(nor indeed any specific approach) but seeks to identify questions that experts should consider when using any predictive model for mutagenicity for application under ICH M7.

Key prerequisites of *in silico* systems that are suitable for regulatory purposes have been previously formulated in the OECD (Q)SAR guidelines ("OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationships models.," 2004). These are summarised in Box 1. Whilst excellent starting points when considering if a model is appropriate, accepting an individual prediction makes additional demands of the model. In order that an expert can effectively review the output from a model, it is important to assess performance against a (validation) test set covering similar chemical space to the compounds that are to be predicted. Such an assessment should also include coverage (the frequency with which a model makes a prediction), and in the occasions where the model cannot make a prediction; there should be a transparent and scientifically rational explanation for when a compound is outside the domain of the model. In addition, most models now provide a measure of confidence (expected accuracy) for each prediction. The approach used and the output generated can vary depending upon the underlying model methodology but irrespective of the approach, the output should be scientifically robust, transparent and accurate. If such methods do not correctly identify occasions where uncertainty is higher and describe why, then they offer little value to the user. Transparency is also key for any prediction since it is hard to review the output of a model that is not entirely comprehensible to the user. Any model that demands of the user to trust the output 'because it is normally right based upon a test set' cannot be effectively challenged by the user and accepting an output 'on faith' is not considered expert review. This consideration and the wording of M7 in terms of relating activity to structure will drive the approaches and choice of descriptors that can be considered acceptable for models that support regulatory submission, although such 'black box' models could still prove of value in other decision-making contexts. A number of possible challenges that an expert could consider when reviewing *in silico* predictions are described in Box 2.

The use of two orthogonal models has been shown to increase sensitivity and coverage (Hillebrecht et al., 2011; Naven et al., 2012). There are several reasons for this. Firstly, expert rule-based systems can absorb the tacit knowledge of the modeller which ideally spans both the biological domain (knowledge of the mechanism of action, route of metabolic activation.), and the chemistry domain (inherent reactivity of a functional group). This effectively allows generalisation from what is known, to a wider context than one could expect from a machine-learnt statistical correlation. Similarly, an expert may be able to tease apart apparently conflicting data that would lead to a statistical approach to see no correlation.[2] In addition, expert rule-based systems can absorb confidential data and still expose sufficient knowledge to allow clear interpretation – something that statistical systems struggle to do if they are to remain transparent. This can be significant; for example, 25% of alerts in one expert rule-based system (Derek Nexus [3]) have been derived from donated confidential data. On the other hand, machine-learnt statistical systems may identify new correlations that have not yet been identified by an expert, and they can potentially retain information from small numbers of compounds that the expert found insufficient from which to construct an alert.

Box 1
OECD Guidelines for the assessment of (Q)SAR models for regulatory use.

> *"To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:*
>
> 1. *a defined endpoint;*
> 2. *an unambiguous algorithm;*
> 3. *a defined domain of applicability;*
> 4. *appropriate measures of goodness-of-fit, robustness and predictivity;*
> 5. *a mechanistic interpretation, if possible."*
>
> ("OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationships models.," 2004)

4. How often do models disagree, and when they do, which one is right?

When two orthogonal models are used to generate predictions they will not always concur; indeed, if they did, then there would be no value in the use of more than one system. While the use of two systems can improve sensitivity, the examination of conflicting predictions can become more burdensome unless the systems have been designed to provide information that supports that analysis. Analysis of Derek Nexus (an expert rule-based system) and Sarah Nexus[3] (a statistical system) against a number of public and proprietary datasets showed that they agreed 70–85% of the time and that when they agreed, accuracy was as high as 90% [Fig. 1]. This is consistent with the reported reproducibility of the *in vitro* assay that the models were built from (Bentzien et al., 2010). When the models disagreed, no one system was clearly superior to the other. Fig. 2 shows the performance of both models against two datasets, one public and one from a data sharing consortium[4] showing that when the two systems disagree, neither system is superior.[5] Attempts to create some machine-learnt rules to guide the user through these 'conflicts' demonstrated the value of confidence or likelihood scores (Judson et al., 2013) in that the less confident prediction was more often incorrect, however the final decision could not be automated and still required expert review (Barber et al., 2014).

[2] For example, the activity observed with acid chlorides in the Ames assay can depend upon the choice of solvent. In aqueous solvents, acid chlorides can hydrolyse to chemically unreactive carboxylic acids and in neutral solvents like acetonitrile they tend to show no activity, whereas direct reaction with DMSO can give rise to highly reactive alkylating species(Cocivera et al., 1978). A statistical system is unlikely to be able to distinguish such subtle effects when learning across a range of compounds tested in the presence of a variety of solvents.

[3] Lhasa Limited, Leeds, UK.

[4] from the intermediates data sharing consortium of 11 companies led by Lhasa Limited.

[5] The dip in predictive performance for the consortium dataset is believed to be driven by the bias of the dataset since the consortium chose to test and share the more challenging compound classes (such as aromatic amines).

[6] This requires the user to be convinced by the model's approach and transparency of any applicability domain definition.

[7] There are a number of possible reasons to challenge the underlying experimental data. The Ames assay has been estimated to be ~85% reproducible for a number of reasons including poor purity - which could drive a false positive result, or low concentration or poor solubility - both of which could yield a false negative result. A compound may have been tested multiple times giving inconsistent results, not tested against 5-strains, tested using non-standard conditions or strains. Data may have been generated in the absence of metabolic activation or using an unusual source for metabolic activation. The data may not be relevant if for example the compound can release histamine (which will then give a positive result which is not driven by mutagenicity).

Box 2
Possible challenges to an *in silico* prediction.

- Expert rule-based system disagrees with statistical system because it uses data or knowledge that is not covered by DNA-reactivity (that can be machine-learnt).
  - *An expert rule-based system may extrapolate based upon functional group reactivity for which there is little or conflicting Ames data.*
- The model has not seen all the fragments of the query compound in a similar context.
  - *Is the query compound in the applicability domain of the model?*[6]
  - *Are there any mis- or unclassified features that could cause concern?*
  - *Insufficient number of relevant training examples*
- Training examples used to derive the prediction are not relevant to the query compound
  - *Positive examples have other expected causes for activity*
  - *Negative examples have known deactivating features that are not present in the query compound*
  - *Experimental data for the training examples is questionable, or incomplete*[7]
- There are more relevant examples that the model did not consider relevant, or are not in the training set
  - *The disclosure of proprietary data may support dismissing an in silico prediction*
- There are close analogues that are incorrectly predicted by the model
  - *eg Muller Class 4 compounds.*
  - *Have the same fragments in a similar environment given incorrect predictions ('misclassified features')?*
- There is relevant data for a fragment that causes a model to return 'out-of-domain'
  - *Evidence that the feature driving an out-of-domain prediction is negative in conjunction with the model's clear assertion that there are no other causes of activity.*
- The alert been poorly constructed?
  - *This may be plausible in complex areas of chemical space where multiple mechanisms can drive activity. Such a challenge would need to be strong and supported by robust and relevant data to override a positive prediction.*
- There are stereoelectronic arguments preventing the mechanism from taking place
  - *This will be hard to use against a positive prediction since it will require a high level of mechanistic knowledge and much supporting data.*

5. Undertaking the expert review

The expert review of a set of predictions does not necessarily need to be onerous. When both *in silico* predictions are in agreement, then a simple review of the information provided by the models and a cursory assessment of the query compound and of closely related known examples may be sufficient to concur with the predictions. Confidence scores or likelihood categories can indicate when a model has less confidence in the outcome which may be the result of a lack of close analogues in the training set or if those close analogues have differing activities. To be of value to the user, those reasons and the structures of close analogues must be

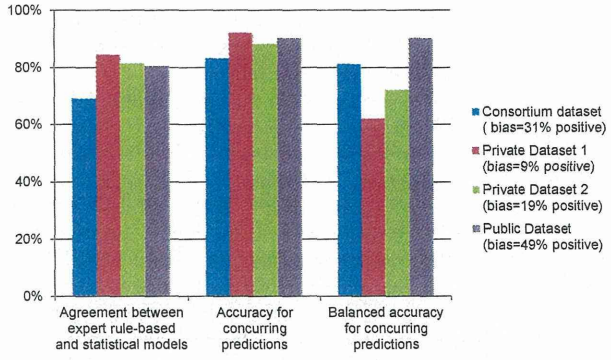

Fig. 1. Frequency and accuracy of concordant predictions from an expert rule-based system (Derek Nexus) and a statistical model (Sarah Nexus) against non-overlapping public and proprietary datasets. [Accuracy = true predictions/all predictions. Balanced accuracy = (sensitivity + specificity)/2].

readily accessible. Some models can explicitly identify reasons for uncertainty and draw attention to specific regions of the molecule that should be assessed more carefully. In the case that no alerts are triggered, a negative prediction can be concluded in accordance to the M7 guidelines which states that "The absence of structural alerts from two complementary (Q)SAR methodologies (expert rule-based and statistical) is sufficient to conclude that the impurity is of no mutagenic concern." This conclusion should not automatically be drawn should a model return 'indeterminate' or 'out of domain').

In the case of Derek Nexus, a further analysis is automatically undertaken looking for "unclassified" or "misclassified" features (Williams and Stalford, 2014). A feature is highlighted as unclassified if it is not present in a similar environment in any compounds contained within a large, curated reference dataset of public compounds. The absence of a feature in a public dataset does not mean that there is any evidence for a positive prediction, but merely that there are less data on which to make a negative prediction. Indeed the alert writer may be aware of proprietary data that cannot be disclosed to the user and are not used for this automated assessment. The fragmentation methodology has been optimised to recognise features that may be presented in an unusual environment (a potential cause of uncertainty), but one which an expert

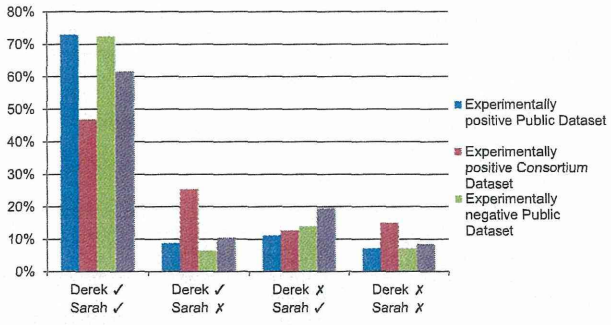

Fig. 2. Analysis of the performance of an expert rule-based (Derek) and a statistical system (Sarah) when used in combination. [Public dataset 10,775 compounds, 49% positive; Consortium dataset 817 compounds, 31% positive. Test compounds were not present in the statistical model training set] Key: Derek ✓ – Derek correctly predicted the outcome, Derek ✗ – Derek incorrectly predicted the outcome. So Derek ✓ Sarah ✗ – the two systems disagreed in their prediction with Derek being correct and Sarah being incorrect.

can often quickly review. Misclassified features are defined as features that are also present in a known positive compound for which no alert was raised by Derek Nexus. There are several reasons why these may arise. A feature may be highlighted based upon knowledge of a single positive experimental result that was dismissed by the expert rule writer, or by it being present in an Ames positive compound with another more plausible reason for activity. Essentially it attempts to emulate an expert who may say – "I predict negative, but I've seen a compound containing this fragment which was reported as active and for which no alert fired". Those 'predicted negative but observed positive' compounds are shown to the user who can then judge whether those compounds give sufficient grounds to challenge the negative prediction. Statistical systems can also provide information to guide the expert. For example, Sarah Nexus exposes machine-learnt knowledge within a self-organising hierarchical network which allows for the organisation of training data in nodes or clusters of compounds sharing common features (Hanser et al., 2014). When a query compound contains a fragment that defines a cluster, this is shown to the user along with those underlying training compounds. If the most similar examples in a cluster have a different signal to the cluster as a whole, there is more uncertainty in the contribution of that feature to the prediction and this is also indicated to the user.

Fig. 3 shows the possible outputs from two systems and the likely conclusions that an expert may come to. In the event of concordant predictions, it is anticipated that the expert's task is likely to be relatively simple. Knowledge of the mechanisms of DNA reactivity and an understanding of chemical reactivity benefits this analysis by helping the user to focus attention upon specific parts of a molecule. If there is some uncertainty in the expert's mind, whether driven by knowledge, experience, or specific structural flags highlighted by a model, then some examination is of benefit, although this need not necessarily be extensive nor documented in detail. Statistical systems that group supporting examples by common fragments are particularly effective in helping the expert identify relevant analogues during this stage. When systems disagree or fail to predict, then a greater level of assessment by the expert is expected and in this case, some documentation of the analysis and conclusions is anticipated. A decision to consider the impurity as positive is not likely to need significant justification since this is a conservative conclusion that will ensure that any perceived risk is appropriately managed. When the systems provide conflicting or uncertain predictions, then the decision to conclude a compound is negative demands a higher level of confidence and a greater level of detail should be documented.

6. Challenging a prediction from an expert rule-based system

Expert rule-based systems are human-defined rules created by scientists with expertise in the endpoint. These are often presented as fragments or more generically as patterns (Markush structures) possibly with additional relevant descriptors for the fragment or the whole compound (e.g. logP). Such systems have the advantage of being easily understood by the expert.

When a positive prediction is made, if that alert is observed to also fire for closely related known negative compounds containing the same feature in the same context (stereoelectronic environment), then an argument may be possible - the rule-writer has not completely captured the significance of this environment to the activity of this feature. In the M7 guidelines, this is specifically identified as a 'Class 4' when the known negative compound is the drug substance or compound related to the drug substance (Müller et al., 2006). In many cases these may be proprietary compounds generated and tested during a synthetic process. The compound(s) should not be likely to have differing activities by virtue of differences in chemical reactivity, activating metabolism, or the presence of deactivating features. This same argument could be applied to undermine a negative prediction should positive compounds containing the feature be known, particularly if the feature is presented in the same environment (a so called 'misclassified' feature).

A positive prediction could be challenged if the alert is not well defined and there are strong reactivity or stereoelectronic arguments that show the proposed mechanism of reaction cannot

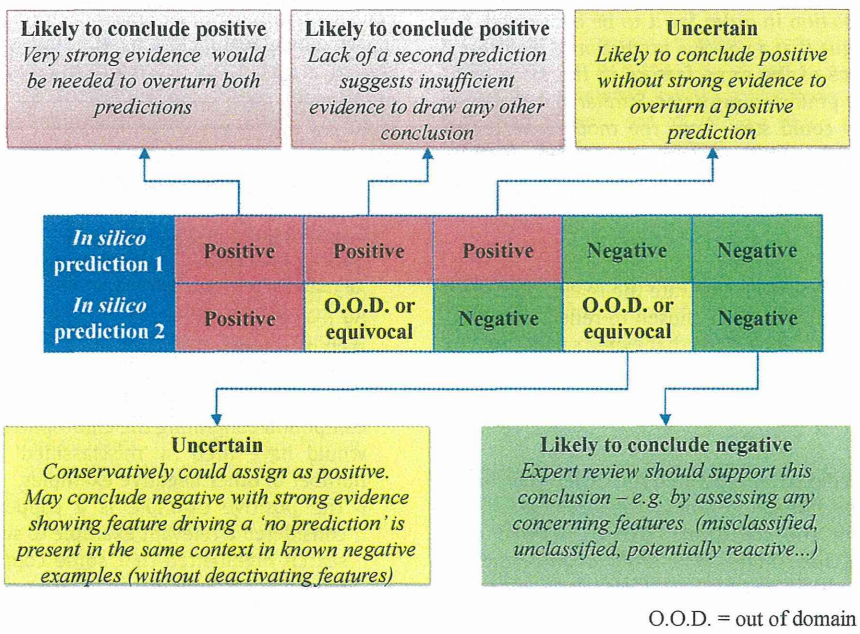

Fig. 3. Decision matrix when evaluating two *in silico* predictions.

reasonably take place for the compound of interest which is supported by relevant (experimental) data. For example, the Benigni-Bossa rule base (Benigni and Bossa, 2008), which has been implemented into some systems, does not have rules that adequately describe the exclusion patterns for tetra-alkyl epoxides which, for steric reasons, do not tend to exhibit DNA reactivity (Wade et al., 1978). There could be a strong argument to challenge a positive prediction in such a case. This illustrates the importance of continuing to develop such systems to capture new knowledge by either modifying or adding alerts to capture this knowledge and explicitly documenting the change within the expert commentary.

A negative prediction could be challenged if there are mis- or unclassified features identified. For example, knowledge of a known positive compound containing a feature not predicted to be active could be sufficient evidence to over-rule a negative prediction in the absence of more relevant data.

7. Challenging a prediction from a statistical system

Since statistical systems can only learn from the data provided, there are a number of reasons that a statistical system could make incorrect predictions. Statistical systems can incorrectly learn as a consequence of the chosen approach, the descriptors used, the methodology applied, the complexity of the endpoint, the presence of activity cliffs, or through limitations of the available training data. For example, a model could incorrectly assign a feature as activating because it is only seen in positive compounds despite it not being the actual cause of activity. Models that apply decision-tree approaches during model learning can identify the most likely cause of activity and not attribute activity to an 'innocent fragment' that is coincident to a mutagenic one. Ultimately however, a statistical model can only learn from the data it is presented with. In such situations, an expert could look at the training examples the model has used and dismiss those having other, more plausible causes of activity.

As with an expert rule-based system, the incorrect prediction of closely related examples could provide grounds to argue that the system has not adequately learnt to predict activity for this class of compound. In order to do this, the model must provide sufficient explanation for the prediction in order for it to be over-ruled. For example a user could argue that a positive prediction is incorrect if it is a direct consequence of the same feature in the same environment in other falsely predicted positives. Similarly, a challenge to a negative prediction could stem from the model's perceived inability to identify the same feature in related positive compounds.

8. Examples

The following worked examples illustrate the reality of making assessments under M7 and have been chosen to reflect the information, thought process and conclusions that experts may come to.

8.1. Methyl sulphate

8.1.1. Model output

Methyl sulphate is predicted as negative by an expert rule-based system but as positive for mutagenicity by a statistical system. The supporting examples for the statistical system consisted of a number of negative long-chain mono-alkyl sulphates and a number of positive polycyclic aromatic benzylic sulphates along with positive di(alkyl)sulphates (Fig. 4).

8.1.2. Expert analysis steps and considerations

- Each of the positive training compounds from the statistical system were subsequently processed through the expert rule-based system.
  - Dialkyl sulphates are predicted to be positive as direct alkylating agents, but the alert specifically excludes mono-alkyl sulphates which are not alkylating agents.
  - The polycyclic aromatic sulphates are all predicted to be positive but for a different reason to that which the statistical system had considered them relevant. The expert commentary and supporting literature references for the polycyclic alert describes a mechanism involving a benzylic carbocation which is common to all the positive compounds (Surh and Miller, 1994) and is absent from all the negative training compounds and from the original query compound.

8.1.3. Summary

- The activity of all the positive training compounds that the statistical system identified can be explained by reasons that do not include the fragment that is common to the query compound, providing the expert with the ability to dismiss them. The supporting negative examples are long-chain mono-alkyl sulphates and considered to be more relevant and result in an expert conclusion to over-rule the positive statistical prediction.

8.1.4. Conclusion

- Negative. This position is supported by knowledge of chemical reactivity and literature comments (Mathison et al., 1995; Wolfenden and Yuan, 2007).

8.2. 4-Chloroisoindoline-1,3-dione

8.2.1. Model output

Both expert rule-based and statistical systems predict the chlorophthalimide to be negative although the former identifies a 'misclassified' feature (Fig. 5). The supporting examples used by the statistical system are derived from two hypotheses and show that (1) similarly substituted chlorobenzenes are Ames negative and (2) there are similar known phthalimides some of which are active.

8.2.2. Expert analysis steps and considerations

- Closer examination of the active phthalimides from the statistical system results in the identification of more plausible causes of activity than the ring system (by visual inspection, confirmed by re-running those examples through the expert rule-based system where alerts were fired for alkyl halide and nitro-aromatic functional groups).
- The expert system had identified a single example of a positive compound containing the chlorophthalimide for which no alert would have fired (a 'misclassified' fragment), along with a number of other negative examples.
  - The positive example is a copper complex and was not considered a relevant example to support activity. A database search highlighted that some copper complexes have been shown to be mutagenic (Feig et al., 1988) although copper itself is not (EMEA, 1998).

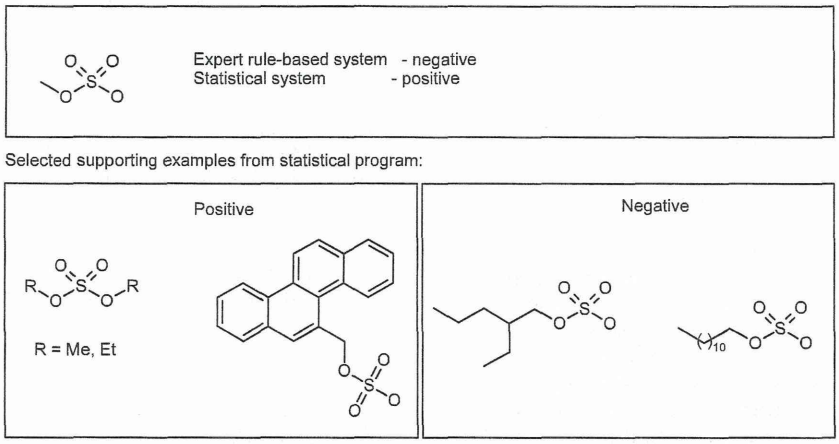

Fig. 4. Supporting information for the assessment of methyl sulphate.

8.2.3. Summary

Both systems make a negative prediction and the compound identified as causing a misclassified feature was dismissed by expert review as not relevant.

8.2.4. Conclusion

Negative. In this case, a known pesticide containing this fragment was subsequently found during a database search. This has been reported as non-mutagenic further supporting this conclusion (University of Hertfordshire, n.d).

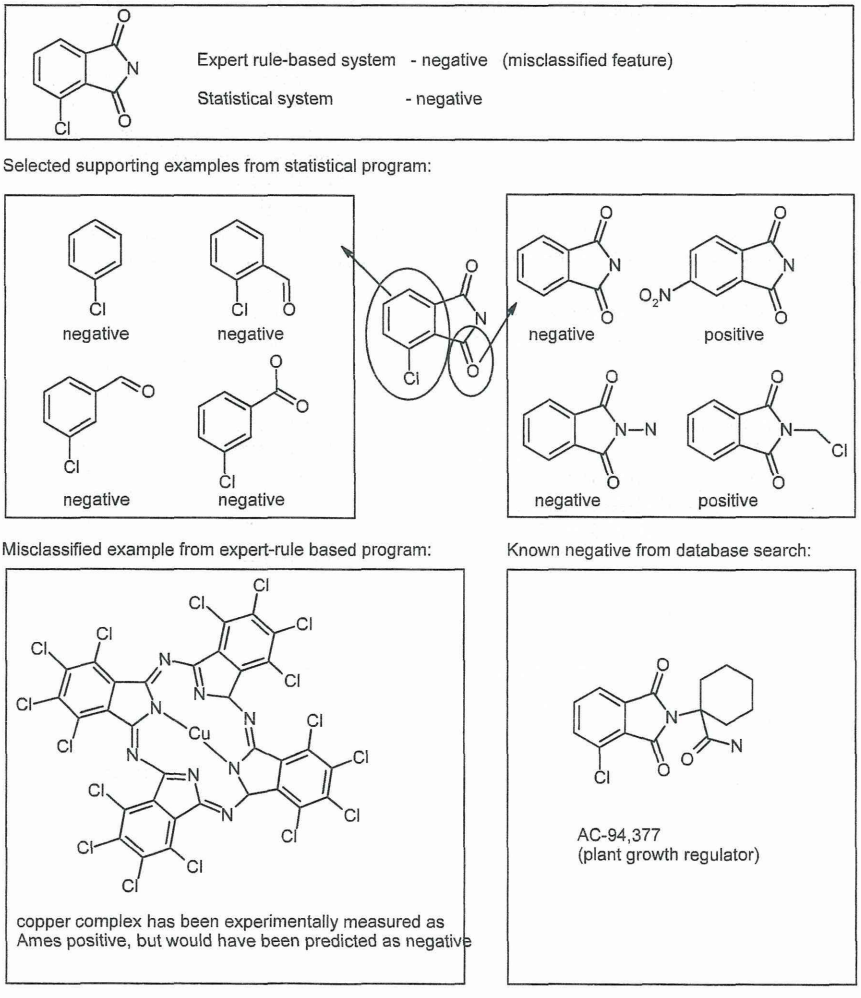

Fig. 5. Supporting information for the assessment of 4-chloroisoindoline-1,3-dione.

8.3. 1-propyl-4,5-dihydrotriazole

8.3.1. Model output

Both expert rule-based and statistical systems predict the reduced triazole to be negative; however, the expert rule-based system identifies a misclassified feature.

8.3.2. Expert analysis steps and considerations

- The statistical system, whilst making a negative prediction, does not show sufficiently relevant training examples to allow an expert to conclude that the negative prediction is reliable.
- The expert rule-based system shows 3 very close analogues that are all active (Fig. 6). Exploring the alerts (in this case by judiciously deleting a bond to give an open-chain system and re-running that compound) resulted in a positive alert showing that acyclic analogues are also known to be active; however, this pattern does not cover cyclic analogues.

8.3.3. Summary

Despite both systems making negative predictions, the presence of misclassified features as well as positive data for several highly similar analogues provides sufficient evidence to over-rule the negative prediction.

8.3.4. Conclusion

Positive. This example has been fed back to the alert developer who has agreed to make modifications to absorb this new knowledge in preparation for the next release showing the importance of reporting mis-predictions back to the software providers.

8.4. 3-Amino-4-chloro-isoxazole

8.4.1. Model output

The expert rule-based system makes a positive prediction. The statistical system provides an equivocal prediction because the compound was considered inside the domain of the model (each part of the molecule had been seen in training compounds) but for which there were insufficient relevant supporting examples from which to draw a conclusion.

8.4.2. Expert analysis steps and considerations

- The positive prediction from the expert rule-based system is well-supported by a detailed evidence-based commentary including a proposed mechanism of action along with relevant references (Fig. 7).
- The equivocal prediction from the statistical-based system does not require detailed analysis since it is superseded by the well-supported positive prediction from the expert rule-based system.

8.4.3. Summary

A positive overall prediction can be easily justified based solely on the expert rule-based prediction.

8.4.4. Conclusion

Positive. A literature search for this compound identified that it has recently been reported as active (Tichenor et al., 2012).

8.5. 6-Methyl adenine

8.5.1. Model output

The statistical system makes a positive prediction based upon some very close analogues, but the expert system does not fire an alert due to an exclusion pattern.

8.5.2. Expert analysis steps and considerations

- The exclusion pattern in the expert rule-based system is driven by the fact that the aniline sits between a fused ring on one side and a ring N on the other, resulting in a combination of stereoelectronic effects that has been shown to reduce the potential for N-hydroxylation (the anticipated first step in the formation of the DNA-reactive species).
  - This effect is well supported by both public and proprietary examples (the latter are not shown within the software).
  - In this case however, the relevance of the specific examples provided by the statistical system were the more persuasive to the reviewer (Fig. 8).
- It should be noted that the alert which is supressed by the exclusion pattern specifically mentions a key paper that shows activity is known for this compound and comments that this alert would not have correctly fired for it (Gorrod et al., 1993).
  - The true cause of activity for this compound is unclear – oxidation of the amino group to the known mutagen 6-N-hydroxylaminopurine has been demonstrated albeit under unusual (forcing) conditions (Clement and Kunze, 1990). Other potential causes of a positive result in an Ames assay have been described including from the potential for 6-methyl adenine to be incorporated in place of adenine (Valinluck et al., 2002) or for bacteria to produce histidine from adenine (Johnston and Roth, 1979) thereby giving a false positive experimental measurement.

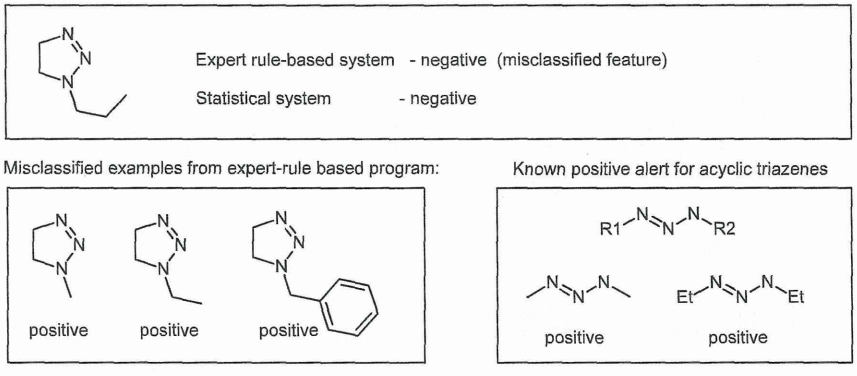

Fig. 6. Supporting information for the assessment of 1-propyl-4,5-dihydrotriazole.

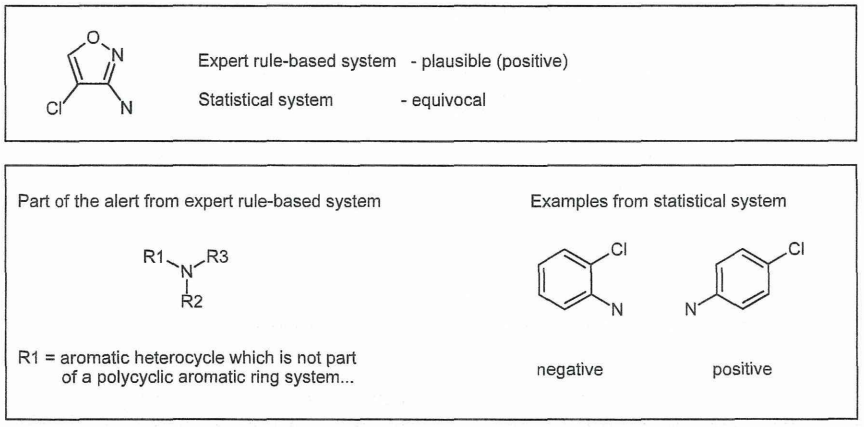

Fig. 7. Supporting information for the assessment of 3-amino-4-chloro-isoxazole.

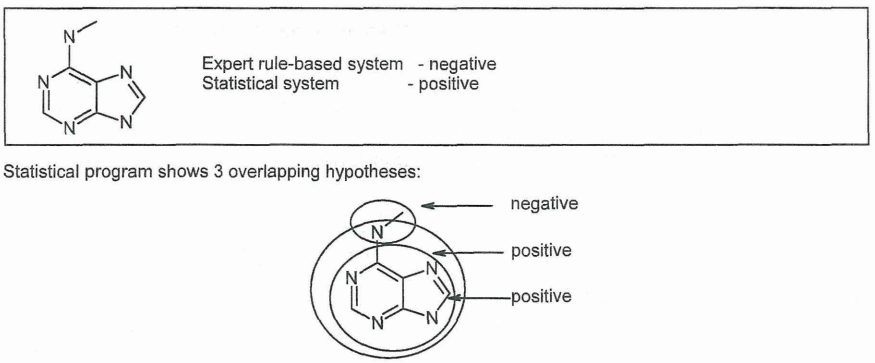

Fig. 8. Supporting information for the assessment of 6-methyl adenine.

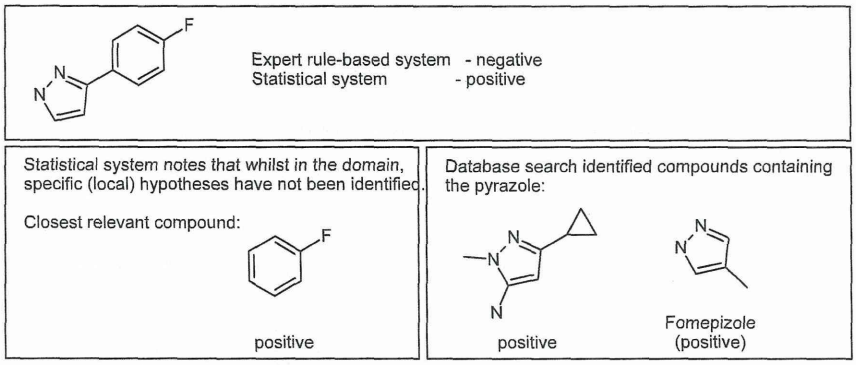

Fig. 9. Supporting information for the assessment of 3-(4-fluorophenyl)-1H-pyrazole.

8.5.3. Summary

The decision to overturn the expert rule-based system was driven by the close analogues identified by the statistical system and was further supported by comments in the expert rule-based system.

8.5.4. Conclusion

Positive.

8.6. 3-(4-fluorophenyl)-1H-pyrazole

8.6.1. Model output

The expert rule-based system identified no alerts and predicted the substituted pyrazole to be negative. The statistical system predicted it to be active, but noted a limited number of relevant training examples (Fig. 9).

8.6.2. Expert analysis steps and considerations

- Inspection of the supporting structures from the statistical system highlighted that insufficient examples were available to support an expert assessment.
  - Supplementing these with a database search of (non-fused) pyrazoles identified two relevant examples, both of which showed activity in Ames assays. Activity of the aminopyrazole could be ascribed to the amino group (running this example through the expert rule-base system added further support to that assessment).
  - The activity of Fomepizole could result following methylene oxidation adjacent to an aromatic ring (not possible with the query compound) suggesting that its activity may not be relevant.

8.6.3. Summary

The lack of a clear prediction and sufficient relevant supporting examples prevented an expert assessment.

8.6.4. Conclusion

Uncertain; should be tested. It was subsequently tested and found to be negative (5-strains).

9. Conclusions

The ability to use *in silico* models to identify potentially DNA reactive compounds is well established and when undertaken with appropriate expert review allows for accurate and scientifically robust predictions to be made with confidence. There are a number of pre-requisites that the expert should demand of a model in order for the predictions to be valuable of which transparency is the most important. The use of two complementary *in silico* systems will at times provide conflicting predictions and this paper aims to describe approaches that may be used to help resolve these. This does not mean that experts will always concur, but it does set out a standard and a framework for such cases to be assessed and presented by an expert.

Transparency document

Transparency document related to this article can be found online at http://dx.doi.org/10.1016/j.yrtph.2015.07.018.

References

Barber, C., Hanser, T., Kruhlak, N.L., Stavitskaya, L., Vessey, J.D., 2014. Establishing best practice for the application of a novel statistical-based model to evaluate potential mutagenic impurities under Ich M7. In: Society of Toxicology, 53rd Annual Meeting. Retrieved from. http://www.lhasalimited.org/Public/Library/Sarah Library/Sarah - Posters/Establishing best practice for use of statistical model for ICH M7.pdf.

Benigni, R., Bossa, C., 2008. Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. Mutat. Res. Rev. Mutat. Res. 659, 248–261. http://dx.doi.org/10.1016/j.mrrev.2008.05.003.

Bentzien, J., Hickey, E.R., Kemper, R. a., Brewer, M.L., Dyekjaer, J.D., East, S.P., Whittaker, M., 2010. An in silico method for predicting Ames activities of primary aromatic amines by calculating the stabilities of nitrenium ions. J. Chem. Inf. Model. 50 (2), 274–297. http://dx.doi.org/10.1021/ci900378x.

Clement, B., Kunze, T., 1990. Hepatic microsomal N-hydroxylation of adenine to 6-N-hydroxylaminopurine. Biochem. Pharmacol. 39 (5), 925–933.

Cocivera, M., Malatesta, V., Woo, K.W., Effio, A., 1978. Mechanism for the reaction involving dimethyl sulfoxide and acetyl chloride studied by nuclear magnetic resonance spectroscopy, 43 (6), 1140–1145. http://dx.doi.org/10.1021/jo00400a025.

Dobo, K.L., Greene, N., Fred, C., Glowienke, S., Harvey, J.S., Hasselgren, C., , et al.Vijayaraj Reddy, M., 2012. In silico methods combined with expert knowledge rule out mutagenic potential of pharmaceutical impurities: an industry survey. Regul. Toxicol. Pharmacol. RTP 62 (3), 449–455. http://dx.doi.org/10.1016/j.yrtph.2012.01.007.

EMEA, 1998. Copper Chloride, Copper Gluconate, Copper Heptanoate, Copper Oxide, Copper Methionate, Copper Sulphate and Dicopper Oxide. Retrieved April 20, 2015, from. http://www.ema.europa.eu/docs/en_GB/document_library/Maximum_Residue_Limits_-_Report/2009/11/WC500013010.pdf.

Feig, A.L., Thederahn, T., Sigman, D.S., 1988. Mutagenicity of the nuclease activity of 1,10-phenanthroline – copper ion. Biochem. Biophys. Res. Commun. 155 (1), 338–343. http://dx.doi.org/10.1016/S0006-291X(88)81090-8.

Gorrod, J.W., Ioannides, C., Lam, S.P., Neville, S., 1993. Mutagenicity testing of 9-N-substituted adenines and their N-oxidation products. Environ. Health Perspect. 101, 21–26. http://dx.doi.org/10.1289/ehp.93101s321.

Greene, N., Dobo, K.L., Kenyon, M.O., Cheung, J., Munzner, J., Sobol, Z., , et al.Wichard, J., 2015. A practical application of two in silico systems for identification of potentially mutagenic impurities. Regul. Toxicol. Pharmacol. 72 (2), 335–349. http://dx.doi.org/10.1016/j.yrtph.2015.05.008.

Hanser, T., Barber, C., Rosser, E., Vessey, J.D., Webb, S.J., Werner, S., 2014. Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge. J. Cheminformatics 6 (1), 21. http://dx.doi.org/10.1186/

1758-2946-6-21.
Hillebrecht, A., Muster, W., Brigo, A., Kansy, M., Weiser, T., Singer, T., 2011. Comparative evaluation of in silico systems for ames test mutagenicity prediction: scope and limitations. Chem. Res. Toxicol. 24 (6), 843–854. http://dx.doi.org/10.1021/tx2000398.
Johnston, H.M., Roth, J.R., 1979. Histidine mutants requiring adenine: selection of mutants with reduced hisG expression in Salmonella Typhimurium. Genetics 92 (May), 1–15.
Judson, P.N., Stalford, S. a., Vessey, J., 2013. Assessing confidence in predictions made by knowledge-based systems. Toxicol. Res. 44 (0), 70–79. http://dx.doi.org/10.1039/c2tx20037f.
Mathison, B.H., Taylor, M.L., Bogdanffy, M.S., 1995. Dimethyl sulfate uptake and methylation of DNA in rat respiratory tissues following acute inhalation. Toxicol. Sci. 28, 255–263. http://dx.doi.org/10.1093/toxsci/28.2.255.
Müller, L., Mauthe, R.J., Riley, C.M., Andino, M.M., Antonis, D. De, Beels, C., , et al.Yotti, L., 2006. A rationale for determining, testing, and controlling specific impurities in pharmaceuticals that possess potential for genotoxicity. Regul. Toxicol. Pharmacol. 44 (3), 198–211. http://dx.doi.org/10.1016/j.yrtph.2005.12.001.
Naven, R.T., Greene, N., Williams, R.V., 2012. Latest advances in computational genotoxicity prediction. Expert Opin. Drug Metab. Toxicol. 8 (12), 1579–1587. http://dx.doi.org/10.1517/17425255.2012.724059.
Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., , et al.Yang, C., 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. ATLA Altern. Lab. Anim. 33, 155–173. Retrieved from. http://www.atla.org.uk/current-status-of-methods-for-defining-the-applicability-domain-of-quantitative-structure-activity-relationships/.
OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-activity Relationships Models, 2004. Retrieved April 20, 2015, from. http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm.
Powley, M.W., 2015. (Q) SAR assessments of potentially mutagenic impurities : a regulatory perspective on the utility of expert knowledge and data submission. Regul. Toxicol. Pharmacol. 71 (2), 295–300. http://dx.doi.org/10.1016/j.yrtph.2014.12.012.
Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R., 2012. Comparison of different approaches to define the applicability domain of QSAR models. Mol. Basel, Switz. 17 (5), 4791–4810. http://dx.doi.org/10.3390/molecules17054791.
Surh, Y.J., Miller, J. a, 1994. Roles of electrophilic sulfuric acid ester metabolites in mutagenesis and carcinogenesis by some polynuclear aromatic hydrocarbons. Chem. Biol. Interact. 92, 351–362. http://dx.doi.org/10.1016/0009-2797(94)90076-0.
Sutter, A., Amberg, A., Boyer, S., Brigo, A., Contrera, J.F., Custer, L.L., , et al.van Gompel, J., 2013. Use of in silico systems and expert knowledge for structure-based assessment of potentially mutagenic impurities. Regul. Toxicol. Pharmacol. RTP 67 (1), 39–52. http://dx.doi.org/10.1016/j.yrtph.2013.05.001.
Tichenor, M.S., Keith, J.M., Jones, W.M., Pierce, J.M., Merit, J., Hawryluk, N., , et al.Breitenbucher, J.G., 2012. Heteroaryl urea inhibitors of fatty acid amide hydrolase: structure-mutagenicity relationships for arylamine metabolites. Bioorg. Med. Chem. Lett. 22 (24), 7357–7362. http://dx.doi.org/10.1016/j.bmcl.2012.10.076.
University of Hertfordshire. (n.d.). Pesticides Properties DataBase. Retrieved April 22, 2015, from http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/2635.htm.
Valinluck, V., Liu, P., Burdzy, A., Ryu, J., Sowers, L.C., 2002. Influence of local duplex stability and N6-methyladenine on uracil recognition by mismatch-specific uracil-DNA glycosylase (Mug). Chem. Res. Toxicol. 15 (12), 1595–1601. http://dx.doi.org/10.1021/tx020062y.
Wade, D.R.R., Airy, S.C.C., Sinsheimer, J.E.E., 1978. Mutagenicity of aliphatic epoxides. Mutat. Res. 58, 217–223. http://dx.doi.org/10.1016/0165-1218(78)90012-5.
Williams, R., Stalford, S., 2014. Making negative predictions for mutagenicity. In: Genetic Toxicology Association Meeting. Newark, Delaware. Retrieved from. http://www.gta-us.org/scimtgs/2014Meeting/posters2014.html.
Wolfenden, R., & Yuan, Y. (2007). Monoalkyl Sulfates as Alkylating Agents in Water, Alkylsulfatase Rate Enhancements, and the "Energy-rich" Nature of Sulfate Half-esters, 104(1), 83–86.

Regulatory Toxicology and Pharmacology 72 (2015) 17–25

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

A feasibility study: Can information collected to classify for mutagenicity be informative in predicting carcinogenicity?

Petko I. Petkov [a], Grace Patlewicz [b,*], Terry W. Schultz [c], Masamitsu Honma [d], Milen Todorov [a], Stefan Kotov [a], Sabcho D. Dimitrov [a], E. Maria Donner [b], Ovanes G. Mekenyan [a]

[a] Laboratory of Mathematical Chemistry (LMC), As. Zlatarov University, Bourgas, Bulgaria
[b] DuPont Haskell Global Centers for Health and Environmental Sciences, Newark, DE 19711, USA
[c] College of Veterinary Medicine, The University of Tennessee, Knoxville, TN 37996-4500, USA
[d] Division of Genetics and Mutagenesis, National Institute of Health Sciences, Tokyo, Japan

ARTICLE INFO

Article history:
Received 5 September 2014
Available online 16 March 2015

Keywords:
Carcinogenicity classification
Integrated Approaches to Testing and Assessment (IATA)
(Q)SAR
Adverse Outcome Pathway (AOP)

ABSTRACT

Carcinogenicity is a complex endpoint of high concern yet the rodent bioassay still used is costly to run in terms of time, money and animals. Therefore carcinogenicity has been the subject of many different efforts to both develop short-term tests and non-testing approaches capable of predicting genotoxic carcinogenic potential. In our previous publication (Mekenyan et al., 2012) we presented an *in vitro–in vivo* extrapolation workflow to help investigate the differences between *in vitro* and *in vivo* genotoxicity tests. The outcomes facilitated the development of new (Q)SAR models and for directing testing. Here we have refined this workflow by grouping specific tests together on the basis of their ability to detect DNA and/or protein damage at different levels of biological organization. This revised workflow, akin to an Integrated Approach to Testing and Assessment (IATA) informed by mechanistic understanding was helpful in rationalizing inconsistent study outcomes and categorizing a test set of carcinogens with mutagenicity data on the basis of regulatory mutagenicity classifications. Rodent genotoxic carcinogens were found to be correctly predicted with a high sensitivity (90–100%) and a low rate of false positives (3–10%). The insights derived are useful to consider when developing future (non-)testing approaches to address regulatory purposes.

1. Introduction

Carcinogenicity is a complex toxicological endpoint of high concern. At the same time the rodent bioassay currently employed to assess carcinogenic potential is costly to run in terms of time, money and number of animals. Therefore carcinogenicity has been the subject of many efforts to develop *in vitro* and *in vivo* short-term tests, specifically capable of predicting genotoxic carcinogenic potential. The available genotoxicity tests assess the potential of substances to cause cancer or heritable diseases in humans. The data generated is used in both the hazard identification and risk characterization of substances for regulatory and product stewardship purposes.

Hazard identification for genotoxicity mainly relies on *in vitro* studies determining mutagenicity of substances in bacteria and in mammalian cells following an initial review of existing literature and Structure Activity Relationship/Quantitative Structure Activity Relationship (SAR/QSAR) pre-screening. Effects such as DNA damage, formation of strand breaks or adducts are other helpful indicators for genotoxicity. *In vivo* studies are also used to evaluate genotoxic potential further and are typically conducted to put *in vitro* observations into perspective.

Given the many different modes of action for mutagenesis, a number of tests are needed to assess whether a chemical is genotoxic or not with any degree of confidence. When combined appropriately, positive results from mutagenicity tests can be used to predict carcinogenicity. Some modes of actions involved in the cancer initiation step (e.g., epigenetic DNA methylation) remain without experimental data support because no appropriate test systems for their identification have yet been developed. This can potentially bring some limitations to the currently employed strategies for predicting carcinogenesis. There have been a number of efforts to investigate strategies for evaluating mutagenicity both from the perspective of classifying a chemical as a mutagen or in directing further work in the assessment of carcinogenic potential (Zeiger, 1998; Kirkland et al., 2005, 2014; Cimino, 2006; Matthews

* Corresponding author at: EPA/ORD/NCCT, 109 T.W. Alexander Dr. B205-01, RTP, NC 27711, USA.
E-mail address: patlewig@hotmail.com (G. Patlewicz).

http://dx.doi.org/10.1016/j.yrtph.2015.03.003
0273-2300/