Cronin, MTD, Bauer, FJ, Bonnell, M, Campos, B, Ebbrell, DJ, Firman, JW, Gutsell, S, Hodges, G, Patlewicz, G, Sapounidou, M, Spinu, N, Thomas, PC and Worth, AP

A scheme to evaluate structural alerts to predict toxicity – Assessing confidence by characterising uncertainties

A scheme to evaluate structural alerts to predict toxicity – Assessing confidence by characterising uncertainties

Mark T.D. Cronin, Franklin J. Bauer, Mark Bonnell, Bruno Campos, David J. Ebbrell, James W. Firman, Steve Gutsell, Geoff Hodges, Grace Patlewicz, Maria Sapounidou, Nicoleta Spînu, Paul C. Thomas, Andrew P. Worth

PII: S0273-2300(22)00136-2

DOI: https://doi.org/10.1016/j.yrtph.2022.105249

Reference: YRTPH 105249

To appear in: *Regulatory Toxicology and Pharmacology*

Received Date: 18 March 2022

Revised Date: 12 July 2022

Accepted Date: 17 August 2022

Please cite this article as: Cronin, M.T.D., Bauer, F.J., Bonnell, M., Campos, B., Ebbrell, D.J., Firman, J.W., Gutsell, S., Hodges, G., Patlewicz, G., Sapounidou, M., Spînu, N., Thomas, P.C., Worth, A.P., A scheme to evaluate structural alerts to predict toxicity – Assessing confidence by characterising uncertainties, *Regulatory Toxicology and Pharmacology* (2022), doi: https://doi.org/10.1016/j.yrtph.2022.105249.

Mark T.D. Cronin: Conceptualization, Investigation, Methodology, Writing - Original Draft, Writing - Review & Editing, Visualization

Franklin J. Bauer: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing

Mark Bonnell: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing

Bruno Campos: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing

David J. Ebbrell: Writing - Review & Editing

James W. Firman: Writing - Review & Editing

Steve Gutsell: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing

Geoff Hodges: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing

Grace Patlewicz: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing

Maria Sapounidou: Writing - Review & Editing

Nicoleta Spînu: Writing - Review & Editing

Paul C. Thomas: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing

Andrew P Worth: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing

A Scheme to Evaluate Structural Alerts to Predict Toxicity – Assessing Confidence By Characterising Uncertainties

Mark T.D. Cronin,[1] Franklin J. Bauer,[2] Mark Bonnell,[3] Bruno Campos,[4] David J. Ebbrell,[1] James W. Firman,[1] Steve Gutsell,[4] Geoff Hodges,[4] Grace Patlewicz,[5] Maria Sapounidou,[1] Nicoleta Spînu,[1] Paul C. Thomas,[2] Andrew P Worth[6]*

[1]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

[2]KREATiS SAS, 23 rue du Creuzat, ZAC de St-Hubert, 38080 L'Isle d'Abeau, France

[3]Science and Risk Assessment Directorate, Environment & Climate Change Canada, 351 St. Joseph Blvd, Gatineau, Quebec K1A 0H3, Canada

[4]Safety and Environmental Assurance Centre, Unilever, Colworth Science Park, Bedfordshire MK44 1LQ, UK

[5]Center for Computational Toxicology and Exposure (CCTE), US Environmental Protection Agency, 109 TW Alexander Dr, RTP, NC 27709

[6]European Commission, Joint Research Centre (JRC), Ispra, Italy

*Author for correspondence: Dr Andrew Worth

Email: Andrew.Worth@ec.europa.eu

ORCID Identifiers:

Mark T.D. Cronin: 0000-0002-6207-4158

Mark Bonnell: 0000-0003-3300-4588

Bruno Campos: 0000-0002-3701-4858

James W. Firman: 0000-0003-0319-1407

Grace Patlewicz: 0000-0003-3863-9689

Nicoleta Spînu: 0000-0002-9465-3090

Paul C. Thomas: 0000-0003-1671-9274

Andrew P Worth: 0000-0002-5303-0452

Abstract

Structure-activity relationships (SARs) in toxicology have enabled the formation of structural rules which, when coded as structural alerts, are an essential tool in *in silico* toxicology. Whilst other *in silico* methods have approaches for their evaluation, there is no formal process to assess the confidence that may be associated with a structural alert. This investigation proposes twelve criteria to assess the uncertainty associated with structural alerts, allowing for an assessment of confidence. The criteria are based around the stated purpose, description of the chemistry, toxicology and mechanism, performance and coverage, as well as corroborating and supporting evidence of the alert. Alerts can be given a confidence assessment and score, enabling the identification of areas where more information may be beneficial. The scheme to evaluate structural alerts was placed in the context of various use cases for industrial and regulatory applications. The analysis of alerts, and consideration of the evaluation scheme, identifies the different characteristics an alert may have, such as being highly specific or generic. These characteristics may determine when an alert can be used for specific uses such as identification of analogues for read-across or hazard identification.

Keywords

Structural alert; structure-activity relationship; toxicity prediction; confidence; uncertainty; evaluation scheme; use case; computational toxicology

Highlights

- Structural alerts are useful tools for predictive toxicology
- 12 criteria to evaluate structural alerts have been identified
- A strategy to determine confidence of structural alerts is presented
- Different use cases require different characteristics of structural alerts

61 **Abbreviations**

62 AChE, acetylcholinesterase; AOP, Adverse Outcome Pathway; EFSA, European Food Safety Authority;
63 HPV, High Production Volume; KE, Key Event; MIE, Molecular Initiating Events; OECD, Organisation
64 for Economic Cooperation and Development; QMRF, QSAR Model Reporting Format; QPRF, QSAR
65 Prediction Report Format; QSAR, quantitative structure-activity relationship; RAAF, Read-Across
66 Assessment Framework; SAR, structure-activity relationship

67

1. Introduction

The concept of the structure-activity relationship (SAR) is fundamental to predictive toxicology (Cronin and Yoon, 2019). As such, SARs have found widespread use in toxicology, risk assessment and other regulatory applications with a particular resurgence of interest with the increasing desire to consider safety without the use of animals (Worth, 2020). Key to enabling SARs as usable *in silico* tools for these applications is the development of structural rules which can then be coded computationally so that they may be applied to identify potential hazard in new molecules (Madden et al., 2021; Cronin et al., 2022). The term "structural alert" is assumed in this paper to represent a fragment or substructure within a molecule that is hypothesised to be responsible for a biological activity from a structural rule. Such a fragment is derived from SAR-based structural rules and may be associated with other structural information such as that relating to substitution patterns or parent structures.

Structural alerts can represent the chemistry which is associated with, for instance, an interaction such as a molecular initiating event (MIE) or key event (KE) in an Adverse Outcome Pathway (AOP) (Allen et al., 2018), an adverse effect (i.e., toxicity that can be observed at an organism or population level) (Siramshetty et al., 2018) or related to a regulatory endpoint (Valsecchi et al., 2019) or indicator of significant toxicity (e.g., as part of the Cramer et al (1978) Decision Tree). The understanding that chemical properties were responsible for toxicological events was well established at the turn of the twentieth century (e.g. Meyer, 1901; Overton, 1901) along with the concept that specific chemical structures could be associated with toxicity (Landsteiner and Jacobs, 1935). The first use of the term "structural alert" is accredited to Ashby (1985) with regard to defining the structural basis of carcinogenicity, a concept that went on to define a series of alerts for genotoxic carcinogenicity (Ashby and Tennant, 1988). Since that time, alerts have been developed in many areas of toxicology for human health and environmental endpoints. The history and use of alerts in toxicology has been well reviewed recently (Cronin and Yoon, 2019; Yang et al., 2020) and a large compilation of alerts is freely available through the OCHEM website (https://ochem.eu; Sushko et al., 2011; 2012).

There are a number of ways of developing the SAR which forms the basis of structural alerts and these are summarised in Table 1, along with their characteristics and strengths and weaknesses. No method is exclusive and, in terms of understanding their use better, no analysis has been performed to determine if or when a particular method may be appropriate. From the outset, it is acknowledged that “expert knowledge” is a subjective term with no clear criteria to define it. In terms of the use of the term “expert knowledge” in this study, it is assumed that the expert would have some training or appreciation of toxicology in the context of hazard identification and be familiar with relevant data for the chemical(s) and endpoint in question.

101 Table 1. Summary of approaches to derive structural-activity relationships, and ultimately structural alerts, for predictive toxicology

Method to derive the structural alert	Description	Characteristics in terms of data for the SAR, methodology and mechanistic understanding	Strengths	Weaknesses	Illustrative example
Expert Knowledge Based on Toxicological Data	Derived from the knowledge of toxicologists who have experience in assessing the data associated with toxicological properties of a series of chemicals	Data: small number of toxicological data on which to base a hypothesis Methodology: Expert judgement and opinion Mechanistic: Presumed high, through precise	Derived from a knowledge based on experimental data, supported by mechanistic information	Slow to develop, no performance statistics; may be a misinterpretation from flawed data or a subjective interpretation of data	Ashby and Tennant (1988) who compiled knowledge on genotoxic carcinogens

		mechanistic definition may not be possible			
Expert Knowledge Based on Mechanistic Understanding	Derived from expert knowledge following (non-statistical) analysis of a data set of chemicals using a mechanistic hypothesis	Data: large number of mechanistic data Methodology: Expert judgement and opinion Mechanistic: Clear mechanistic hypothesis	Based on expert knowledge (preferably from multiple sources) and potentially creating a broad set of alerts, supported by data or mechanistic understanding. Can be extended broadly without extensive toxicological data.	Labour intensive to develop and requires expert knowledge across a complete mechanism of action or dataset	Enoch and Cronin (2010) and Enoch et al. (2011) who derived alerts for DNA and protein binding respectively on the basis of electrophilic chemistry; Bauer et al (2018) who derived a decision tree on six classes of mechanisms of action, termed MechoA
Data-Driven Approaches	Use of statistical analyses to determine fragments	Data: Large data sets required for analysis	A rapid method, with readily available performance statistics.	Requirement for large data sets to achieve significant results.	Wedlake et al. (2020) used a Bayesian approach to develop alerts for *in vitro*

	associated with a particular toxicity	Methodology: data mining and machine learning of toxicological data Mechanistic: Not possible unless assigned after alert development	The data on which the alerts are derived from are available	Prone to limited validation (usually restricted to curation). Difficult to assign mechanistic knowledge or validity to the alerts derived as they may be in an uninterpretable "black box" form. Often the fragments are overlapping and require rationalisation	data related to MIEs; Claesson and Minidis (2018) to develop alerts for reactive metabolite formation; Cui et al. (2019) alerts from fingerprints for drug-induced rhabdomyolysis
Chemotype Enrichment	Use of statistical analysis to determine which structural fragments may	Data: Large data sets	Rapid to apply. Provides a statistical outcome to	Currently limited by the need for relatively large data sets and the	Wang et al. (2019; 2021) investigated ToxCast

	be significantly associated with a toxicity or effect	Methodology: Data mining of high throughput data Mechanistic: Driven by the mechanistic hypothesis of the data	demonstrate the strength of relationship between the activity and structure. Use of readily available alerts.	fragments already available	endpoints using ToxPrint Chemotypes
Hybrid Approaches Combining Statistical Analysis and Expert Analysis	For purpose here is to use statistical analysis (such as clustering approaches) to find groups within data to be used as leads for expert analysis. This will not produce a comprehensive set of alerts but may find SARs	Data: Many toxicological data Methodology: Clustering of data following by expert judgement and opinion Mechanistic: No mechanistic understanding unless	A rapid approach to derive knowledge / hypotheses. Supported by data and mechanistic understanding	Evaluating the hypotheses from data mining can be slow and requires expert knowledge.	Hewitt et al. (2013) who applied expert knowledge to the results of cluster analyses on a database of hepatotoxicity data to derive usable alerts for liver toxicity. Wang et al. (2019) used a ToxPrint chemotype enrichment analysis to identify >20 distinct chemical

	(which can be optimised) that would not be obtained by expert knowledge alone.	applied after alert development			substructural features as significantly enriched for the sodium-iodide symporter inhibition.

102

As well as the description of methods to develop structural alerts in Table 1, other characteristics of alerts could be considered to improve their use including their definition, underlying data source(s), potential domain, mechanistic relevance, coverage and performance. Whilst these are likely to be crucial for the successful use of structural alerts, they are seldom defined, although several recent studies have demonstrated that careful development of alerts can improve performance and relevance (Amberg et al., 2019; Benigni, 2021; Kalgutkar, 2020; Kalgutkar and Driscoll, 2020). In addition, the different uses of structural alerts e.g., for hazard assessment, grouping and read-across, screening etc. have not been fully described. As such, a better understanding of the properties, specifically the strengths and weaknesses, of alerts should increase confidence in their application and hence improve opportunities for acceptance, especially for regulatory purposes.

Despite the extensive development and use of structural alerts, their importance, and reliance on them in many use cases, no standardised agreed means of describing them and assessing their utility in terms of their reliability and robustness has been developed. This is in contrast with related approaches where assessment formats have been put in place, such as read-across (e.g., the Read-Across Assessment Framework (RAAF) (ECHA, 2017)) and quantitative structure-activity relationships (QSARs) (e.g. the Organisation for Economic Cooperation and Development (OECD) Principles for the Validation of QSARs (OECD, 2007), QSAR Model Reporting Format (QMRF), QSAR Prediction Report Format (QPRF) (Worth, 2010)). The lack of an agreed approach has potentially reduced confidence in the application of SARs. As such a means of evaluating structural alerts would enable confidence to be assigned to them, ensure their optimal usage and enhance their acceptability.

One means of understanding confidence in computational toxicology tools has been through the characterisation and definition of uncertainty. For example, Schultz et al. (2019) have defined the uncertainties associated with read-across and Cronin et al. (2019) have detailed areas of uncertainty, variability and bias of QSARs for toxicity prediction. The purpose of these analyses was not to conclude that a particular approach should, or should not, be used, but to assist in the validation process, identify aspects of a model that may be associated with significant levels of uncertainty and determine

the overall confidence that may be assigned to a model. This approach to understanding uncertainty provides the opportunity to determine the type and level of confidence required for a predictive toxicology approach to be "fit-for-purpose" (Belfield et al. 2021).

One of the most recognised set of criteria in health sciences and toxicology to define confidence that may be associated with evidence to support a conclusion, i.e., causation, are the Bradford Hill criteria (Hill, 1965). These were adapted by Meek et al. (2014), amongst others, to assist in a weight of evidence framework for mode of toxicological action which are closely aligned to the issue of evaluating structural alerts. The revised criteria included assessment of biological concordance, essentiality of KEs, concordance of empirical observations among KEs, consistency and analogy. Whilst these adapted Bradford Hill criteria cannot be mapped directly for the assessment of structural alerts, they provide a starting point e.g., assessment of mechanisms, underlying evidence and definition. Likewise, there is as yet no agreement of the level of quantification of uncertainty that can, or should, be applied. Schultz et al. (2019) reviewed this topic as regards to read-across and concluded at the current time a simple "high, moderate, low" scheme was the most practical. It is also noted that, with regard to AOPs, more quantitative schemes have been proposed with six (Collier et al. 2016) and seven levels of "evidence" respectively (Patlewicz et al. 2013; Becker et al. 2017). Indeed, a "scientific confidence framework" has been developed by Patlewicz et al. (2015) to support the use of AOPs for regulatory purposes. This formalises a number of criteria (seven in total) that were developed by Patlewicz et al. (2013) based on analogous assessment schemes for biomarkers and QSAR. These, and other, studies demonstrate that confidence in the use of strategies for using non-animal data can be assessed in a meaningful manner to support their use. The acceptable level of uncertainty for a particular purpose, e.g. a regulatory decision, remains difficult to ascertain and is likely to be context dependent.

Given the lack of a defined set of criteria to assess structural alerts for toxicity, the aim of this investigation was to develop a scheme for their critical evaluation. Specifically, we aimed to determine how criteria for describing the confidence in structural alerts for the prediction of toxicity could be

developed based on the assessment of the uncertainties of the alerts. Reference was made to adapted Bradford Hill criteria (i.e., to assess the likelihood of causation) and other schemes for computational toxicology, with the objective of assessing and numerically scoring the overall confidence that may be placed in an alert. Further, use cases for structural alerts were reviewed with the objective of determining the characteristics of alerts that may be required for certain applications in predictive toxicology.

2. **Methods**

2.1 Development of Criteria to Define the Uncertainty Associated with Structural Alerts for Toxicity Prediction

A set of criteria was created to define the properties of, and uncertainty associated with, structural alerts for toxicity prediction. This task was performed by the authors using expert analysis to address particular aspects of structural alerts, in part with reference to the adapted Bradford Hill criteria, which can be summarised as follows:

- Description and definition of the domain of the structural alert
- Evidence of causality e.g., mechanisms of action
- Concordance and consistency of biology e.g., supporting data
- Performance of the structural alert

In order to make the criteria usable for the evaluation of structural alerts, the broad themes stated above were defined by a larger number of definable criteria deemed practical for the description of the uncertainties of a structural alert.

Provisional Scheme for Assigning a Confidence Score to a Structural Alert

Following definition of the criteria for the uncertainty associated with a structural alert, each was categorised with definitions for low, moderate and high uncertainty to make it into a practical and workable scheme. Should any particular criterion be irrelevant to the alert, then this would be defined "not applicable".

In order to provide the possibility of creating an overall score, individual criteria were ranked according to their potential importance when using a structural alert. The ranking was performed semi-quantitatively and undertaken using expert opinion and interpretation.

2.2 Assessment of Use Cases for Structural Alerts

The use cases for structural alerts to predict toxicity were scoped, representing in particular both regulatory use and application within industry. Specifically, use cases were sought for different applications of structural alerts with the overall aim of predicting toxicity. For each use case the desirable characteristics of an alert were defined. The desirable characteristics were based around the criteria for definition of uncertainties and were defined as low, moderate or high. The aim of this exercise was to define and identify the types of structural alerts that are most suited for a particular use case, such that these properties could be defined by the developer / user as a means to demonstrate the applicability of an alert, group of alerts or *in silico* profiler.

3. Results and Discussion

This study aimed to develop a scheme to evaluate the uncertainty associated with structural alerts for the prediction of toxicity such that confidence in their use could be assigned. In order to develop such a scheme, cognisance was taken of a number of approaches starting with the definition of uncertainty as provided by European Food Safety Authority (EFSA) which defined uncertainty with regard to toxicological assessment as "*all types of limitations in available knowledge that affect the range and*

202 *probability of possible answers to an assessment question*" (EFSA, 2018). The EFSA Guidance is based
203 around identifying, assessing, describing and, in some cases, quantifying uncertainty and it is this
204 definition that was applied by Cronin et al. (2019) to defining the uncertainty and other properties of
205 QSAR models.

206

207 *3.1 Uncertainty Assessment Criteria for Structural Alerts*

208 The assessment of criteria relating to uncertainty was performed with the intention of providing a
209 scheme that would assist in the evaluation of structural alerts and to determine the types of
210 uncertainty that may be acceptable for defined scenarios. The development of criteria focused on the
211 definition and domain(s) (in terms of the biology/toxicology predicted, chemical structure and
212 properties, requirement for metabolic activation etc,) of an alert, its mechanistic relevance,
213 performance and the level of evidence supporting the alert. In total, twelve assessable criteria were
214 identified that covered the main aspects of uncertainty of a structural alert, these are described in
215 detail and with their relevance to uncertainty in Table 2.

216 The first criterion (as stated in Table 2) for the assessment of structural alerts relates to its "Purpose"
217 which will ensure that a proper use case scenario has been assigned. The following five criteria
218 (Structural Description, Property Domain, Toxicity or Relationship to Adversity, Species Specificity,
219 Metabolic Domain) attempt to define uncertainty associated with the definition of the alert and its
220 applicability. It is essential that a structural alert must be adequately defined in terms of chemical
221 structure or toxicophore, otherwise it will be difficult or impossible to use. Its description should be
222 explicit and ideally comprise any confounding or influencing factors e.g., that may promote a change,
223 increase or decrease in activity. It is important to note that slight differences in structure may be
224 associated with large changes in activity and toxic effects, this is often termed an "activity cliff"
225 (Maggiora, 2006). Such minor differences in structure may affect reactivity, and hence endpoints such
226 as skin sensitisation (Pestana et al., 2022) or receptor binding, notable for reproductive effects (Mori

et al., 2018). To be accurate, structural alerts must encode this information, to avoid over-prediction. The definition of the domain of alerts is assisted by consideration of all data, for instance *in chemico* data have been utilised to define the domains of a number of reactive mechanisms associated with skin sensitisation (Richarz et al., 2014; Rodriguez-Sanchez et al.,(2013); Nelms et al., (2013)).

The definition of domain associated with physico-chemical properties will allow for cut-offs e.g., for solubility or volatility to be incorporated which will account, in part at least, for elements of toxicokinetics. At the current time, this aspect of the domain is seldom characterised. However, a broad (or no) physico-chemical property domain will extend the coverage of an alert, and strict cut-offs will restrict coverage, i.e., general or highly specific respectively. The definition of domain in terms of physico-chemical properties must implicitly be derived from training set data and hence is likely to forge a link with species specificity. In most cases, physico-chemical properties are likely to be related to the toxicokinetics of a compound, i.e. an alert may indicate the toxicodynamic possibility of initiating toxicity, but this may be tempered by adverse toxicokinetic properties. The incorporation of a physico-chemical property and / or descriptor domain may ultimately allow for some form of quantification, as demonstrated recently with regard to determining groupings of potency for repeated dose toxicity (Yang et al., 2021), increasing reactivity or bioavailability that may be associated with skin sensitisation (Natsch et al., 2015) or the Cramer Classes for systemic toxicity (Cramer et al., 1978).). Alternatively, structural alerts without physico-chemical properties can be used in combination with QSAR models, where the structural alert guides the user to select the appropriate QSAR model that relates to the mechanism predicted by the structural alert (e.g. a QSAR model for non-polar narcosis), while the QSAR itself incorporates the physico-chemical properties allowing for quantitative prediction. The toxicity, endpoint or adverse effect predicted should be defined, along with the species to which it is relevant. With regard to definition of species, this will be dependent on the training set and endpoint. There are also examples, e.g., alerts for the inhibition of acetylcholinesterase where the alert will be very broadly applicable, and even a statement such as “any species with acetylcholinesterase” may be seen as appropriate. The species applicability of some

alerts may also be defined by extrapolation, using for instance protein orthology databases (LaLone et al., 2016), which is of course leading to higher uncertainty, but at the same time greatly enhancing the species applicability domain of the alert. In terms of metabolism, it is acknowledged that some alerts implicitly imply metabolism and this is captured in the description of the alert. An example being for the DNA reactivity of an aromatic amine, which implicitly includes a metabolic step to the nitrenium ion or nitroso derivative (Bauer et al., 2018; Enoch and Cronin, 2010). However, not all metabolic transformations are captured implicitly in alerts, with some requiring knowledge of metabolism or use of a metabolic simulator and the alert is only found in the metabolite e.g. some phenols can be oxidised to the corresponding quinone which may be a skin sensitiser, whilst the alert is often associated with the quinone alone (Bajot et al., 2011). There are also many direct acting, non-metabolically activated, alerts for toxicity. The purpose of this criterion is that the requirement (or not) for metabolism should be stated, or if this knowledge is not known it should be acknowledged as an uncertainty. The uncertainty is not in the requirement for metabolism, but whether it is known and stated unambiguously.

The evidence of causality of an alert i.e., that it is plausible, is captured partially by mechanistic relevance with two criteria (Mechanistic Interpretation and Mechanistic Causality) and related to the criteria describing the availability of corroborating or supporting evidence. Mechanistic relevance is important to provide evidence of causality, i.e., that it is toxicologically meaningful, and hence transparency of an alert. In this case Mechanistic Interpretation ascertains the confidence in there being a recognisable mechanism of action that can be associated to the SAR and, ultimately, structural alert. Mechanistic Causality is whether the description of the structural alert, in terms of chemistry or properties, is related to the mechanism of action. Reference to AOPs is highly useful in this context (OECD, 2017), particularly with regard to MIEs which may drive structural alerts (Cronin and Richarz, 2017). The two criteria are not independent and assessing Mechanistic Causality is not possible without knowledge of Mechanistic Interpretation, or at very least knowledge of a potential mechanism and / or MIE. This is important to demonstrate the veracity of an alert, although it is

acknowledged that full mechanistic interpretation may not be possible for all alerts i.e., when the mechanisms are unknown or debated. Supporting evidence is addressed with two criteria. Corroborating Evidence relates to relevant biological data, e.g., *in vivo* assays, or *in vitro* data relating and confirming a mechanism or adverse outcome directly, that support the structural alert. Corroborating Evidence can also relate to high-throughput or high content data, for instance to explore alerts associated with MIEs and KEs (Wang et al. 2019; 2021). Supporting Evidence is other evidence or data streams, which may have lower levels of biological complexity, e.g., other *in vitro* data, high content screening, omics outputs etc., that support weight of evidence to provide the mechanistic relevance of the structural alert. Supporting Evidence can, however, also include other information such as data from related endpoints, non-standard data etc., for instance the use of mutagenicity data to support the assessment of skin sensitisation (Mekenyan et al., 2010).

The final two criteria to consider (Coverage, Performance) are objective and will assist in understanding how an alert can be used. Coverage can be defined as the number of hits the alert has in a chemically diverse database featuring the alert; this is, of course, reliant on the nature of a database and is relative only to that and for a specific alert. It will give general information on whether an alert is general in nature i.e., high coverage, or specific, i.e. low coverage. Performance can be assessed with a number of statistical criteria, e.g. Cooper statistics (Cooper et al., 1979), Fisher's exact test (as exemplified in Wang et al., 2019); it is noted that there are few alerts associated with the absence of a given mechanism of toxicity – although they could, for instance, be derived from machine learning – and the "negatives" in Cooper statistics should only be considered when there is a negative alert, but should not be considered for positive alerts with the absence of an alert analogous to a negative outcome, hence prediction of "negatives", i.e. non-toxic molecules, should be ignored in this situation. Dependent on the use of the alert, some scenarios, e.g., low false negative rate, may be preferred. With particular reference to data-driven methods of determining structural alerts, there may be a need to consider the use of test and training sets to assess the performance and significance of an alert if there is no underlying expert knowledge at the outset, similar to the development of

other types of *in silico* models. It is obvious that such statistics are dependent on the quality and extent of any underlying data set as well as how strictly the alert is defined both in terms of chemical structure and physico-chemical properties. As such, these criteria should not be considered to exclude alerts, but will provide an estimate of the confidence provided by associated data i.e. if there are few data to support and alert, it may indicate that further data should be sought.

310 Table 2. Definitions and relevance to uncertainty of the properties relating to structural alerts

Criteria	**Definition and Relevance to Uncertainty**
Purpose	The purpose, or potential use, of the structure alert with regard to regulatory assessment, product development etc. and will usually be stated by the user. For low uncertainty the stated use should be clear and unambiguous e.g., for hazard identification relating to toxicity prediction or to facilitate grouping and read-across. The characteristics of the alert should be appropriate for use.
Structural Description	The functional group, or other chemical substructure, that is defined as the structural alert is unambiguously described including any modulating factors and the local molecular environment e.g., substitution patterns on a ring, branching or unsaturation on an alkyl chain etc. Clear and unambiguous definition will enable transparency and documentation.
Property Domain	The domain of the alert defined in terms of relevant physico-chemical properties (e.g., solubility, volatility), molecular descriptors (e.g. 2D, 3D properties such as dimensions), molecular properties (e.g. toxicokinetics (e.g. clearance) and any other relevant property. It is assumed that the domain of the alert will be defined on the training set, if available.

Toxicity or Relationship to Adversity	The definition of the toxicological effect that is elicited, or the adverse effect that may be related to a MIE or KE in an AOP that is associated with the structural alert. This will provide clear indication of the use of the structural alert.
Species Specificity	The structural alert is associated with effects to a particular species, taxa or group of organisms and, if required, life stage.
Metabolic Domain	Consideration of whether the alert requires, or does not require, metabolic activation.
Mechanistic Interpretation	The structural alert is associated with a recognisable and / or understandable mechanism of toxic action, in addition to, where possible, an AOP.
Mechanistic Causality	The definition of the structural alert in terms of structural chemistry, physico-chemical properties etc., is related to the MIE or KE of the mechanism / AOP in a comprehensible/plausible fashion. If possible, the structural alert should relate to the mechanism of action in terms of the chemistry that underpins the interaction with physiological / biochemical processes. E.g. a structural alert for covalent DNA binding should be related to an organic chemistry reactive mechanism. It is noted that an alert may be mechanistically interpretable, but lacks mechanistic causality.

Coverage	The coverage is the relative proportion of hits a structural alert would have within a defined chemical inventory.
Performance	The performance of a structural alert can be defined in terms of its predictivity, or ability to match compounds known to be associated with that effect. Ideally structural alerts will have a good prediction rate for positives, and low false positive prediction rate. However, this is dependent in part at least on the purpose of the structural alert i.e., toxicity prediction versus grouping or screening.
Corroborating Evidence	The availability of source toxicological, effect or other data that support, or were used to create, the structural alert, e.g., that it may be directly relevant to a toxicological endpoint, adverse effect, MIE etc.
Supporting Evidence	The availability of additional information that may support a weight of evidence approach e.g. data from omics or *in vitro* assays, or data from other endpoints or non-standard tests, that support the structural alert and provide evidence for the mechanism of action or related to an AOP, but which may not have been considered in the development of the alert. Direct mechanistic relevance may be difficult for many endpoints.

311

312

3.2 Scheme to Assign a Confidence Score to a Structural Alert

A key component of the scheme to define uncertainty was the possibility of investigating the (semi-) quantitative assignment of confidence to an alert. To achieve this, the twelve assessable criteria were defined in terms of low, moderate or high confidence, as reported in Table 3. From the outset, it is important to state that low confidence in one or more criteria may be acceptable under certain circumstances. The purpose, in line with the ethos applied by EFSA (2018) with regard to uncertainty, is to highlight areas where improvement in confidence may be achievable to improve the acceptability of a prediction involving a structural alert for a specific purpose. The scheme will also allow for the comparison of the reliability of alerts within, for instance, weight of evidence approaches. Whilst the current scheme assigns confidence into one of three classes i.e., low, moderate and high, it is acknowledged that more classifications could be assigned and an analysis of the advantages and disadvantages of including more classifications is provided by Cronin et al. (2019) with regard to the assessment of the uncertainty of QSARs.

The criteria for assessment of structural alerts have different levels of relevance for a given purpose. Table 4 provides a putative evaluation and ranking of the criteria with regard to their use, for instance, for hazard assessment where a point of departure may, or may not, be required. Some criteria, e.g., the definition of the alert, are essential to the use of an alert. Others could have lower confidence, especially with regard to evidence for causality i.e., mechanistic and metabolic understanding. Whilst mechanistic understanding is desirable, the absence of complete mechanistic understanding should not preclude the use of an established and plausible structural alert. A similar argument can be made of metabolic understanding – i.e., in many cases this may be obvious or can be implied, but lack of complete knowledge of how metabolism may affect an alert should not preclude its use. In addition, metabolic competency may be species dependent and may activate or inactivate a MIE. A further set of criteria, mainly related to the properties and supporting information of the alert, are considered less critical for the evaluation of confidence.

The application of the criteria and relative assessment of confidence is provided for three different types of alerts in the Supplementary Information Tables S1 – S3 respectively. The structural alerts considered are for aliphatic alcohols with reference to acute toxicity across multiple environmental species (taken from Sapounidou et al. 2021 and analogous to that from Verhaar et al. 1992), the ability of aromatic amines to bind to DNA (Enoch and Cronin, 2010) and the inhibition of acetylcholinesterase (AChE) by 1-indanone (Figure 2 in Wedlake et al. (2020)). There are significant differences between these alerts in that those for the aliphatic alcohol and aromatic amine moieties are based on considerable expert knowledge and are well-supported by experimental data. The alert for AChE inhibition is data-driven being derived from data from *in vitro* assays. The differences are reflected in the scores. For instance, Tables S1 and S2 indicate both the alcohol and amine alerts are well defined with a strong mechanistic background. However, low confidence was apparent in the lack of information on coverage and performance. As both alerts are intended for grouping, rather than direct toxicity prediction, this may be deemed acceptable if used appropriately. Table S3 indicates the alert for AChE inhibition has less direct toxicological relevance but is well characterised in terms of coverage and performance.

The relative confidence that can be associated with the three structural alerts is demonstrated graphically as “radar plots” in Fig 1. The two alerts based on expert knowledge (from Sapounidou et al. (2021) and Enoch et al. (2011)) have the same “confidence profile” as defined by the criteria with low confidence associated with the lack of documented coverage and performance of these alerts. The data-driven alert from Wedlake et al. has a different confidence profile, with lower confidence associated with the lack of primary data anchored to the alert.

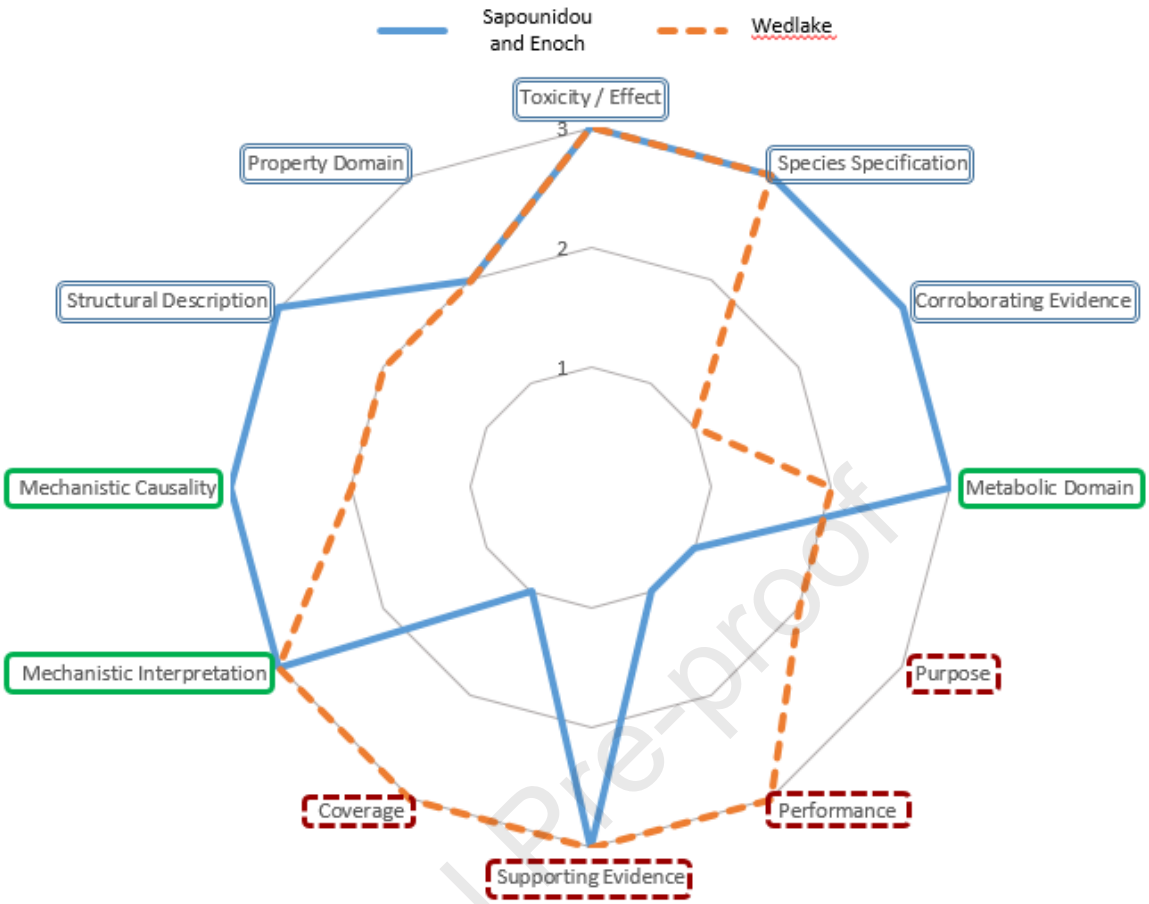

Figure 1. Radar plot representing the "confidence profile" associated with knowledge driven alerts (from Sapounidou and Enoch) in blue as compared to the data-driven alert (from Wedlake) in orange (dashed line). The confidence criteria are ordered according to the relative importance as stated in Table 4, with the essential criteria at the top of the radar plot in blue boxes (double line), moderately importantly in green boxes (single line) at the centre of the plot and lower importance at the base of the plot in red boxes (dashed line). The criteria have been scored from 3 (low uncertainty / high confidence) to 1 (high uncertainty / low confidence).

Given the possibility of ranking the relative criteria for the assessment of the confidence for structural alerts according to their relevance and importance as shown in Table 4, it may also be possible to allocate some type of weighting to create a score for a particular alert that takes account of the particular levels of confidence. This can be converted to give a "confidence score" for a particular alert.

A proposed scheme to add a weighting to each of the criteria is given in Supplementary Information Tables S1 – S3 and has been applied to the three alerts. For clarity, the weightings in Tables S1-S3 correspond to Table 4, i.e. essential criteria are given a weighting of 10, desirable criteria a weighting of 5 and optional criteria a weighting of 2. At this time the weightings are arbitrary and any uptake of such weighting will require consideration with regard to their use and purpose. For instance, it is anticipated that alerts could be aligned with different characteristics for hazard identification (where a highly specific, data rich alert may be required) as opposed to prioritisation and screening (where a broader alert, not necessarily mechanistically-based, may be acceptable). As such, not only different weightings, but different (semi-quantitative) weights could be applied. Where a rapid screening tool is required, for instance for the evaluation of a chemical inventory, then the most relevant characteristics of alerts will be coverage and an understanding of the false prediction rate (particularly the possibility of not identifying particular effects). To assign compounds to a particular QSAR, as in the Sapounidou et al., (2021) scheme, then much greater emphasis will be placed on the mechanistic understanding, or the relevance to the known molecular initiating events.

The weightings in the scheme are on a scale up to 10 with the higher weighting being associated with those criteria deemed more essential in Table 4. Such an analysis has the effect of emphasising the important uncertainties associated with an alert. The weighting has had the effect of emphasising the relatively low confidence (or high uncertainty) that is associated with the alert from Wedlake with regard to its structural description and the lack of primary (*in vivo*) supporting evidence. This, of course, does not preclude the use of this alert, but demonstrates where further information and / or knowledge could be provided to increase confidence in its use. In addition, there may be possibilities for using alerts not supported by *in vivo* data for specific purposes, such as to confirm an MIE. The essentiality of some criteria will depend on the use case, as noted above and it is unlikely that a single list covering all use cases can be developed.

The Mean Confidence Score and a “Weighted Confidence Score” are also provided for the three alerts in Tables S1-S3. The Weighted Confidence Score is calculated as:

$$\text{Weighted Confidence Score} = \frac{\sum \text{weighted confidence scores for each criterion}}{\sum \text{weightings for each criterion}} \quad (1)$$

The resulting scores are on a scale from 3 (greatest confidence) to 1 (lowest confidence). Having a single number for a Confidence Score is in some ways appealing i.e., a number can provide information on confidence, but runs a very high risk of being misleading if misinterpreted. It is not intended that a higher score implies any alert to be "better" than any other alert, but that it may be better defined in certain characteristics which could make it more amenable for various use cases. The Confidence Scores for the knowledge-based alerts (Sapounidou and Enoch) are higher than for the data-driven alert, however this does not take account of other factors such as speed of development. It should also be emphasised that a single score for confidence may mask an unacceptable uncertainty in one, or a small number, of areas. Thus, close examination of radar plots, such as Figure 1, is helpful and inevitably leads to the question of what the desirable characteristics of an alert for a specific purpose are, which is considered in the next section. Since weightings in any scheme are defined by the user, they can be adjusted to emphasise any particular aspect of the evaluation.

415 Table 3. Definitions of the properties relating to structural alerts and their relevance to confidence

Criterion	**Confidence**	**Relevance to the Structural Alert in Terms of Possible Uncertainty Affecting Confidence**
Purpose	High	The purpose of the structural alert is clearly and unambiguously stated, e.g., toxicity prediction or grouping.
	Moderate	The purpose of the structural alert is broad or ambiguous.
	Low	The purpose of the structural alert is not stated.
Structural Description	High	Unambiguous description of the functional group and / or molecular fragment including modulating factors.
	Moderate	Structural alert is loosely defined with regard to its chemical structure with little or no information regarding modulating factors.
	Low	Poor, or no, description of the structural alert with regard to its chemical structure or modulating factors.
Property Domain	High	A well-defined domain in terms of the complete molecular environment and ranges of physico-chemical and / or structural properties.
	Moderate	Some, but incomplete, definition of the domain for the complete molecular environment. No, or incomplete, definition of the ranges of physico-chemical and / or structural properties.

	Low	No, or very ambiguous, definition of the domain for the complete molecular environment and the ranges of physico-chemical and / or structural properties.
Toxicity or Relationship to Adversity	High	The endpoint, toxicity or adverse effect(s) is clearly and unambiguously stated.
	Moderate	The endpoint, toxicity or adverse effect(s) is general and lacks specificity e.g. in terms of organ or species.
	Low	The endpoint, toxicity or adverse effect(s) is not known or stated.
Species Specificity	High	The species, taxa or groups of organisms, in addition to relevant life stage if important, to which the structural alert is relevant are identified and clearly stated.
	Moderate	There is some evidence and documentation that the structural alert is associated with the species to which it pertains.
	Low	No evidence is presented for a species-specific response to the structural alert.
Metabolic Domain	High	The metabolic domain is clearly and unambiguously stated e.g., the alert defines whether a chemical does or does not require metabolic activation.
	Moderate	The metabolic domain is ambiguous or poorly defined.
	Low	The metabolic domain is not known or stated.

Mechanistic Interpretation	High	The structural alert is strongly associated with a well-recognised and documented mechanism of action, e.g., a well-developed or OECD endorsed AOP.
	Moderate	The structural alert is possibly associated with a mechanism of action.
	Low	There is no mechanism of action or no documentation associated with the structural alert.
Mechanistic Causality	High	The chemistry captured by the structural alert is strongly associated with the MIE and / or a KE of the mechanism of action.
	Moderate	There is possible, but unsubstantiated, evidence that the chemistry of the structure may be associated with the mechanism of action, for instance evidence of correlation but not causality.
	Low	The chemistry captured by the structure alert has no documented association with the mechanism of action.
Coverage	High	The structural alert has relatively low coverage of alert-specific chemical space which could imply a limited and well-defined domain.
	Moderate	The structural alert has general coverage of alert-specific chemical space with a moderately broad domain.
	Low	The structural alert has high, or undefined, coverage of alert-specific chemical space indicating a broad, unspecific alert.

Performance	High	A statement relating to the predictive performance of the structural alert to assist in understanding the purpose of the alert, i.e., good performance measured by few false positives / negatives for hazard identification, or biased to ensure few false negatives for screening in a tiered approach.
	Moderate	The structural alert has modest (i.e. greater than random but is not 100% accurate) predictive performance.
	Low	The structural alert is not able to distinguish between active and inactive chemicals.
Corroborating Evidence	High	Multiple and confirmatory toxicological data to support the structural alert.
	Moderate	Few toxicological data exist to support the structural alert.
	Low	No toxicological data are available to support the structural alert e.g. for a statistical approach or one derived on hypothetic mechanisms.
Supporting Evidence	High	Multiple and confirmatory evidence from mechanistic information to confirm the mechanistic hypothesis.
	Moderate	Few data exist to support the mechanistic interpretation of the structural alert.
	Low	No mechanistic information is available to support the structural alert.

416

417

418 Table 4. Proposed relative importance of the confidence criteria in the scheme for the assessment of
419 structural alerts relating to acceptable levels of confidence. In this case, the attributes are for hazard
420 identification supporting risk assessment.

Criteria	**Comment**
Essential Attributes of a Structural Alert – Must be Associated with High Confidence (where possible)	
Structural Description	The alert must be explicitly defined in terms of its chemical structure, structural domain and which species it is relevant to.
Property Domain	
Toxicity or Relationship to Adversity	
Species Specificity	
Corroborating Evidence	
Desirable Attributes of a Structural Alert – Preferably Associated with High Confidence (where possible)	
Metabolic Domain	The mechanistic and metabolic relevance of an alert increases its transparency and potential acceptance.
Mechanistic Interpretation	
Mechanistic Causality	
Optional Attributes of a Structural Alerts – Where Possible Associated with High / Moderate Confidence	

Purpose	Statistical analysis and source data increase the credibility, or otherwise, of a structural alert.
Coverage	
Performance	
Supporting Evidence	

421

422

423 *3.3 Use Cases and the Desired Properties for Structural Alerts*

424 Five use case scenarios for structural alerts are described below, with attributes noted in Table 5.
425 These do not encompass all uses, but are representative of the types of applications for which
426 structural alerts may be used. These include those for regulatory use and industry specific uses,
427 namely:

428 • Hazard identification through direct prediction of toxicity to support risk assessment, e.g.,
429 giving weight to a particular adverse outcome.
430 • Mechanism-based analogue identification, e.g., to select similar compounds or analogues as
431 part of a read-across to enable mechanistic justification, assignment of a particular chemical
432 to a QSAR, such as a reactive or specific mechanism.
433 • Category identification e.g., assigning a compound as a chemical class-based analogue for High
434 Production Volume (HPV) chemicals.
435 • Predictions of effects, or identification of hazard, leading to classification and labelling in a
436 regulatory context.
437 • Screening and/ or prioritisation e.g., to identify or highlight potentially hazardous compounds
438 in a regulatory context or as part of product development.

439

Table 5 provides an estimate of the ideal minimum levels of confidence that might be required for each of the twelve uncertainty criteria. From the outset, it is clear that different levels of confidence are acceptable for different use case scenarios. Those associated with providing input into hazard identification i.e., direct prediction of toxicity and read-across ideally have higher confidence. Lower confidence may be acceptable for screening and prioritisation.

Some characteristics of structural alerts should be definitive regardless of use case and hence be associated with high confidence. Examples of these include the structural description of the fragment, the endpoints to which it relates and the species relevance. In addition, the definition and understanding of confidence in structural alerts suggests that different use case scenarios could potentially utilise different characteristics of structural alerts. The identification of analogues, for instance as a primary categorisation tool for read-across, requires highly defined structural alerts with good mechanistic understanding. The purpose here is to identify very closely related chemicals as defined by their structural alerts that would support a robust argument for similarity. The use of structural alerts for assignment of compounds to chemical classes could have lower confidence in terms of structural definition. This would allow for a larger number of compounds to be grouped together, and associated with this could be lower mechanistic understanding with the possible expectation of sub-categorisation later on to allow for efficient analogue selection. However, for uses such as hazard identification or prioritisation lower confidence may be acceptable to allow for the identification of potential toxicants, with the possibility of false positives being ameliorated by further evidence or testing.

Knowledge from the scheme for the evaluation of confidence of structural alerts can also help indicate how to use alerts. For instance, the aliphatic alcohol and aromatic amine alerts (assessed in Tables S1 and S2 respectively) are associated with low confidence for their coverage and performance, as these statistics are not known. In addition, they can be considered as quite broadly defined, thus likely to capture or identify many analogues in a read-across scenario. In such a situation sub-categorisation is recommended, for instance using similarity indices (Mellor et al., 2019). Thus, the evaluation of

466 confidence through the scheme presented does not preclude the use of any structural alert but will
467 assist in the identification of how and where they can be used optimally and justifiably. Other aspects
468 to be considered are the definition of the various domains i.e., structural, mechanistic and metabolic.
469 As noted above and in Table 4, high confidence in the structural definition is a pre-requisite for use,
470 whilst mechanistic and, in particular, metabolic definition may be more aspirational.

471

472 Table 5. Ideal levels of confidence and characteristics of structural alerts in different use case scenarios. The ideal levels of confidence as defined in Table 3.

Criteria					
	Hazard Identification Supporting Risk Assessment	**Mechanism-Based Analogue Identification**	**Category Identification e.g., Chemical Class-Based Analogue for HPV Chemicals**	**Predictions Leading to Classification and Labelling**	**Screening and/ or Prioritisation**
Structural Description	High	High	High	High	High
Property Domain	High	High	High	Moderate	Moderate
Toxicity or Relationship to Adversity	High	High	High	High	High
Species Specificity	High	High	High	High	Moderate (depends on if alerts is endpoint

					agnostic or endpoint specific)
Metabolic Domain	High / Moderate	High / Moderate	High / Moderate	Moderate	Moderate (now to include metabolites as a key improvement of prioritization for some endpoints)
Purpose	High	High	Moderate	Moderate	Moderate
Mechanistic Interpretation	High	High	High	High	High
Mechanistic Causality	High	High	Moderate/ High	Moderate/ High	Moderate / High
Coverage	High	High	Low	Low	Low
Performance	High	High	Moderate	Moderate	Moderate

Corroborating Evidence	High	High	Moderate	Moderate	Moderate
Supporting Evidence	Moderate	Moderate	Moderate / Low	Low	Low

473

4. Conclusions

A scheme is proposed that characterises the uncertainty associated with structural alerts in an attempt to understand the confidence that may be associated with them. Twelve criteria have been considered that account for the quality and usability of an alert for a specific purpose. These criteria have been ranked according to how essential they are for a particular use case. Assessment of existing alerts suggests that those derived directly from expert knowledge have different uncertainties to those from data-driven analyses. This does not discount any particular method of alert creation, rather these findings can be used to reduce uncertainty through finding further data and information to increase confidence in the use of these predictive approaches as well as allowing for increased confidence on decisions made on the alerts and for benchmarking existing alerts.

Declaration of Interests

No authors declare any conflicts of interest.

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect the views or positions of the US Environmental Protection Agency. Reference to commercial products or services does not constitute endorsement.

References

Allen the, Goodman JM, Gutsell S, Russell PJ (2018) Using 2D structural alerts to define chemical categories for Molecular Initiating Events. *Toxicol. Sci*. 165: 213-223.

Amberg A, Anger LT, Bercu J, Bower D, Cross KP, Custer L, Harvey JS, Hasselgren C, Honma M, Johnson C, Jolly R, Kenyon MO, Kruhlak NL, Leavitt P, Quigley DP, Miller S, Snodin D, Stavitskaya L, Teasdale A,

498 Trejo-Martin A, White AT, Wichard J, Myatt GJ (2019) Extending (Q)SARs to incorporate proprietary
499 knowledge for regulatory purposes: is aromatic N-oxide a structural alert for predicting DNA-reactive
500 mutagenicity? *Mutagenesis* 34: 67-82.

501 Ashby J (1985) Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity.
502 *Environ. Mutagen*. 7: 919−921.

503 Ashby J, Tennant RW (1988) Chemical structure, *Salmonella* mutagenicity and extent of
504 carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by
505 the U. S. NTP. *Mutat. Res.* 204: 17–115

506 Bajot F, Cronin MTD, Roberts DW, Schultz TW (2011) Reactivity and aquatic toxicity of aromatic
507 compounds transformable to quinone-type Michael acceptors. *SAR QSAR Environ. Res.* 22: 51–65.

508 Bauer FJ, Thomas PC, Fouchard SY, Neunlist SJM (2018) High-accuracy prediction of mechanisms of
509 action using structural alerts. *Comput. Toxicol*. 7: 36–45.

510 Becker RA, Dellarco V, Seed J, Kronenberg JM, Meek B, Foreman J, Palermo C, Kirman C, Linkov I,
511 Schoeny R, Dourson M, Pottenger LH, Manibusan MK (2017) Quantitative weight of evidence to assess
512 confidence in potential modes of action. *Regul. Toxicol. Pharmacol.* 86: 205-220.

513 Belfield SJ, Enoch SJ, Firman JW, Madden JC, Richarz A-N, Schultz TW, Cronin MTD (2021)
514 Determination of “Fitness-for-Purpose” of *in silico* models to predict (eco-)toxicological endpoints for
515 regulatory use. *Regul. Toxicol. Pharmacol.* 123: 104956.

516 Benigni R (2021) *In silico* assessment of genotoxicity. Combinations of sensitive structural alerts
517 minimize false negative predictions for all genotoxicity endpoints and can single out chemicals for
518 which experimentation can be avoided. *Regul. Toxicol. Pharmacol*. 126: 105042.

519 Claesson A, Minidis A (2018) Systematic approach to organizing structural alerts for reactive
520 metabolite formation from potential drugs. *Chem. Res. Toxicol.* 31: 389-411.

521 Collier ZA, Gust KA, Gonzalez-Morales B, Gong P, Wilbanks MS, Linkov I, Perkins EJ (2016) A weight of
522 evidence assessment approach for adverse outcome pathways. *Regul. Toxicol. Pharmacol.* 75: 46-57.

523 Cooper JA 2nd, Saracci R, Cole P (1979) Describing the validity of carcinogen screening tests. *Br. J.*
524 *Cancer*. 39: 87-89.

525 Cramer GM, Ford RA, Hall RL (1978) Estimation of toxic hazard – a decision tree approach. *Food*
526 *Cosmet. Toxicol*. 16, 255–276.

527 Cronin MTD, Enoch SJ, Madden JC, Rathman JF, Richarz A-N, Yang C (2022) Review of *in silico*
528 toxicology approaches to support the safety assessment of cosmetics-related materials. *Comput.*
529 *Toxicol.* 21: 100213.

530 Cronin MTD, Richarz A-N (2017) Relationship between Adverse Outcome Pathways and chemistry-
531 cased in silico models to predict toxicity. *Appl. in Vitro Toxicol.* 3: 286-297.

532 Cronin MTD, Richarz A-N, Schultz TW (2019) Identification and description of the uncertainty,
533 variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity
534 prediction. *Regul. Toxicol. Pharmacol*. 106: 90-104.

535 Cronin MTD, Yoon M (2019) Computational methods to predict toxicity. In: Balls M, Combes R, Worth
536 A (eds) *The History of Alternative Test Methods in Toxicology*. Academic Press, London, pp. 287-300.

537 Cui X, Liu J, Zhang J, Wu Q, Li X (2019) *In silico* prediction of drug-induced rhabdomyolysis with
538 machine-learning models and structural alerts. *J. Appl. Toxicol*. 39: 1224-1232.

539 ECHA (European Chemicals Agency), 2017. *Read-across Assessment Framework (RAAF)*.
540 https://echa.europa.eu/documents/10162/13628/raaf_en.pd

541 EFSA (European Food Safety Authority) Scientific Committee, Benford, D, et al., 2018. Guidance on
542 uncertainty analysis in scientific assessments. *EFSA J*. 16: 5123.

543 Enoch SJ, Cronin MTD (2010) A review of the electrophilic reaction chemistry involved in covalent DNA
544 binding. *Crit. Rev. Toxicol.* 40: 728-748.

545 Enoch SJ, Ellison CM, Schultz TW, Cronin MTD (2011) A review of the electrophilic reaction chemistry
546 involved in covalent protein binding relevant to toxicity. *Crit. Rev. Toxicol.* 41: 783-802.

547 Hewitt M, Enoch SJ, Madden JC, Przybylak KR, Cronin MTD (2013) Hepatotoxicity: A scheme for
548 generating chemical categories for read-across, structural alerts and insights into mechanism(s) of
549 action. *Crit. Rev. Toxicol.* 43: 537–558.

550 Hill AB (1965) The environment and disease: Association or causation? *Proc. Roy. Soc. Med*. 58: 295–
551 300.

552 Kalgutkar, AS (2020) Designing around structural alerts in drug discovery. *J. Med. Chem*. 63: 6276-
553 6302.

554 Kalgutkar, AS, Driscoll JP (2020) Is there enough evidence to classify cycloalkyl amine substituents as
555 structural alerts? *Biochem. Pharmacol*. 174: 113796.

556 LaLone CA, Villeneuve DL, Lyons D, Helgen HW, Robinson SL, Swintek JA, Saari TW, Ankley GT (2016)
557 Editor's Highlight: Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS): A web-
558 based tool for addressing the challenges of cross-species extrapolation of chemical toxicity. *Toxicol.*
559 *Sci.* 153: 228–245.

560 Landsteiner K, Jacobs J (1935) Studies on the sensitization of animals with simple chemical
561 compounds. *J. Exp. Med*. 61: 643–656.

562 Madden JC, Enoch SJ, Paini A, Cronin MTD (2020) A review of *in silico* tools as alternatives to animal
563 testing: Principles, resources and applications. *ATLA* 48: 146-172.

564 Maggiora GM (2006) On outliers and activity cliffs - why QSAR often disappoints. *J. Chem. Inf. Model*.
565 46: 1535.

566 Meek ME, Palermo CM, Bachman AN, North CM, Lewis, RJ (2014) Mode of action human relevance
567 (species concordance) framework: Evolution of the Bradford Hill considerations and comparative
568 analysis of weight of evidence. *J. Appl. Toxicol. 34:* 595–606.

569 Mekenyan O, Patlewicz G, Dimitrova G, Kuseva C, Todorov M, Stoeva S, Kotov S, Donner EM (2010)
570 Use of genotoxicity information in the development of Integrated Testing Strategies (ITS) for skin
571 sensitization. *Chem. Res. Toxicol.* 23: 1519–1540.

572 Mellor CL, Marchese Robinson RL, Benigni R, Ebbrell D, Enoch SJ, Firman JW, Madden JC, Pawar G,
573 Yang C, Cronin MTD (2019) Molecular fingerprint-derived similarity measures for toxicological read-
574 across: Recommendations for optimal use. *Regul. Toxicol. Pharmacol.* 101: 121-134.

575 Meyer H (1901) Zur Theorie der Alkoholnarkose. *Arch. Exp. Pathol. Pharmakol*. 46: 338–346.

576 Mori T, Ito T, Liu S, Ando H, Sakamoto S, Yamaguchi Y, Tokunaga E, Shibata N, Handa H, Hakoshima T
577 (2018) Structural basis of thalidomide enantiomer binding to cereblon. Sci. Rep. 8: 1294.

578 Natsch A, Emter R, Gfeller H, Haupt T, Ellis E (2015) Predicting skin sensitizer potency based on *in vitro*
579 data from KeratinoSens and kinetic peptide binding: Global versus domain-based assessment. *Toxicol.*
580 *Sci.* 143: 319–332.

581 Nelms MD, Cronin MTD, Enoch SJ, Schultz TW (2013) Experimental verification, and domain definition,
582 of structural alerts for protein binding: epoxides, lactones, nitroso, nitros, aldehydes and ketones. *SAR*
583 *QSAR Environ. Res.* 24: 695-709.

584 OECD (Organisation for Economic Cooperation and Development), 2007. *Guidance Document on the*
585 *Validation of (Quantitative) Structure-Activity Relationships*. ENV/JM/MONO(2007)2. OECD, Paris, pp.
586 154.

587 OECD (Organisation for Economic Cooperation and Development), 2017. *Revised Guidance Document*
588 *on Developing and Assessing Adverse Outcome Pathways*. ENV/JM/MONO(2013)6. OECD, Paris, pp.
589 32.

590 Overton CE (1901) *Studien über die Narkose zugleich ein Beitrag zur allgemeinen Pharmakologie*.
591 Gustav Fischer: Jena, Germany.

592 Patlewicz G, Simon T, Goyak K, Phillips RD, Rowlands JC, Seidel SD, Becker RA (2013) Use and validation
593 of HT/HC assays to support 21st century toxicity evaluations. *Regul. Toxicol. Pharmacol* 65: 259-268,

594 Patlewicz G, Simon TW, Rowlands JC, Budinsky RA, Becker RA (2015) Proposing a scientific confidence
595 framework to help support the application of adverse outcome pathways for regulatory purposes.
596 *Regul. Toxicol. Pharmacol.* 71: 463-477.

597 Pestana C, Enoch SJ, Firman JW, Madden JC, Spînu N, Cronin MTD (2022) A strategy to define
598 applicability domains for read-across. *Comput. Toxicol*. 22: 100220.

599 Richarz AN, Schultz TW, Cronin MTD, Enoch SJ (2014) Experimental verification of structural alerts for
600 the protein binding of sulfur-containing compounds. *SAR QSAR Environ. Res.* 25: 325-341.

601 Rodriguez-Sanchez N, Schultz TW, Cronin MTD, Enoch SJ (2013) Experimental verification of structural
602 alerts for the protein binding of cyclic compounds acting as Michael acceptors. *SAR QSAR Environ. Res.*
603 24: 963-977.

604 Sapounidou M, Ebbrell DJ, Bonnell MA, Campos B, Firman JW, Gutsell S, Hodges G, Roberts J, Cronin
605 MTD (2021) Development of an enhanced mechanistically-driven mode of action classification scheme
606 for adverse effects in environmental species. *Environ. Sci. Technol.* 55: 1897-1907.

607 Schultz TW, Richarz A-N, Cronin MTD (2019) Assessing uncertainty in read-across: Questions to
608 evaluate toxicity predictions based on knowledge gained from case studies. *Comput. Toxicol*. 9: 1-11.

609 Siramshetty VB, Preissner R, Gohlke B-O (2018) Exploring activity profiles of PAINS and their structural
610 context in target-ligand complexes. *J. Chem. Inf. Model*. 58: 1847-1857.

611 Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A,
612 Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M,
613 Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D,
614 Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko

615 IV (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model
616 development and publishing of chemical information. *J. Comput.-Aided. Mol. Des.* 25: 533-554.

617 Sushko I, Salmina E, Potemkin VA, Poda G, Tetko IV (2012) ToxAlerts: A web server of structural alerts
618 for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf. Model*. 52: 2310-
619 2316.

620 Valsecchi C, Grisoni F, Consonni V, Ballabio D (2019) Structural alerts for the identification of
621 bioaccumulative compounds. *Integr. Environ. Assess. Manage*. 15: 19- 28.

622 Verhaar H J M, van Leeuwen CJ, Hermens JLM (1992) Classifying environmental pollutants. 1.
623 Structure-Activity-Relationships for prediction of aquatic toxicity. *Chemosphere* 25: 471-491.

624 Wang J, Hallinger DR, Murr AS, Buckalew AR, Lougee RR, Richard AM, Laws SC, Stoker TE (2019) High-
625 throughput screening and chemotype-enrichment analysis of ToxCast phase II chemicals evaluated for
626 human sodium-iodide symporter (NIS) inhibition. *Environ. Int.* 126: 377-386.

627 Wang J, Richard AM, Murr AS, Buckalew AR, Lougee RR, Shobair M, Hallinger DR, Laws SC, Stoker TE
628 (2021) Expanded high-throughput screening and chemotype-enrichment analysis of the phase II: e1k
629 ToxCast library for human sodium-iodide symporter (NIS) inhibition. *Arch. Toxicol*. 95: 1723-1737.

630 Wedlake AJ, Folia M, Piechota S, AlltheTEH, Goodman JM, Gutsell S, Russell PJ (2020) Structural alerts
631 and random forest models in a consensus approach for receptor binding Molecular Initiating Events.
632 *Chem. Res. Toxicol.* 33: 388-401.

633 Worth AP (2020) Computational modelling for the sustainable management of chemicals. *Comput.*
634 *Toxicol.* 14: 100122.

635 Worth AP (2010) The role of QSAR methodology in the regulatory assessment of chemicals. In Puzyn
636 T, Lesczynski J, Cronin MTD (eds) *Recent Advances in QSAR Studies: Methods and Applications*.
637 Springer, Dordrecht, The Netherlands, pp. 367-382.

638 Yang H, Lou C, Li W, Liu G, Tang Y (2020) Computational approaches to identify structural alerts and
639 their applications in environmental toxicology and drug discovery. *Chem. Res. Toxicol.* 33: 1312−1322.

640 Yang C, Rathman JF, Magdziarz T, Mostrag A, Kulkarni S, Barton-Maclaren TS (2021) Do similar
641 structures have similar No Observed Adverse Effect Level (NOAEL) values? Exploring
642 chemoinformatics approaches for estimating NOAEL bounds and uncertainties. *Chem. Res. Toxicol*. 34:
643 616-633.

644

A Scheme to Evaluate Structural Alerts to Predict Toxicity – Assessing Confidence By Characterising Uncertainties

Highlights

- Structural alerts are useful tools for predictive toxicology
- 12 criteria to evaluate structural alerts have been identified
- A strategy to determine confidence of structural alerts is presented
- Different use cases require different characteristics of structural alerts

A Scheme to Evaluate Structural Alerts to Predict Toxicity – Assessing Confidence By Characterising Uncertainties

Funding body information

All authors were funded via their own institutions.

Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: