from the features associated with positive contributions to the prediction since these features are represented in experimentally negative compounds (e.g., Acedoben and Acetanilide); e) good accuracy of analogs' predictions that supports the reliability of the prediction. The review of the expert rule-based system also confirms the negative outcome, given the absence of structural alerts of potential concern for mutagenicity, and supporting evidence coming from the experimental data for the closer analogs in the alert reference set, which are negative and correctly predicted as negative by the model.

For this molecule, the expert review notes that the potential metabolism of the target impurity should be also considered since the chemical contains an aromatic amide functional group, that may be bioactivated to a primary aromatic amine [29]. Certain primary aromatic amines are mutagenic, and the position of ring substituents influences the chemical's mutagenic potential [30]. Ahlberg and co-workers analyzed a series of functional groups in different positions relative to the amino group to determine whether they are potentially activating [30]. In the case of the target impurity, the carboxylic acid in the meta position and the bromine in the para position are not considered activating. Therefore, even if the chemical undergoes metabolic activation resulting in a primary aromatic amine, the metabolite is unlikely to be mutagenic given the presence of these two ring substituents.

The expert review considerations confirm the negative outcome provided by the statistical- and expert rule-based models and the call's reliability score is increased for the individual models to RS3 (prediction with expert review). The outcome of the two models can then be combined to derive an overall assessment of the target impurity. For the combination of the two negative results, the current expert review also considers the low risk of missing a mutagenic impurity according to the analysis published by Amberg et al. [4]. This analysis using a large bacterial mutation data set shows that, when both statistical-based and expert rule-based methods generate a negative (in domain) assessment,

such mutagenic risk is equal to 8.1% (6% by Dobo et al. [6]). The expert review can thus conclude that the target impurity is predicted as not mutagenic, i.e., negative for bacterial *in vitro* mutagenicity (Ames test), and the confidence of the prediction is increased to a high level. As such, the target impurity 2-bromo-5-acetamidobenzoic acid is assigned to the ICH M7 Class 5, and a standard report is generated including all the considerations of the expert review that were duly mapped throughout the ICH M7 protocol workflow.

Case study 2: ICH M7 assessment of 1-chloro-2-nitrobenzene

The ICH M7 hazard assessment performed for 1-chloro-2-nitrobenzene identifies available experimental data in the Leadscope toxicity database: positive bacterial mutagenicity and carcinogenicity data. These data are organized by the tool in the summary table illustrated in Fig. 14. The target impurity is preliminarily assigned to the ICH M7 Class 1 by the standardized workflow, which automatically sets a high confidence for the assignment.

An expert review of the positive data is conducted to confirm the overall positive outcome as illustrated in Fig. 15. By clicking on the experimental data bacterial mutagenicity node, the data used in the assessment can be inspected alongside all the mutagenicity studies identified for the target impurity. An expert review of the available data determines that there is clear evidence of the mutagenic activity of the target impurity and therefore the experimental data reliability is increased from a reliability score of RS5 to a reliability score of RS1. To further confirm the positive outcome, the *in silico* predictions for 1-chloro-2-nitrobenzene are analyzed. The two complementary (Q)SAR methodologies provide consistent positive outcomes. The positive prediction given by the statistical model is driven by the correspondence of target impurity with an experimentally positive training compound; the model features identified by the model adequately cover the structure of the target impurity, with the aromatic nitro and cyclic nitro moieties providing higher contribution to the prediction. This results in a clear positive call. The expert rule-based methodology further supports the

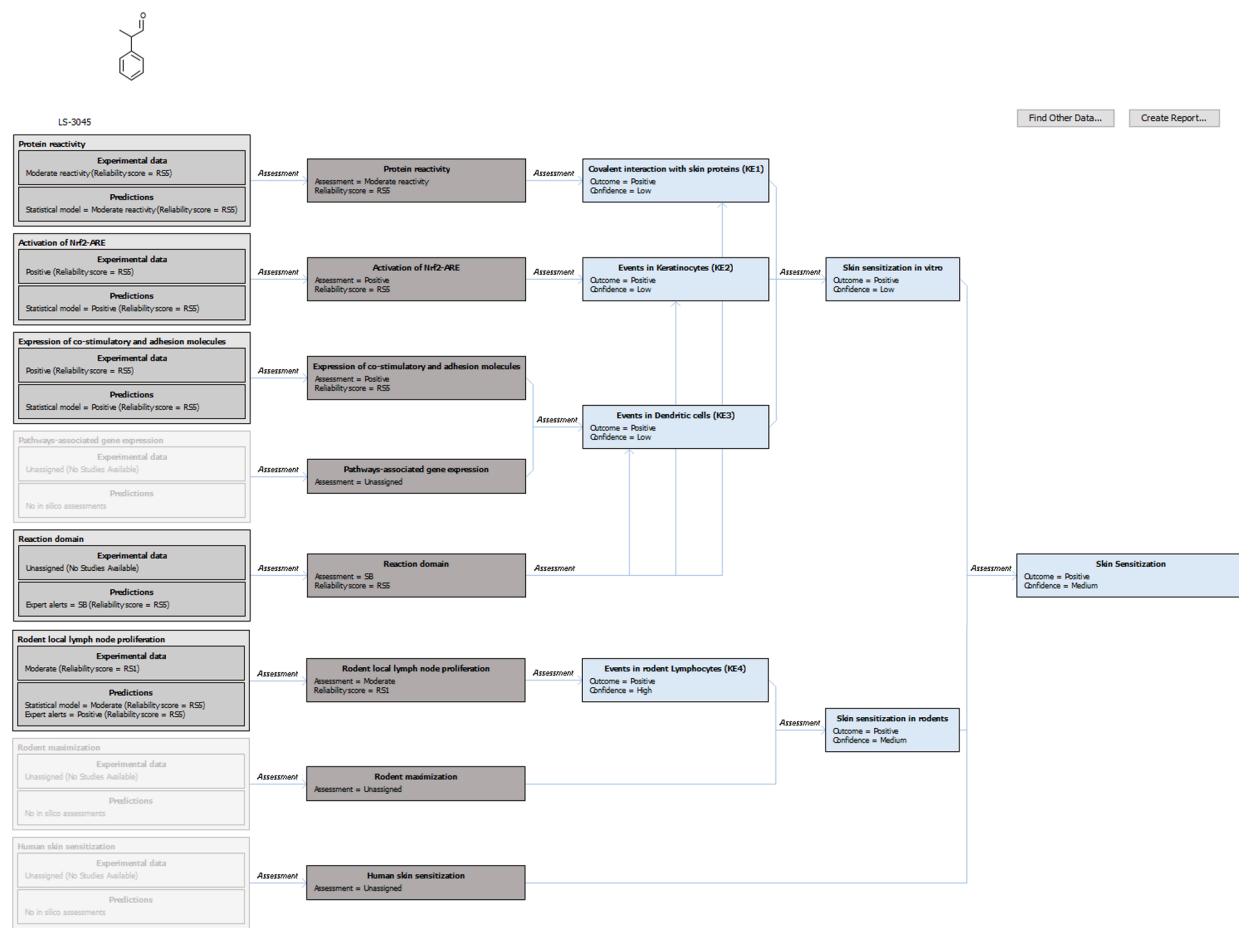


Fig. 11. Implementation of the skin sensitization *in silico* protocol's hazard assessment framework within the platform.

Table 5
Summary of the results from the database

Database	Mapped to effects/mechanisms	Number of chemicals	Number of studies	Number of tests
Carcinogenicity	Rodent carcinogenicity	5,700	18,084	27,099
Genetic toxicology	Bacterial mutation	12,694	41,914	288,280
	Mouse Lymphoma Chromosome aberration <i>in vitro</i>	5,921	11,764	16,227
	Micronucleus <i>in vitro</i> Chromosome aberration <i>in vivo</i>	1,298	794	1,065
Skin sensitization	Micronucleus <i>in vivo</i>	1,026	2,679	3,054
	Micronucleus <i>in vivo</i>	4,078	9,148	12,010
	Protein reactivity	271		
	Activation of Nrf2-ARE	281		
	Expression of co-stimulatory and adhesion molecules	239		
	Reaction domain	458		
	Rodent local lymph node proliferation	2,176	3,266	1,893
Rodent maximization	54			
Human skin sensitization	151			

positive outcome because of the identification of the aromatic nitro structural alert of potential concern for mutagenicity. In addition, the target impurity belongs to the alert reference set and this steers the positive expert rule-based prediction. Based on the above

considerations, the current expert review concludes that the target impurity 1-chloro-2-nitrobenzene is mutagenic, i.e., positive for bacterial mutation (Ames test) with a high confidence.

The positive experimental carcinogenicity data and corresponding studies identified by the tool for the target impurity are next inspected together with all the carcinogenicity studies. Such studies provide clear evidence of the carcinogenic activity of 1-chloro-2-nitrobenzene and therefore the experimental data reliability is increased from a reliability score of RS5 to a reliability score of RS1.

The ICH M7 protocol workflow allows for the combination of the positive bacterial mutation and carcinogenicity results that lead to an ICH M7 Class 1 assignment with high-confidence for the target impurity 1-chloro-2-nitrobenzene. Experimental data, *in silico* predictions and considerations of the expert review are all structured in a standardized report to be shared with third parties. A reviewer can then use the detailed information organized in such report to formulate an independent assessment.

Case study 3: Genetic toxicology assessment of m-xylylenediamine

When the genetic toxicology protocol is applied on m-xylylenediamine, the tool performs database searches and run *in silico* models for each effect/mechanism defined by this hazard assessment framework [16], summarized in Fig. 15.

The protocol window, as shown in Fig. 17, then guides the expert review of the experimental data and *in silico* results as gathered by the tool for each effect/mechanism.

First, the genetic mutation potential is assessed by considering available information for bacterial mutation and mammalian gene

Table 6
Summary of *in silico* performance results

Effect/ mechanism	Computational models	Type of model	Type of validation	Count	Concordance (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NVP (%)	Comment
Bacterial mutation	Bacterial mutation v2	Statistical	Cross validation (5%)	9,254	85	85	86	88	83	
Bacterial mutation	Bacterial mutationv2	Statistical	External validation	388	83	82	83	56	95	Reported in [33]
Bacterial mutation	Bacterial mutation v7	Expert rules	Internal validation	11,528	87	87	88	89	86	
Mouse Lymphoma	MLA Activated v2	Statistical	Cross validation (5%)	675	76	75	76	72	80	
Mouse Lymphoma	MLA unactivated v2	Statistical	Cross validation (5%)	752	76	79	74	73	80	
Chromosome aberration <i>in vitro</i>	In Vitro Chrom Ab CHL v2	Statistical	Cross validation (3%)	874	77	80	74	78	76	
Chromosome aberration <i>in vivo</i>	In Vivo Chrom Ab Comp v2	Statistical	Cross validation (2%)	285	77	80	74	78	76	
Micronucleus <i>in vivo</i>	In vivo micronucleus mouse v2	Statistical	Cross validation (5%)	1001	76	75	76	60	87	3 sub-models
Micronucleus <i>in vivo</i>	In vivo micronucleus mouse	Statistical	External validation	42	80	67	84	57	89	91% coverage; Reported in [34]
Protein reactivity	DPRA v2	Statistical	Cross validation	176	87	93	71	90	79	Categorical model. The sensitivity and specificity of the DPRA categorical model was calculated based on the binary values of positive and negative, where positive reactivity values are defined as a mean % depletion > 6.38% (low, moderate and high reactivity), and the no or minimal reactivity class (mean % depletion <= 6.37%) is negative.
Activation of Nrf2-ARE	KeratinoSens v2	Statistical	Cross validation (10%)	234	78	83	71	79	76	
Expression of co- stimulatory and adhesion molecules	h-CLAT v2	Statistical	Cross validation	179	75	76	72	91	46	4 sub-models
Rodent local lymph node proliferation	LLNA v2	Statistical (categorical)	Cross validation	843	80	85	73	82	77	Categorical model. The sensitivity and specificity of the LLNA categorical model was calculated based on the binary values of positive and negative, where positive values are defined as a EC3 % <= 100% (weak, moderate and strong/extreme sensitizers), and the non- sensitizers (EC3 % > 100%) are negative.
Rodent local lymph node proliferation	LLNA v2	Expert rules	Internal validation	843	85	80	92	93	77	The sensitivity and specificity of the LLNA categorical model was calculated based on the binary values of positive and negative, where positive values are defined as a EC3 % <= 100% (weak, moderate and strong/extreme sensitizers), and the non- sensitizers (EC3 % > 100%) are negative.


Integrated hazard assessment		Individual models				
Regulatory protocol: ICH M7						
Structure	Laboratory Data	Expert rule based system	Statistical based system	M7 Class assessment	M7 Class confidence	Additional supportive evidence and comments
 2-bromo-5-acetamidobenzoic acid		Negative No Alerts	Negative PPP = 0.173	Class 5 - Non-mutagenic	Medium	The impurity lacks reactive mutagenic or carcinogenic potential

Fig. 12. Summary of the preliminary results of the ICH M7 hazard assessment of 2-bromo-5-acetamidobenzoic acid.

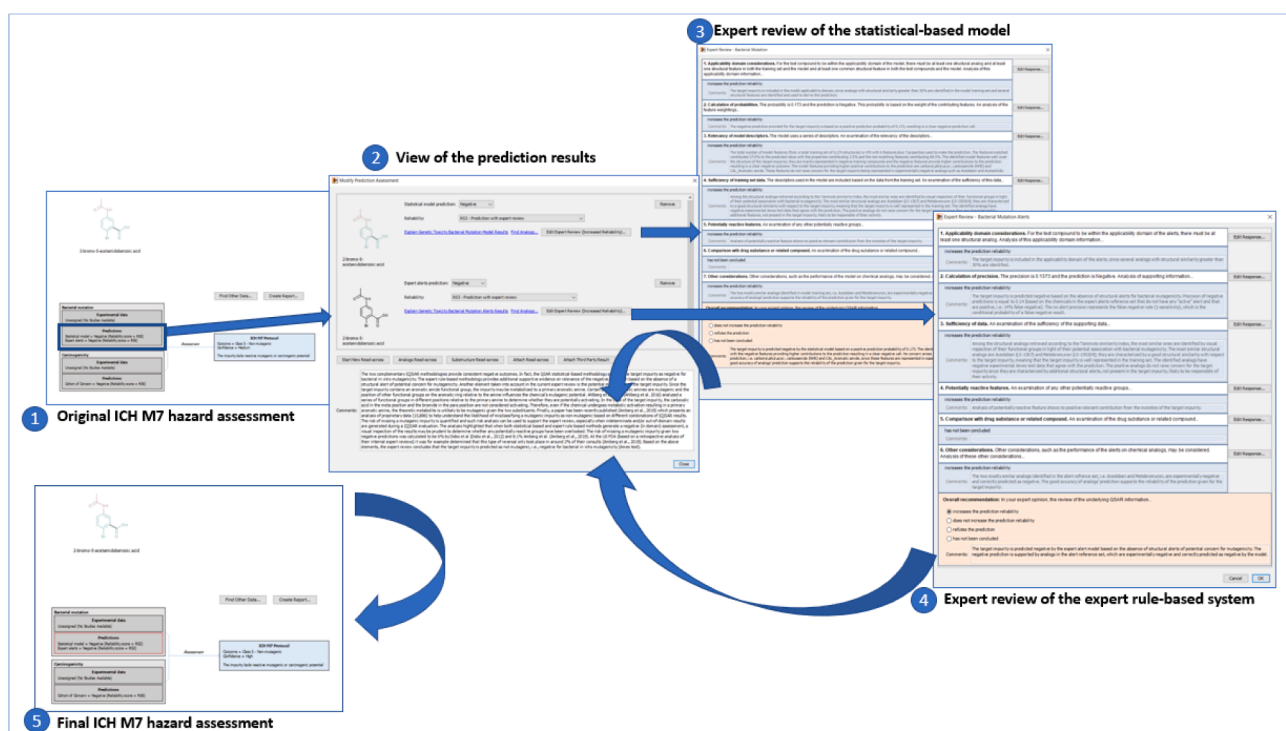


Fig. 13. ICH M7 hazard assessment workflow for 2-bromo-5-acetamidobenzoic acid.

mutation. The negative bacterial mutation experimental data is inspected. It indicates clear evidence of non-mutagenic activity for the target chemical according to studies compliant with the OECD 471 guideline's requirements [3]. As such, the default reliability score of RS5 is increased to RS1. To further confirm the negative outcome, the *in silico* predictions for m-xylylenediamine are next analyzed. The two complementary (Q)SAR methodologies provide consistent negative outcomes. Expert review sets the reliability score of these *in silico* results to RS3, whereas the overall negative bacterial mutation assessment that also accounts for the available experimental data (RS1) can be associated with an RS1 score.

For the mouse lymphoma assessment, the protocol shows that negative experimental data are available and *in silico* predictions are generated; an expert review concludes that there is sufficient evidence to increase the reliability of the experimental data to RS1 whereas the *in silico* predictions can be associated with an RS3 score. This is combined

in the protocol workflow into an overall negative mouse lymphoma assessment with a reliability score of RS1.

The bacterial mutation and the mouse lymphoma assessments are used to derive the overall negative gene mutation potential. Given the reliability scores that have been set during the expert review, the confidence of this negative result is automatically assigned by the protocol to "High" [16].

Next the clastogenicity / aneugenicity *in vitro* endpoint (see Fig. 17) is assessed by inspecting the available chromosome aberration *in vitro* experimental data and *in silico* results. An expert review confirms the initial negative assessment that is associated with an RS1 score. These results are used to derive the negative assessment for the clastogenicity / aneugenicity *in vitro* endpoint with a medium confidence due to the lack of information on micronucleus *in vitro*.

The next step consists in reviewing the assessment of the clastogenicity/aneugenicity *in vivo* potential (see Fig. 16). No experimental


Integrated hazard assessment		Individual models				
Regulatory protocol: ICH M7						
Structure	Laboratory Data	Expert rule based system	Statistical based system	M7 Class assessment	M7 Class confidence	Additional supportive evidence and comments
 1-chloro-2-nitrobenzene	Carcinogenicity: Positive Bacterial Mutation: Positive	Positive 22: aromatic nitro (0.91)	Positive PPP = 0.825	Class 1 - Mutagenic carcinogen	High	Laboratory carcinogenicity and mutagenicity data was used to determine the classification: mutagenic carcinogen.
<div style="display: flex; justify-content: space-around;"> Explain... Export Table... Generate Full Reports ICH M7 Summary Report </div>						

Fig. 14. Summary of the preliminary results of the ICH M7 hazard assessment of 1-chloro-2-nitrobenzene.

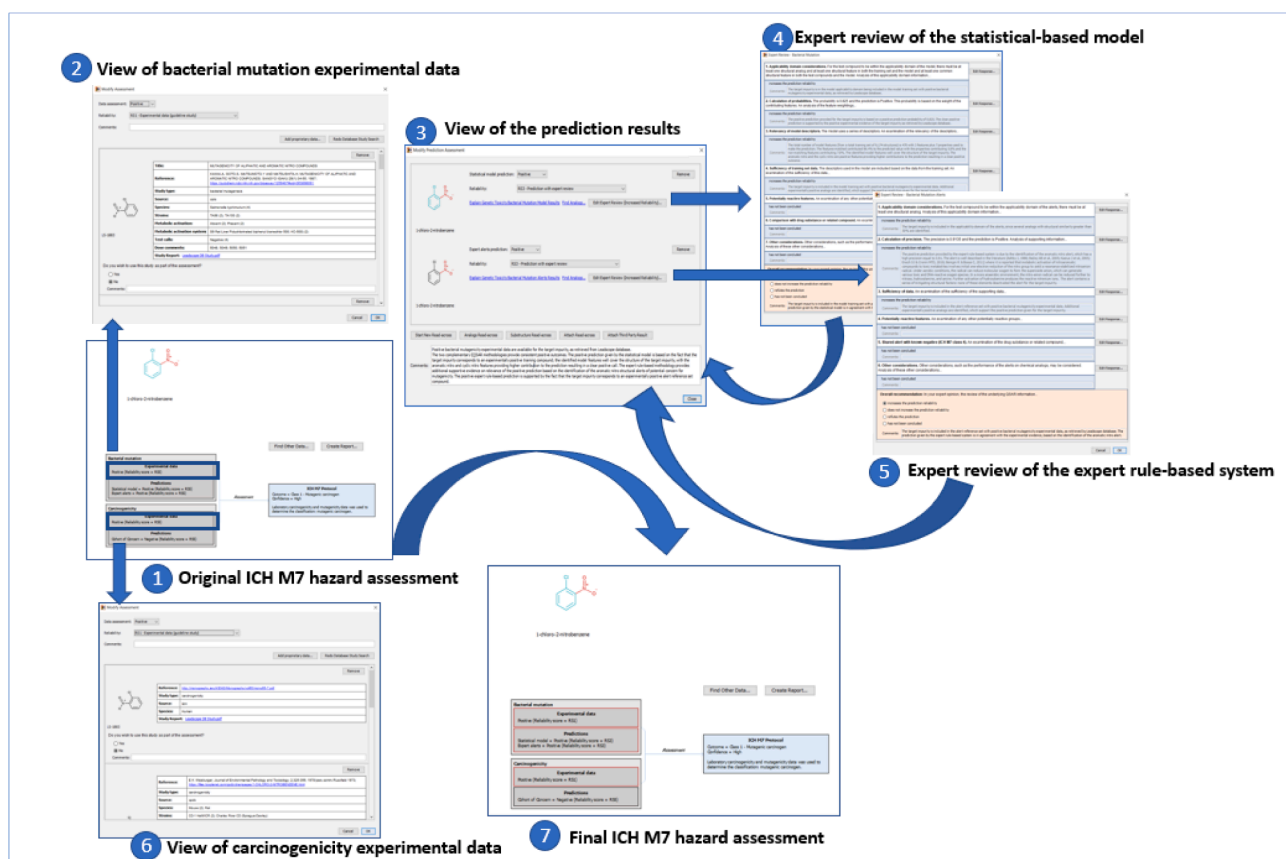


Fig. 15. ICH M7 hazard assessment workflow for 1-chloro-2-nitrobenzene.

evidence is available for m-xylylenediamine concerning chromosome aberration *in vivo*, whereas the *in silico* result (i.e., a statistical-based system) provides an out-of-domain result triggered by the absence of similar structures in the training set. Expert review of this prediction suggests that it is feasible to overturn the out-of-domain outcome into a negative call based on the following elements: a) a low positive prediction probability (PPP = 0.124); b) a good coverage of the m-xylylenediamine structure; c) negative features contributing to the prediction leading to a clear negative call; d) no features associated with a positive contribution. Since only one model is used to predict chromosome aberration *in vivo* without any sufficiently similar analogs, the reliability score of the chromosome aberration *in vivo* prediction is set to RS5. An expert review of the micronucleus *in vivo* experimental data and *in silico*

results increases the experimental reliability score to RS1 and overturns the out-of-domain prediction into a negative prediction with RS5 score (based on the good coverage of the structure of m-xylylenediamine in addition to the lack of any reactive potential). A negative assessment with RS1 score is then set for the micronucleus *in vivo* assessment. The results for chromosome aberration *in vivo* (negative RS5) and micronucleus *in vivo* (negative RS1) are combined for the negative assessment of the overall clastogenicity / aneugenicity *in vivo* potential, with corresponding "Medium" confidence set by the tool according to the protocol rules [16].

The clastogenicity / aneugenicity *in vitro* (negative, medium confidence) and *in vivo* (negative, medium confidence) sub-endpoints are then combined into a single clastogenicity / aneugenicity endpoint. This

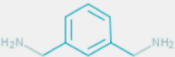
Integrated hazard assessment: Genetic Toxicity						
Structure	Genetic Toxicity assessment	Genetic Toxicity confidence	Gene Mutation assessment	Gene Mutation confidence	Clastogenicity / Aneugenicity assessment	Clastogenicity / Aneugenicity confidence
 m-xylylenediamine	Negative	Low	Negative	Medium	Negative	Low

Fig. 16. Summary of the preliminary results of the genetic toxicology protocol of m-xylylenediamine.

is assessed as negative with a “Medium” confidence as proposed by the tool.

Finally, all the results of the sub-endpoints (see Fig. 16) are automatically combined into a single overall negative assessment of “Genetic Toxicity” with a medium confidence score, confirmed by the expert review.

The genetic toxicology assessment is then saved in a report summarizing the results alongside the elements considered in the expert reviews. The report consists of a single editable word document including an executive summary (covering materials and methods used for the prediction of each effect/mechanism; any rules and principles used to combine the information; results for the individual effects/mechanisms and associated reliability scores; results for the endpoints and associated confidence scores) and the hazard assessment framework view (broken down into a series of graphs) and any comments included in the assessment. In addition, a zip file containing an appendix of information is created including full study and *in silico* reports for each individual prediction.

Case study 4: Skin sensitization assessment of bis-GMA

The following case study describes the assessment of bis-GMA (CAS 1565-94-2) using the implemented version of the skin sensitization protocol. The software returns an assessment of positive with high confidence for skin sensitization in humans. The main endpoints, “skin sensitization in rodents” and “skin sensitization *in vitro*” are assessed as positive with a high and low confidence levels, respectively. Fig. 18 provides an overview of the workflow used to derive the overall assessment.

The “Explain” function is used to understand the basis for the positive prediction and the confidence level. It is important to explore any experimental data or *in silico* predictions that are used in the assessment and how reliable the data are. *In silico* predictions are used to assess all the relevant mechanisms/effects while experimental data are available for the LLNA and h-CLAT assessments. A high-level overview shows that the h-CLAT experimental data disagrees with the positive LLNA experimental assessment. Further, the positive h-CLAT statistical model outcome does not support the negative experimental h-CLAT assessment. Both these results are initially assigned a reliability score of RS5 and the ‘Expression of co-stimulatory and adhesion molecules’ endpoint is left unassigned given conflicting assessments of the same reliabilities, Fig. 19. This prompts an expert review.

The skin sensitization protocol outlines factors that could lead to false negative results in the h-CLAT experimental system and discusses the exclusion of chemicals with a Log P value >3.5 from the applicability domain of the h-CLAT test [17,31]. The calculated ALogP value of bis-GMA (3.73) marginally falls within this range and a false negative experimental result is suspected. This non-applicability of the experimental system is reflected in the test guideline [32] and does not support

an increase in the reliability of the experimental h-CLAT data. An explanation of the positive statistical prediction shows that the result is within the applicability domain of the model with 10 structural features mapping to the Bis-GMA structure and a predicted probability of 0.79. The feature weighting, relevancy of the model descriptors and sufficiency of training set data are evaluated as part of the expert review process as prompted by the checklist of items to consider for an expert review, Fig. 20.

The feature which contributes most to the positive weight is the acrylate group. The activity of the training set examples cannot be explained through any moiety other than the (meth)acrylate feature, and a potential reaction mechanism could be postulated based on this feature. This supports the relevancy of the structural moiety and an increased prediction reliability. Of note however, the training set data are non-aromatic structures, which lack diversity and the influence of the aromatic ring on the activity cannot be adequately assessed. Overall, the expert review of the statistical model’s positive prediction confirms such result given the weighting of features and the mechanistic basis which could be attributed to the acrylate feature. The level of evidence supports an increase in reliability to an RS3 level for this positive result. Since the RS3 reliability is higher than the RS5 reliability of the experimental data, the positive prediction is used in the assessment. The “Events in Dendritic cells” endpoint automatically changes to a positive outcome, with a medium confidence level. Fig. 21 shows a manual override of the result, the associated documentation and the updated “Events in Dendritic cells” node.

After working through the assessment of the “Events in Dendritic cells” endpoint and understanding the negative experimental results and the evidence presented by the *in silico* methods, one of the two following approaches could be taken. Either, an evaluation of *in vitro* endpoints which were predicted by *in silico* models (“Covalent interaction with skin proteins” and “Events in Keratinocytes”) could be continued or any additional experimental data could be assessed. The latter approach is adopted here since a high-quality experimental result would be sufficient for a regulatory assessment, particularly where any conflicting information can be explained. An assessment of the rodent LLNA result is made based on an experimental study linked to the test structure contained in the database. This study indicates weak sensitization potential and is assigned a reliability score of 1. Clicking on the experimental results for the LLNA returns the studies that are available for expert review, including their references and a comment field to document any findings, Fig. 22.

The LLNA statistical model (predicting weak potency class) and expert alert results (acrylates and methacrylate alert matched) further support the positive assessment and potency classification. The assessment of the “skin sensitization in rodents” endpoint is therefore assessed as positive (weak potency) with high confidence. At this point, there does not appear to be conflicting evidence across the framework and the positive “sensitization *in vitro*” assessment supports the overall positive

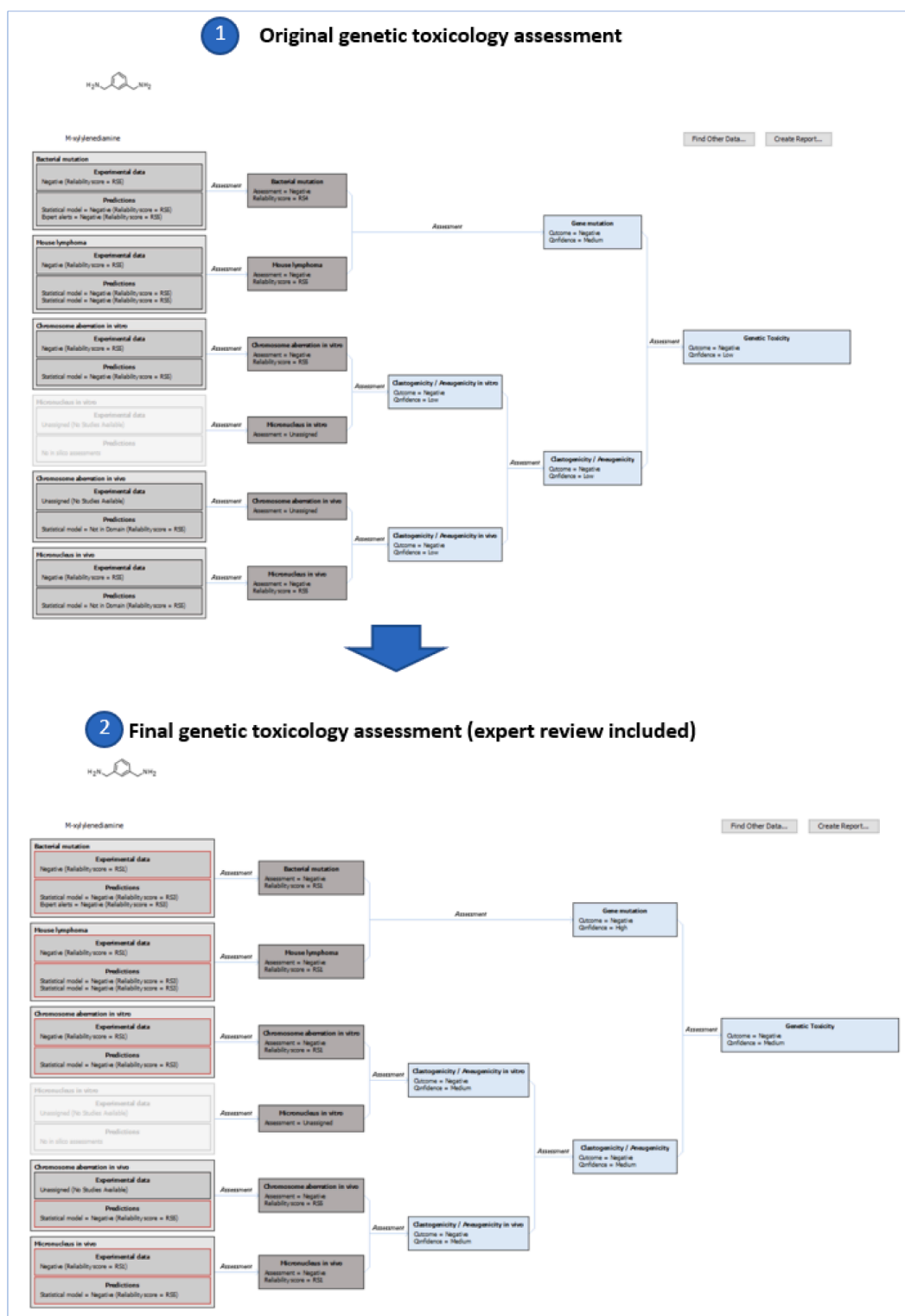


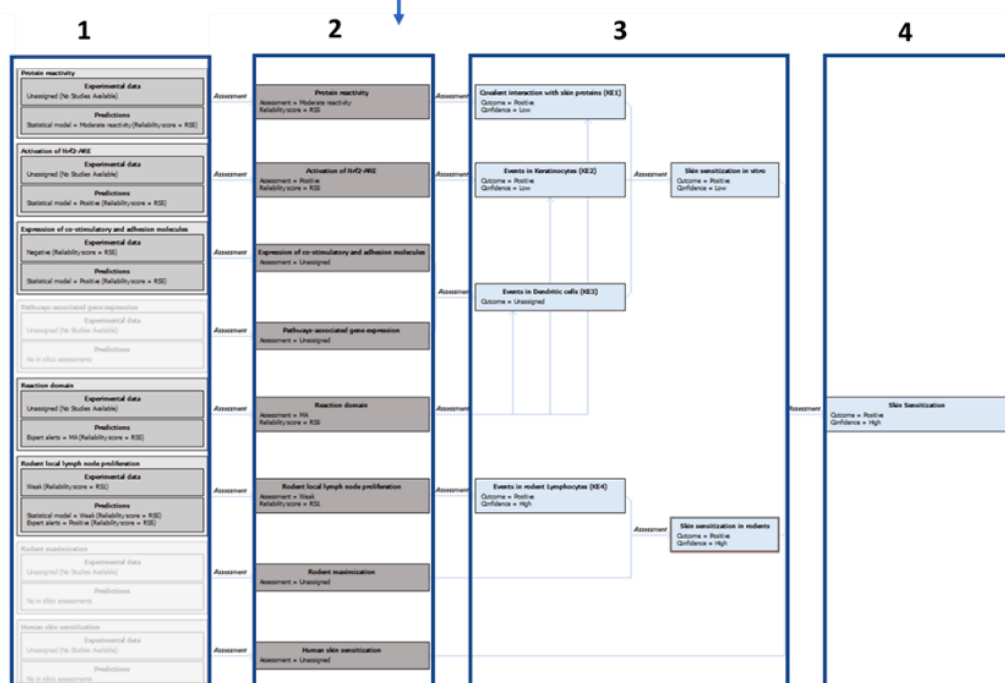
Fig. 17. Genetic toxicology assessment workflow for m-xylylenediamine. The original genetic toxicology assessment is reviewed by inspecting and analyzing: i) genetic mutation potential; ii) clastogenicity *in vitro*; iii) clastogenicity *in vivo*.

outcome with high confidence for skin sensitization hazard of Bis-GMA. However, it is prudent to review the statistical models predicting protein reactivity and activation of the Nrf2-ARE pathway (keratinocyte activation) to confirm that there is no conflicting information. Similar to the h-CLAT assessment, clicking on the nodes which contains the model predictions returns the explain model, find analogs and expert review fields (see Fig. 17). The DPRA model predicts moderate protein

reactivity with a probability of 0.921 which is driven primarily by the acrylate feature (see Fig. 23).

The result is similar to the Keratinosens™ statistical model's prediction (see Table 6 for additional information on the models). In both cases, the expert review supported an increase in reliability to an RS3 level and the "covalent interaction with skin proteins" and the "Events in Keratinocytes" endpoints are both assessed as positive with medium

Integrated hazard assessment		Individual models				
Integrated hazard assessment: Skin Sensitization						
Structure	Skin Sensitization in Humans assessment	Skin Sensitization in Humans confidence	Skin Sensitization in Rodents assessment	Skin Sensitization in Rodents confidence	Skin Sensitization in vitro assessment	Skin Sensitization in vitro confidence
Bis-GMA <chem>C=C1C=CC(=O)OC1=O</chem>	Positive	High	Positive	High	Positive	Low
<input type="button" value="Explain..."/> <input type="button" value="Export Table..."/> <input type="button" value="Generate Full Reports"/>						



1. Evaluate which mechanisms/effects are predicted and whether experimental data or *in silico* models are used for the assessment. Assess underlying data to evaluate reliability and relevance. Access the checklist of items to consider as part of the expert review and document findings.
2. Manually update the reliability scores or assessment for mechanisms/effects to reflect the expert review findings.
3. Review how the assessments of mechanisms and effects are used to derive the assessment of sub-endpoints and their associated confidence scores
4. Review how the assessments of the overall endpoint was made and how the assessments of the sub-endpoints were used to derive the overall assessment

Fig. 18. Workflow to derive an assessment of skin sensitization using the implemented skin sensitization protocol.

confidence. Together with the h-CLAT assessment, the “skin sensitization *in vitro*” endpoint is assessed as positive, with medium confidence and the system explains that this assessment is based on an encoded rule of “at least two positive assessments aligned” for the “sensitization *in vitro*” endpoint. The “sensitization *in vitro*” assessment supports the final assessment of positive, with high confidence which is based primarily on LLNA results as the key study.

Discussion

The tool described in this paper addresses many critical issues that enable the use of integrated approaches for toxicological hazard assessment to be used across industrial and regulatory applications. By being based on commonly agreed principles and procedures, such as *in*

silico toxicology protocols, the platform provides an approach that is defensible to colleagues and peers. By incorporating transparent metrics of reliability, relevance and confidence, the approach supports many different applications, from regulatory submissions to screening chemicals, that tolerate differing levels of uncertainty. The visual platform is transparent, clearly showing the steps in the assessment process, with details available at any stage on demand. The ability to interact with the platform supports a thorough expert review. Such a review may modify any conclusions on any of the effects, mechanisms, or derived toxicological endpoints. The deviations from the default assessment (described in the protocols) are recorded along with their justifications. Automatically documenting this entire process, including the source materials (experimental and *in silico* results) and the entire decision-making process, and tracking the expert review rationale ensure the results are

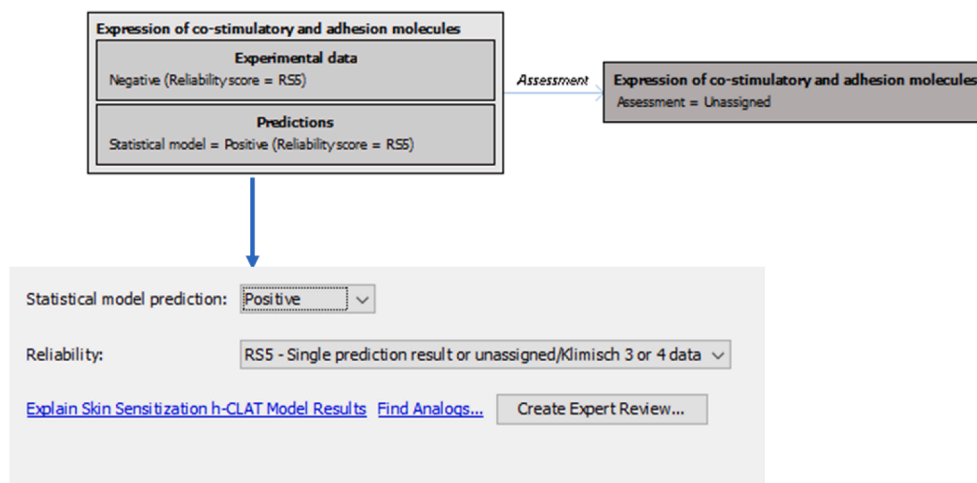


Fig. 19. Assessment node for the “Expression of co-stimulatory adhesion molecules” based on the h-CLAT method. Clicking on the result allows access to underlying information.

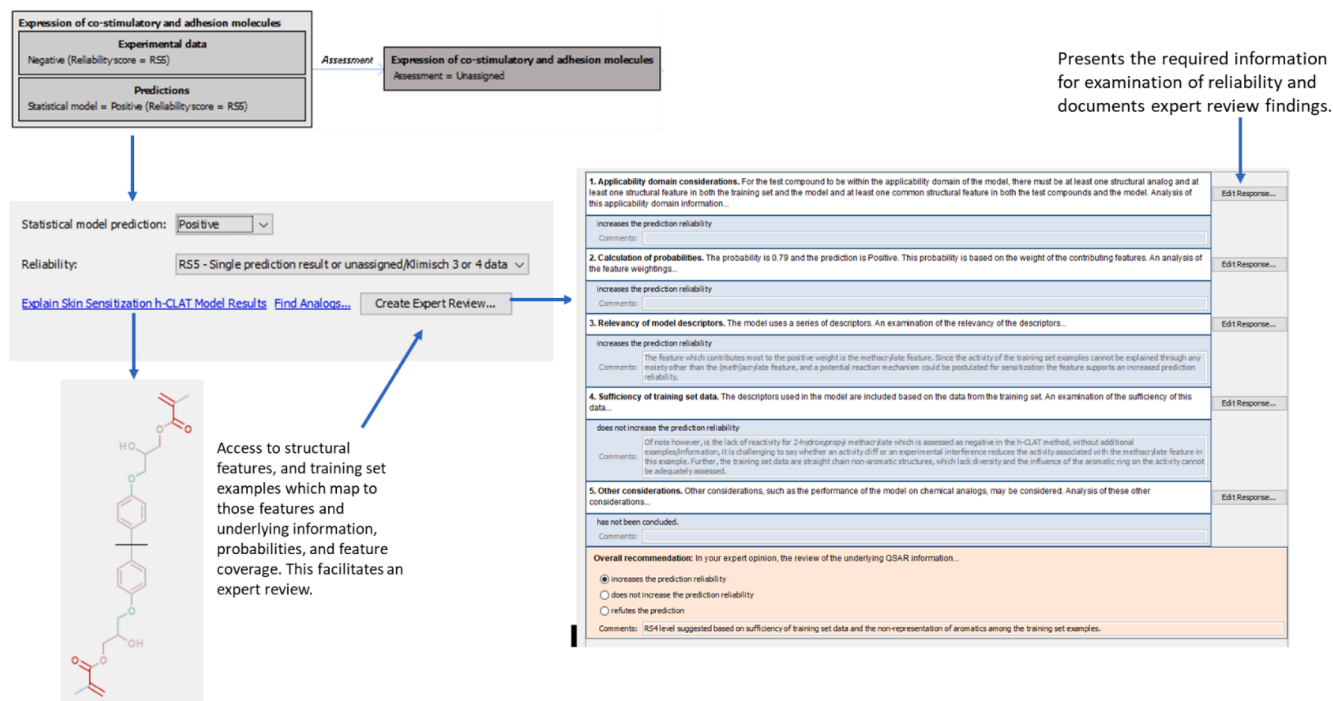


Fig. 20. Access to underlying information allows an expert review to be performed in a consistent manner.

effectively, completely, and consistently reported. At the same time, this automatic approach avoids any transcription errors. The documentation supports an outside review and would also enable a third party to repeat the process. The visual platform is based on a consistent organization and will support many different regulatory guidelines and protocols. Hence, this approach may be easily adopted when new regulations and *in silico* toxicology protocols are developed.

As mentioned earlier, an expert review plays an important role throughout the assessment process which is invariably subjective. To mitigate this concern, a series of guidelines and case studies were introduced and detailed in different *in silico* protocol-related papers [1,4,5,16,17]. Using these commonly agreed principles for performing such a review, a more consistent approach will support the application of *in silico* assessments across different regulatory frameworks and jurisdictions.

This paper has outlined a series of case studies based on publicly

available databases and *in silico* models. The framework described also supports the integration of proprietary experimental data to use in the assessment of individual effects/mechanisms. Proprietary experimental data on chemicals analogs can also be utilized in this framework, by introducing a read-across prediction and as part of an expert review of the *in silico* results. In addition, proprietary models can also be used to assess individual effects/mechanisms. This may be helpful when the test chemical falls outside the chemical space from which the public models were built. When such assessments are performed for external groups, such as regulatory agencies, it may be necessary to disclose the model's training set to be transparent. Such disclosure often makes the use of these proprietary models impractical. There are also approaches that avoid the use of proprietary data directly yet incorporate knowledge derived from proprietary database, such as the SAR fingerprinting approach. [30]

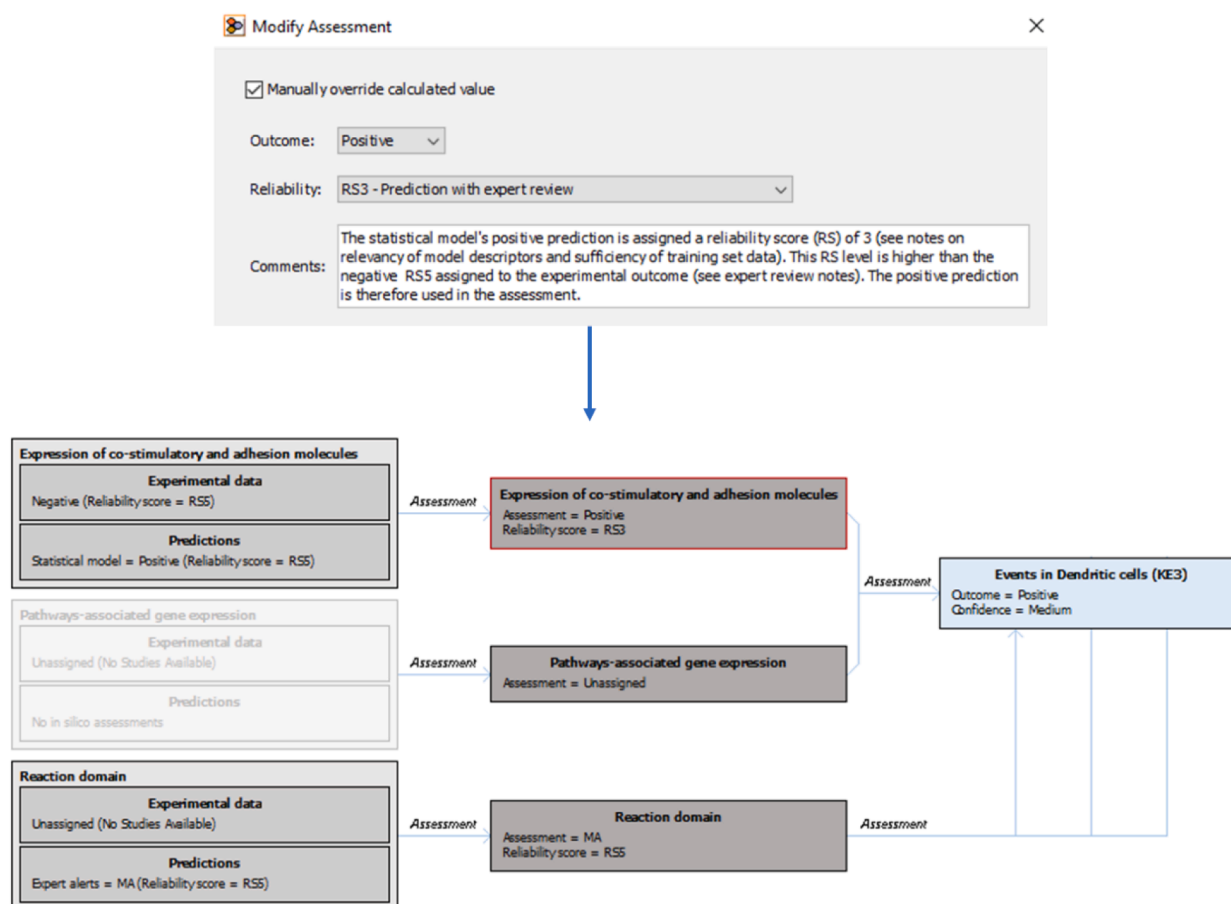


Fig. 21. Influence of expert review findings on the assessment of “Events in Dendritic cells” endpoint.

The screenshot shows a 'Modify Assessment' dialog box. The 'Sensitization' is set to 'Weak' and 'Reliability' is 'RS1 - Experimental data (guideline study)'. The 'Comments' field is empty. There are buttons for 'Add proprietary data...' and 'Redo Database Study Search'. A table provides details for a study call:

Study call:	Weak
Title:	NICEATM LLNA Potency Category
Reference:	https://ntp.niehs.nih.gov/whatwestudy/niceatm/test-method-evaluations/immunotoxicity/llna/index.html
Study type:	local lymph node assay
Dose comments:	NICEATM LLNA EC3 (%): 45.3

Below the table, there is a chemical structure and the text 'LS-123696'. A question asks 'Do you wish to use this study as part of the assessment?' with 'Yes' selected. The 'Comments' field contains 'Conducted according to guidelines.'

Fig. 22. Review of experimental study and assessment of reliability.

Conclusion

The integrated assessment of toxicological endpoints, where a battery of experimental and *in silico* methods are combined, is important to current and future toxicological hazard assessments. It provides a more mechanistically interpretable approach that also supports the 3Rs. The

successful application of such approaches to hazard assessment will require the adoption of quality-driven standards and processes. Tools that support such assessments in an efficient, transparent, defensible, and repeatable manner, such as the visual and interactive platform described in this paper, will be essential to support hazard assessment based on these new methods.

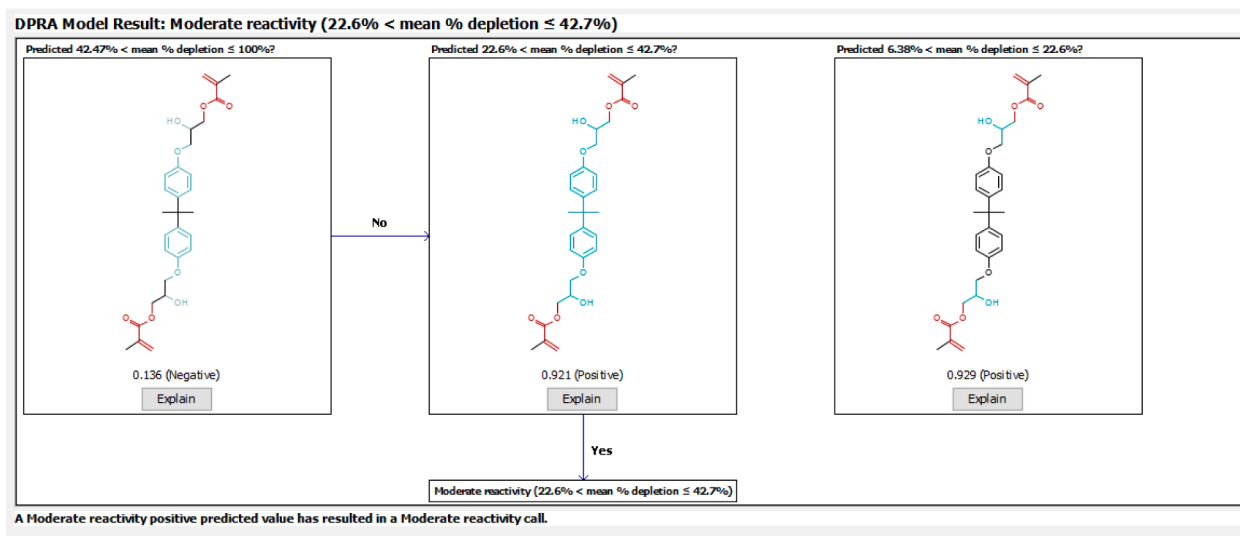


Fig. 23. A categorical model is used to predict the protein reactivity of Bis-GMA.

CRedit authorship contribution statement

Glenn J. Myatt: Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition. **Arianna Bassan:** Writing – original draft, Writing – review & editing. **Dave Bower:** Investigation, Writing – review & editing. **Candice Johnson:** Writing – original draft, Writing – review & editing. **Scott Miller:** Software, Writing – review & editing. **Manuela Pavan:** Writing – original draft, Writing – review & editing. **Kevin P. Cross:** Conceptualization, Software, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number R44ES026909. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- [1] G.J. Myatt, E. Ahlberg, Y. Akahori, D. Allen, A. Amberg, L.T. Anger, A. Aptula, S. Auerbach, L. Beilke, P. Bellion, R. Benigni, J. Bercu, E.D. Booth, D. Bower, A. Brigo, N. Burden, Z. Cammerer, M.T.D. Cronin, K.P. Cross, L. Custer, M. Dettwiler, K. Dobo, K.A. Ford, M.C. Fortin, S.E. Gad-McDonald, N. Gellatly, V. Gervais, K.P. Glover, S. Glowienke, J. Van Gompel, S. Gutsell, B. Hardy, J. S. Harvey, J. Hillegass, M. Honma, J.-H. Hsieh, C.-W. Hsu, K. Hughes, C. Johnson, R. Jolly, D. Jones, R. Kemper, M.O. Kenyon, M.T. Kim, N.L. Kruhlak, S.A. Kulkarni, K. Kümmerer, P. Leavitt, B. Majer, S. Masten, S. Miller, J. Moser, M. Mumtaz, W. Muster, L. Neilson, T.I. Oprea, G. Patlewicz, A. Paulino, E. Lo Piparo, M. Powley, D.P. Quigley, M.V. Reddy, A.-N. Richarz, P. Ruiz, B. Schilter, R. Serafimova, W. Simpson, L. Stavitskaya, R. Stidl, D. Suarez-Rodriguez, D. T. Szabo, A. Teasdale, A. Trejo-Martin, J.-P. Valentin, A. Vuorinen, B.A. Wall, P. Watts, A.T. White, J. Wichard, K.L. Witt, A. Woolley, D. Woolley, C. Zwickl, C. Hasselgren, *In silico* toxicology protocols, *Regul. Toxicol. Pharmacol.* 96 (2018) 1–17, <https://doi.org/10.1016/j.yrtph.2018.04.014>.
- [2] ICH, M7 (R1) Assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk, European Medicines Agency, 2017. https://database.ich.org/sites/default/files/M7_R1_Guideline.pdf.
- [3] OECD, Test No. 471: Bacterial Reverse Mutation Test, OECD Publishing, Paris, 1997. Doi:10.1787/9789264071247-en.
- [4] A. Amberg, R.V. Andaya, L.T. Anger, C. Barber, L. Beilke, J. Bercu, D. Bower, A. Brigo, Z. Cammerer, K.P. Cross, L. Custer, K. Dobo, H. Gerets, V. Gervais, S. Glowienke, S. Gomez, J. Van Gompel, J. Harvey, C. Hasselgren, M. Honma, C. Johnson, R. Jolly, R. Kemper, M. Kenyon, N. Kruhlak, P. Leavitt, S. Miller, W. Muster, R. Naven, J. Nicolette, A. Parenty, M. Powley, D.P. Quigley, M. V. Reddy, J.C. Sasaki, L. Stavitskaya, A. Teasdale, A. Trejo-Martin, S. Weiner, D. S. Welch, A. White, J. Wichard, D. Woolley, G.J. Myatt, Principles and procedures for handling out-of-domain and indeterminate results as part of ICH M7 recommended (Q)SAR analyses, *Regul. Toxicol. Pharmacol.* 102 (2019) 53–64, <https://doi.org/10.1016/j.yrtph.2018.12.007>.
- [5] A. Amberg, L. Beilke, J. Bercu, D. Bower, A. Brigo, K.P. Cross, L. Custer, K. Dobo, E. Dowdy, K.A. Ford, S. Glowienke, J. Van Gompel, J. Harvey, C. Hasselgren, M. Honma, R. Jolly, R. Kemper, M. Kenyon, N. Kruhlak, P. Leavitt, S. Miller, W. Muster, J. Nicolette, A. Plaper, M. Powley, D.P. Quigley, M.V. Reddy, H.-P. Spirkel, L. Stavitskaya, A. Teasdale, S. Weiner, D.S. Welch, A. White, J. Wichard, G.J. Myatt, Principles and procedures for implementation of ICH M7 recommended (Q)SAR analyses, *Regul. Toxicol. Pharmacol.* 77 (2016) 13–24, <https://doi.org/10.1016/j.yrtph.2016.02.004>.
- [6] K.L. Dobo, N. Greene, C. Fred, S. Glowienke, J.S. Harvey, C. Hasselgren, R. Jolly, M.O. Kenyon, J.B. Munzner, W. Muster, R. Neft, M. Vijayaraj Reddy, A.T. White, S. Weiner, *In silico* methods combined with expert knowledge rule out mutagenic potential of pharmaceutical impurities: an industry survey, *Regul. Toxicol. Pharmacol.* 62 (3) (2012) 449–455, <https://doi.org/10.1016/j.yrtph.2012.01.007>.
- [7] C. Barber, A. Amberg, L. Custer, K.L. Dobo, S. Glowienke, J. Van Gompel, S. Gutsell, J. Harvey, M. Honma, M.O. Kenyon, N. Kruhlak, W. Muster, L. Stavitskaya, A. Teasdale, J. Vessey, J. Wichard, Establishing best practise in the application of expert review of mutagenicity under ICH M7, *Regul. Toxicol. Pharmacol.* 73 (1) (2015) 367–377, <https://doi.org/10.1016/j.yrtph.2015.07.018>.
- [8] M.W. Powley, (Q)SAR assessments of potentially mutagenic impurities: a regulatory perspective on the utility of expert knowledge and data submission, *Regul. Toxicol. Pharmacol.* 71 (2) (2015) 295–300, <https://doi.org/10.1016/j.yrtph.2014.12.012>.
- [9] C.M. Ellison, P. Piechota, J.C. Madden, S.J. Enoch, M.T.D. Cronin, Adverse Outcome Pathway (AOP) informed modeling of aquatic toxicology: QSARs, read-across, and interspecies verification of modes of action, *Environ. Sci. Technol.* 50 (7) (2016) 3995–4007, <https://doi.org/10.1021/acs.est.5b05918>.
- [10] T.M. Martin, D.M. Young, C.R. Lilavois, M.G. Barron, Comparison of global and mode of action-based models for aquatic toxicity, *SAR QSAR Environ. Res.* 26 (3) (2015) 245–262, <https://doi.org/10.1080/1062936X.2015.1018939>.
- [11] OECD, Guidance Document on the Reporting of Defined Approaches to Be Used within Integrated Approaches to Testing and Assessment, OECD Environment, Health and Safety Publications, Paris, 2016. doi:10.1787/9789264274822-en.
- [12] OECD, Guidance Document on the Reporting of Defined Approaches and Individual Information Sources to be Used within Integrated Approaches to Testing and Assessment (IATA) for Skin Sensitisation, OECD Environment, Health and Safety Publications, Paris, 2016. doi:10.1787/9789264279285-en.
- [13] US EPA, Integrated Risk Information System, US EPA. (2021). <https://www.epa.gov/assessing-and-managing-chemicals-under-tscs/strategic-plan-reduce-use-ver-tebrate-animals-chemical> (accessed April 9, 2021).
- [14] OECD, OECD Test Guidelines for Chemicals, (2021). <https://www.oecd.org/chemicalsafety/testing/oecdguidelinesforthetestingofchemicals.htm> (accessed April 22, 2021).
- [15] OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Environment, Health and Safety Publications, Paris, 2007. doi:10.1787/9789264085442-en.
- [16] C. Hasselgren, E. Ahlberg, Y. Akahori, A. Amberg, L.T. Anger, F. Atienzar, S. Auerbach, L. Beilke, P. Bellion, R. Benigni, J. Bercu, E.D. Booth, D. Bower,

- A. Brigo, Z. Cammerer, M.T.D. Cronin, I. Crooks, K.P. Cross, L. Custer, K. Dobo, T. Doktorova, D. Faulkner, K.A. Ford, M.C. Fortin, M. Frericks, S.E. Gad-McDonald, N. Gellatly, H. Gerets, V. Gervais, S. Glowienke, J. Van Gompel, J.S. Harvey, J. Hillegass, M. Honma, J.H. Hsieh, C.W. Hsu, T.S. Barton-Maclaren, C. Johnson, R. Jolly, D. Jones, R. Kemper, M.O. Kenyon, N.L. Kruhlak, S.A. Kulkarni, K. Kümmerer, P. Leavitt, S. Masten, S. Miller, C. Moudgal, W. Muster, A. Paulino, E. Lo Piparo, M. Powley, D.P. Quigley, M.V. Reddy, A.N. Richarz, B. Schilter, R. D. Snyder, L. Stavitskaya, R. Stidl, D.T. Szabo, A. Teasdale, R.R. Tice, A. Trejo-Martin, A. Vuorinen, B.A. Wall, P. Watts, A.T. White, J. Wichard, K.L. Witt, A. Woolley, D. Woolley, C. Zwickl, G.J. Myatt, Genetic toxicology in silico protocol, *Regul. Toxicol. Pharmacol.* 107 (2019) 104403, <https://doi.org/10.1016/j.yrtph.2019.104403>.
- [17] C. Johnson, E. Ahlberg, L.T. Anger, L. Beilke, R. Benigni, J. Bercu, S. Bobst, D. Bower, A. Brigo, S. Campbell, M.T.D. Cronin, I. Crooks, K.P. Cross, T. Doktorova, T. Exner, D. Faulkner, I.M. Fearon, M. Fehr, S.C. Gad, V. Gervais, A. Giddings, S. Glowienke, B. Hardy, C. Hasselgren, J. Hillegass, R. Jolly, E. Krupp, L. Lomnitski, J. Magby, J. Mestres, L. Milchak, S. Miller, W. Muster, L. Neilson, R. Parakhia, A. Parenty, P. Parris, A. Paulino, A.T. Paulino, D.W. Roberts, H. Schlecker, R. Stidl, D. Suarez-Rodriguez, D.T. Szabo, R.R. Tice, D. Urbisch, A. Vuorinen, B. Wall, T. Weiler, A.T. White, J. Whitenour, J. Wichard, D. Woolley, C. Zwickl, G.J. Myatt, Skin sensitization in silico protocol, *Regul. Toxicol. Pharmacol.* 116 (2020) 104688, <https://doi.org/10.1016/j.yrtph.2020.104688>.
- [18] NTP, National Toxicology Program U.S. Department of Health and Human Services - Data & Resources, (2021). <https://ntp.niehs.nih.gov/data/index.html>.
- [19] NIH, Download Carcinogenic Potency Database (CPDB) Data, (2021). <https://www.nlm.nih.gov/databases/download/cpdb.html> (accessed June 11, 2021).
- [20] D. Bower, K. Cross, G. Myatt, Organisation of Toxicological Data in Databases, in: D. Neagu, A.-N. Richarz (Eds.), *Big Data in Predictive Toxicology*, Royal Society of Chemistry, Cambridge, 2020, pp. 108–165, <https://doi.org/10.1039/9781782623656-00108>.
- [21] Leadscope, Instem - Computational Toxicology, 2021. <https://www.instem.com/solutions/insilico/computational-toxicology.php>.
- [22] G. Roberts, G.J. Myatt, W.P. Johnson, K.P. Cross, P.E. Blower, LeadScope: software for exploring large sets of screening data, *J. Chem. Inf. Comput. Sci.* 40 (2000) 1302–1314, <https://doi.org/10.1021/ci0000631>.
- [23] K.P. Cross, G. Myatt, C. Yang, M.A. Fligner, J.S. Verducci, P.E. Blower, Finding discriminating structural features by reassembling common building blocks, *J. Med. Chem.* 46 (22) (2003) 4770–4775, <https://doi.org/10.1021/jm0302703>.
- [24] C. Yang, K. Cross, G.J. Myatt, P.E. Blower, J.F. Rathman, Building predictive models for protein tyrosine phosphatase 1B inhibitors based on discriminating structural features by reassembling medicinal chemistry building blocks, *J. Med. Chem.* 47 (24) (2004) 5984–5994, <https://doi.org/10.1021/jm0497242>.
- [25] P. Blower, K. Cross, M. Fligner, G. Myatt, J. Verducci, C. Yang, Systematic analysis of large screening sets in drug discovery, *Curr. Drug Discovery Technol.* 1 (2004) 37–47, <https://doi.org/10.2174/1570163043484879>.
- [26] S.J. Enoch, M.T.D. Cronin, C.M. Ellison, The use of a chemistry-based profiler for covalent DNA binding in the development of chemical categories for read-across for genotoxicity, *Altern. Lab. Anim.* 39 (2) (2011) 131–145, <https://doi.org/10.1177/026119291103900206>.
- [27] A.O. Aptula, D.W. Roberts, Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: general principles and application to reactive toxicity, *Chem. Res. Toxicol.* 19 (2006) 1097–1105, <https://doi.org/10.1021/tx0601004>.
- [28] ECHA, Read-Across Assessment Framework (RAAF), 2017. doi:10.2823/619212.
- [29] B. Testa, J.M. Mayer, *Hydrolysis in Drug and Prodrug Metabolism: Chemistry, Biochemistry, and Enzymology*, 1st ed., John Wiley & Sons, Ltd, 2003. doi: 10.1002/9783906390444.
- [30] E. Ahlberg, A. Amberg, L.D. Beilke, D. Bower, K.P. Cross, L. Custer, K.A. Ford, J. Van Gompel, J. Harvey, M. Honma, R. Jolly, E. Joossens, R.A. Kemper, M. Kenyon, N. Kruhlak, L. Kuhnke, P. Leavitt, R. Naven, C. Neilan, D.P. Quigley, D. Shuey, H.-P. Spirkl, L. Stavitskaya, A. Teasdale, A. White, J. Wichard, C. Zwickl, G.J. Myatt, Extending (Q)SARs to incorporate proprietary knowledge for regulatory purposes: a case study using aromatic amine mutagenicity, *Regul. Toxicol. Pharmacol.* 77 (2016) 1–12, <https://doi.org/10.1016/j.yrtph.2016.02.003>.
- [31] O. Takenouchi, M. Miyazawa, K. Saito, T. Ashikaga, H. Sakaguchi, Predictive performance of the human Cell Line Activation Test (h-CLAT) for lipophilic chemicals with high octanol-water partition coefficients, *J. Toxicol. Sci.* 38 (4) (2013) 599–609, <https://doi.org/10.2131/jts.38.599>.
- [32] OECD, Test No. 442E: In Vitro Skin Sensitisation: In Vitro Skin Sensitisation assays addressing the Key Event on activation of dendritic cells on the Adverse Outcome Pathway for Skin Sensitisation, OECD Publishing, Paris, 2018. doi:10.1787/9789264264359-en.
- [33] C. Landry, M.T. Kim, N.L. Kruhlak, K.P. Cross, R. Saiakhov, S. Chakravarti, L. Stavitskaya, Transitioning to composite bacterial mutagenicity models in ICH M7 (Q)SAR analyses, *Regul. Toxicol. Pharmacol.* 109 (2019) 104488, <https://doi.org/10.1016/j.yrtph.2019.104488>. In press.
- [34] J.W. Yoo, N.L. Kruhlak, C. Landry, K.P. Cross, A. Sedykh, L. Stavitskaya, Development of improved QSAR models for predicting the outcome of the in vivo micronucleus genetic toxicity assay, *Regul. Toxicol. Pharmacol.* 113 (2020) 104620, <https://doi.org/10.1016/j.yrtph.2020.104620>.