Johnson, C, Anger, LT, Benigni, R, Bower, D, Bringezu, F, Crofton, K, Cronin, MTD, Cross, KP, Dettwiler, M, Frericks, M, Melnikov, F, Miller, S, Roberts, DW, Suarez-Rodrigez, D, Roncaglioni, A, Lo Piparo, E, Tice, RR, Zwickl, C and Myatt, GJ

Evaluating Confidence in Toxicity Assessments Based on Experimental Data and In Silico Predictions

Evaluating Confidence in Toxicity Assessments Based on Experimental Data and *In Silico* Predictions

Candice Johnson[a*], Lennart T. Anger[b], Romualdo Benigni[c], David Bower[a], Frank Bringezu[d], Kevin Crofton[e], Mark T.D. Cronin[f], Kevin P. Cross[a], Magdalena Dettwiler[g], Markus Frericks[h], Fjodor Melnikov[b], Scott Miller[a], David W. Roberts[f], Diana Suarez-Rodrigez[i], Alessandra Roncaglioni[j], Elena Lo Piparo[k], Raymond R. Tice[l], Craig Zwickl[m], Glenn J. Myatt[a]

a) Instem, 1393 Dublin Rd, Columbus, OH 43215, USA
b) Genentech, Inc., 1 DNA Way, South San Francisco, CA, 94080, USA
c) Alpha-PreTox, via G.Pascoli 1, 00184 Roma, Italy
d) Merck Healthcare KGaA , Frankfurter Str. 250, U009/101
e) R3Fellows LLC, Durham, NC
f) School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, L3 3AF, UK
g) Idorsia Pharmaceuticals Ltd, Hegenheimermattweg 91, 4123 Allschwill, Switzerland
h) BASF SE, APD/ET, Li 444, Speyerer St 2, 67117 Limburgerhof, Germany
i) FStox consulting LTD, 2 Brooks Road Raunds Wellingborough NN9 6NS
j) Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy
k) Chemical Food Safety Group, Nestlé Research, Lausanne, Switzerland
l) RTice Consulting, Hillsborough, NC 27278, USA
m) Transendix LLC, 1407 Moores Manor, Indianapolis, IN 46229, USA

*Corresponding author. E-mail address: cjohnson@leadscope.com (C. Johnson)

Table of contents

Abbreviations

IATA – integrated approaches to testing and assessment

DPRA – Direct Peptide Reactivity Assay

h-CLAT – Human Cell Line Activation Test

U-SENS™ – U937 Cell Line Activation Test

Nrf2– NF-E2-related factor 2

ARE – Antioxidant Responsive Element

LLNA – Local Lymph Node Assay

GPMT – Guinea Pig Maximization Test

HMT – Human Maximization test

HRIPT– Human Repeat Insult Patch tests

KE – Key Event

2o3 DA – 2 out of 3 defined approach

CD86 – Cluster of Differentiation 86

RS – Reliability score

SI – Stimulation index

QMM – Quantitative Mechanistic Model

DPT – Diagnostic Patch Testing

CD86 – Cluster of Differentiation 86

Abstract

Understanding the reliability and relevance of a toxicological assessment is important for gauging the overall confidence and communicating the degree of uncertainty related to it. The process involved in assessing reliability and relevance is well defined for experimental data. Similar criteria need to be established for *in silico* predictions, as they become increasingly more important to fill data gaps and need to be reasonably integrated as additional lines of evidence. Thus, *in silico* assessments could be communicated with greater confidence and in a more harmonized manner. The current work expands on previous definitions of reliability, relevance, and confidence and establishes a conceptional framework to apply those to *in silico* data. The approach is used in two case studies: 1) phthalic anhydride, where experimental data are readily available and 2) 4-hydroxy,3-propoxybenzaldehyde, a data poor case which relies predominantly on *in silico* methods, showing that reliability, relevance, and confidence of *in silico* assessments can be effectively communicated within integrated approaches to testing and assessment (IATA).

1. Introduction

Computational tools are increasingly used to either directly support toxicological assessments or contribute to the weight of evidence[1]. The combination of advancements in technology, increasing understanding of toxicological processes, and the availability of robust data to support models lead to improved model predictivity. Currently, several lines of evidence often contribute to an overall endpoint assessment and computational methods are routinely used to fill data gaps. Hence, clarification is needed of the review process that results in a measure of confidence in a hazard assessment. Quantification of confidence is particularly important as it addresses the context in which such assessments can be made. A regulatory submission may require high confidence assessments while a lower level of confidence may be sufficient for other applications, such as for prioritization or screening of chemicals. The level of confidence in an assessment can also provide a basis for planning additional testing.

Myatt et al.[2] introduced a scoring method that assesses the reliability of a hazard identification based on both experimental data and *in silico* approaches. Further, a confidence score, which takes into account the reliability, relevance, and coverage of information was presented. We build on the previous work by Myatt and colleagues by further defining these terms and illustrating how they are considered in practice. When used within frameworks that consider multiple lines of evidence, such as an Integrated Approach to Testing and Assessment (IATA) or the recently published *in silico* protocols[3,4], reliability and relevance depend on whether an experimental result or an *in silico* assessment is being reviewed. The work that follows illustrates the application of these terms and how they are used to assign confidence to an assessment conducted based on experimental data and *in silico* predictions. Using the presented conceptional framework, the hazard assessment for skin sensitization[3] was applied to the analysis of phthalic anhydride (data rich compound) and 4-hydroxy-3-propoxybenzaldehyde (data poor compound). Skin sensitization potential of the two compounds was assessed based on experimental data collected from published literature and on *in silico* predictions generated using Leadscope models. Both the experimental data and *in silico* results were evaluated for their reliability and relevance and a final confidence in the assessment was assigned. The requirements for a transparent expert review or interrogation of model results are highlighted. We demonstrate that the framework facilitates the effective communication of reliability, relevance, and confidence of *in silico* predictions.

2. Conceptual framework

The conceptual framework was previously developed by Myatt et al.[2] We further expand on the definitions of reliability, relevance, and confidence and provide worked examples demonstrating the application of the principles.

2.1 Context

The following terms will be used to facilitate discussion throughout this section: 'experimental level', 'compound level', '*in silico* model level', and '*in silico* prediction level'. Table 1 shows the relationship of these terms either to one another, or the endpoint that is being assessed.

Table 1. Definition of levels at which reliability, relevance, and coverage are considered

Discussion level	**Context of discussion**
Experimental level	Refers to tests/assays. Reliability and relevance at this level describe the relationship between the experimental system and the endpoint, discussed further in sections 2.2.1 and 2.3.1
Compound level	Reliability and relevance at this level describe the relationship between the substance being tested and the experimental system, discussed further in section 2.3.2
In silico model level	Reliability and relevance at this level describe the relationship between the model and the endpoint of interest, discussed further in sections 2.2.2 and 2.3.3
In silico prediction level	Reliability and relevance at this level describes the relationship between the specific *in silico* model and the chemical structure being evaluated, discussed further in sections 2.2.3 and 2.3.3

2.2 Reliability

2.2.1 Experimental level reliability

At the experimental level, the term reliability in its conventional meaning is defined by the Organisation for Economic Co-operation and Development (OECD) and refers to the extent of reproducibility of results within and among laboratories over time for a test performed using the same standardized protocol[5]. This definition addresses primarily experimental studies conducted according to internationally standardized

and validated test guidelines to support regulatory risk assessment. Data generated in non-standard studies, conducted for example within academia, may also be included in hazard identification. In addition to the quality of the test, the availability of adequately described experimental procedures and results contribute to data reliability[4]. Thus, the following factors are considered when assessing the reliability of experimental data[2]:

- Whether the **test** was compliant with internationally accepted best practice guidelines such as, the OECD principles of Good Laboratory Practices (GLP) or Good *In Vitro* Methods Practices (GIVIMP) standards[6],
- Whether the **data** were generated using accepted test guidelines,
- Whether the **data** were available for independent inspection, and the method description was of a high quality allow independent repetition of the experiment if required,
- Concordance with other studies relevant for the assessment,
- Deviations from the test protocol and the transparent discussion of outliers, extreme values, and reliability. Non-standard tests may be supported by further parameters of the test like statistical power, verification of measurement methods and data, and control of experimental variables that could affect measurements. The addition of adequate positive and negative control substances also contribute to the reliability of a test.

There are different degrees of reliabilities ranging from RS1 to RS5, where RS1 is the highest reliability score, Table 2. Reliability scores of RS1 and RS2 are assigned only to experimental data and map to Klimish scores 1 and 2. RS5 (which maps to Klimish scores of 3 or 4) may be assigned to experimental studies that are of lower quality or which deviate markedly from a testing guideline. An expert review of the experimental study may support the conclusion of such studies, which could increase the reliability score to RS3.[2,7] The discussion is limited to experimental data at this point.

198 Table 2. Descriptions of reliability scores[8]

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

199 *2.2.2 In silico model level reliability*

200 *In silico* models are derived from experimental data and therefore model reliability is reflected in the
201 reliability of the training data. However, as opposed to the test method, for which reliability is
202 characterized by intra- and inter laboratory variability for a single compound, for a global *in silico* model
203 the term refers primarily to the accuracy of the prediction for a number of structurally diverse chemicals.
204 Further, experimental variability is embedded in the models and the prediction uncertainty cannot be
205 smaller than the experimental error that is contained in the training set used to build the model. The
206 transparency of the model is considered as it is critical for an expert review of the prediction. The reliability
207 of an *in silico* model is illustrated by the OECD *in silico* model validation principles[9]. According to these

principles, an *in silico* model requires an "unambiguous algorithm" enabling an expert review of the prediction produced by the model (Principle 2) and performance (goodness-of–fit, robustness, and predictivity) of a model demonstrated for a training set and for an appropriate test set (Principle 4).

2.2.3 In silico prediction level reliability

The reliability of an *in silico* prediction measures the extent that an *in silico* result is predictive of an experimental result, within the system which the model predicts. Reliability of an individual model may vary for structurally different chemicals and is higher for a chemical for which structural features are appropriately represented in the training set; in other words, the query compound is sufficiently similar to compounds used for model development. Assessment of the similarity between the query and the training compounds is warranted in models with a defined applicability domain. Further, a higher reliability is assigned to predictions derived from mechanistic descriptors associated with the biological activity underlying the assessed endpoint. Different individual models may have limited predictiveness (reflected in a low RS5 score); however, combining multiple independent models in an ensemble approach may improve predictiveness and thus reliability as compared to single models (RS4), Table 2. An expert review could further increase this reliability to RS3. Myatt et al.,[2,7] provide a more comprehensive overview of the reliability scores.

The following criteria are considered in an expert review of reliability and support the assignment of an RS3 score, which is the highest reliability score that can be obtained for an *in silico* prediction. These criteria are reproduced from Myatt et al., 2018[2].

- Is the chemical within the applicability domain of the model?
- Do structural features map to a diverse group of compounds and is there a potential (reaction) mechanism associated with the feature? If the features map to a congeneric or homologous series, does the test compound belong to this series? Diversity of chemicals matching a feature increases the confidence that the feature is associated with activity.
- Review of training set examples that matches structural descriptors - are other moieties potentially responsible for biological activity?

- The model inherits the reliability of the experimental data from the training set. This implies that the applicability of experimental reliability criteria to the training set examples should be also considered.

- Is there information from the literature to support the assessment?

2.3 Relevance

2.3.1 Experimental level relevance

Experimental level relevance describes whether a method is meaningful and useful for a purpose and is the extent to which a test correctly measures/predicts the effect/mechanism of interest in general terms, not at a specific compound level. For example, an assessment of skin sensitization can include skin permeability. However, predictivity of the test for this specific endpoint is limited and thus relevance of the assessment is low if no other experimental data are available. Relevance also includes a consideration of the accuracy (e.g., its sensitivity and specificity) of a test.[7] Experimental level relevance criteria to be considered when assessing the results from an experimental study also include whether the reported species and experimental endpoints are appropriate for regulatory purposes.

2.3.2 Compound level relevance

The limitations of a test method are also considered aspects of relevance.[5] Typically, method-related limitations are observed at the compound level and may sometimes expand across a chemical class.

The following is a non-exhaustive list of compound level relevance criteria to be considered when assessing the results from an experiment study.

- Does the test article represent the substance being assessed? For example, if an active ingredient only makes up 5% of an organic solvent-based formulation, it is difficult to attribute the activity to an individual ingredient.

- Were appropriate doses/concentrations tested?

- Did the test designed take into consideration the physical and chemical properties of the compound (e.g., purity, stability, solubility)?

- Did the test system cover the mechanism of activity targeted by the compound?

- Did the test system provide metabolic capability adequate for the compound, if required?

In some cases, the relevance criteria outlined above are addressed in a test guideline and it is important to note whether or not deviations from these criteria also lead to non-adherence to the test guideline (a measure of reliability) so that the same study limitation is not overly weighted in the overall assessment of confidence.

2.3.3 In silico model and prediction level relevance

A (Q)SAR model's relevance is based on the relevance of the mechanism or effect that the model predicts and so the (Q)SAR model inherits the relevance of the experimental system. A model built on human effect data; for example, may be considered more relevant than one which predicts the result of an animal study or *in vitro* assay. In lieu of human effect models, multiple mechanisms that lead to a biological effect and therefore multiple (Q)SARs or combinations thereof in respective AOPs may be needed to predict more complex endpoints. As such, an *in silico* prediction could be considered relevant when derived from training set data that are obtained from experimental studies that adhere to experimental level relevance criteria.

The degree of relevance is considered in deriving an assessment of confidence, Section 2.5. Similar to reliability, an evaluation of relevance is conducted during an expert review. The relevance of an assessment may be decreased based on expert review findings. However, if the expert review does not identify any limitations in the relevance of the study, the assessment is considered with standard relevance. Table 3 provides a summary of the discussion on reliability, and relevance.

285

286 Table 3. Definitions summarizing reliability, and relevance at various levels of discussion

	Experimental level	Compound level	*In silico* model level	*In silico* prediction level
Reliability	The reproducibility of results within and among laboratories over time for a test performed using the same standardized protocol	Not applicable	The accuracy of the prediction for a number of structurally diverse chemicals	The extent that an *in silico* result is predictive of an experimental result, within the system which the model predicts
Relevance	Whether a method is meaningful and useful for a purpose and is the extent to which a test correctly measures/predicts the effect/mechanism of interest	The limitations of a method for testing a specific compound	A (Q)SAR model's relevance is based on the relevance of the mechanism or effect that the model predicts	An *in silico* prediction could be considered relevant when derived from training set data that are obtained from experimental studies that adhere to experimental level relevance criteria.

287

288 ## 2.4 Completeness of information

289 Assessment of a specific regulatory endpoint assumes evaluation of a number of toxicology studies and
290 other tests (experimental results or *in silico* predictions). This reflects the fact that a number of
291 toxicological manifestations are associated with one endpoint. In addition, multiple mechanisms could
292 trigger the same toxicological manifestation. A generic hazard assessment framework proposed by Myatt
293 et al.[2] illustrates principles how the toxicological information is assembled within the assessment. This
294 framework has been implemented in the assessment of specific regulatory endpoints: genotoxicity[3] and

skin sensitization[2]. It is important to consider that most of the possible pathways by which the apical endpoint can occur are being evaluated. This coverage of molecular pathways and effects is given consideration when evaluating the confidence in the assessment of the apical endpoint.

2.5 Confidence

The reliability, relevance, and coverage of information determine the level of confidence in the assessment. Confidence could be logically defined into categories of high, medium, low, or no confidence. The following definitions apply to the levels of confidence.

- A high confidence rating suggests that there is sufficient evidence that the assessment provided an accurate conclusion, and further research is unlikely to increase the confidence.
- A medium confidence rating suggests that there is adequate evidence that the assessment provided an accurate conclusion, but further research might increase the confidence.
- A low confidence rating suggests an accurate conclusion is lacking and further research is needed to support a robust conclusion and to improve its confidence.
- A no confidence rating suggests that further research is needed in order to derive an assessment.

While not appropriate for the regulatory submissions, the low confidence rating could be useful for prioritization, identification of the most relevant testing candidates, and to determine data gaps. Typically, in the case of no confidence, data are either unavailable, discordant with no supporting information, or there is no relevance/reproducibility. While decisions cannot be made in these cases, the data may be useful for discussion as seeking solutions may advance testing paradigms. In all cases, a weight of evidence analysis by an expert is suggested.

3. Case Studies

The following sections describe the analysis of phthalic anhydride and 4-hydroxy-3-propoxybenzaldehyde using an implementation of the skin sensitization protocol[3], Leadscope Enterprise version 3.8, skin sensitization integrated hazard assessment (v1.0). Version 1 of the skin sensitization hazard assessment includes the following statistical models: Direct Peptide Reactivity Assay (v1.0), Human Cell Line Activation Test (h-CLAT) (v2.0), KeratinoSens ™ (v2.0), Local Lymph Node (v.2.0). The following alerts sets are also included: Local Lymph Node Assay Expert Alerts (v2.0), Reaction Domain Alerts (v1.0). Here we note that

in the derivation of the skin sensitization *in vitro* endpoint, the '2 out of 3' defined approach (2o3 DA) to skin sensitization hazard identification is used in relation to OCED TG 497[10], and within the IATA defined by Johnson et al., 2020[3] which includes an analysis of the structure activity relationship of the test structure with known examples, and an evaluation of other adverse outcome pathway (AOP) endpoints.

The principles describing reliability, relevance, and coverage, which were described above, are applied to the phthalic anhydride and 4-hydroxy-3-propoxybenzaldehyde cases to provide practical examples of the confidence derivation. Further, reliability scores described in Myatt et. al.[2] are used to communicate the reliability of the assessments.

3.1 Skin sensitization hazard assessment framework

The skin sensitization hazard assessment framework will be used to illustrate, through two case studies, how the previously described reliability, relevance, coverage, and confidence, may be assessed. Throughout these discussions, experimental data will be identified and evaluated. In addition, different *in silico* models will also be used. They include statistical-based models built on named substructural features and phys-chem properties descriptors, that generate a probability of a positive value. This probability is translated into a positive/negative prediction using cut-offs. For example, a prediction greater than 0.5 is assigned to positive and less than 0.5 assigned to negative, but for value close to 0.5 the uncertainty may be higher based on the distribution of predictivity. An assessment of chemical similarity may be used to rank analogs based on their structural similarity to the test chemicals. For this assessment, the chemical structures represented by molecular fingerprints converting structural features into bit vectors[11–14]. These abstract representations of chemicals allow easy computational processing and comparison. Chemical dissimilarities can be calculated by standard methods applying Tanimoto, Dice, or equivalent distance measures[15,16]. However, it should be noted that similarity scores calculated using different methods may give different results and agreement between different methods applied could increase confidence in the similarity assessment. Other factors, such as water solubility, molecular size, pKa and log K_{ow} should also be considered in accordance with the OECD guidance on grouping of chemicals[17].

Figure 1 shows the hazard assessment framework for skin sensitization[3]. The mechanisms and effects that were assessed in the following examples include: protein reactivity, activation of biochemical

pathways (Nrf2-ARE pathway), expression of co-stimulatory and adhesion molecules, rodent LLNA proliferation, rodent maximization, human skin sensitization (gray boxes). These were assessed using either experimental data and/or *in silico* models. An expert review was performed on the study data and the *in silico* predictions and a reliability score was assigned to the assessment. The results of the individual assessments and their corresponding reliability scores were used to assess the toxicological endpoints related to skin sensitization and to assign confidence scores (blue boxes). Relevance and coverage were also considered in the evaluation of the confidence level as highlighted by the following examples.

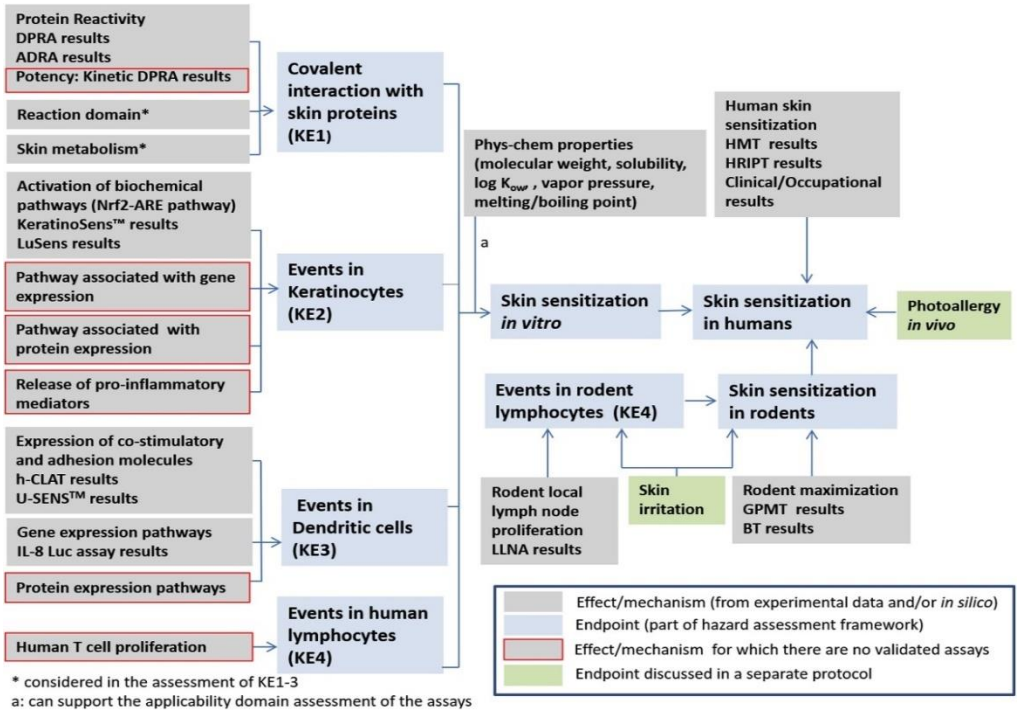
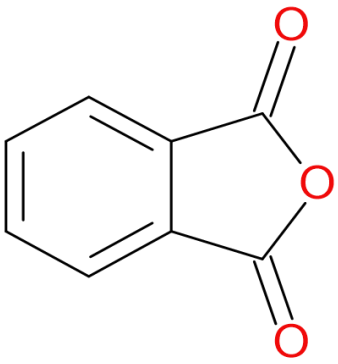


Figure 1. Skin sensitization hazard assessment framework[3]

3.2 Phthalic anhydride case study

3.2.1 Chemistry

Phthalic anhydride (CAS# 85-44-9) is a white solid used in the synthesis of resins and plastics.[18] The chemical structure is shown in Figure 2.



Property	Value
Molecular weight	148.1 g/mol
Water solubility	6.2 g/L @ 25°C (experimental)
Log K_{ow}	1.6 (measured)

Figure 2: Chemical structure and properties of phthalic anhydride[18]

3.2.2 Covalent interaction with skin proteins

The Direct Peptide Reactivity Assay (DPRA) is an *in chemico* method addressing covalent binding to proteins which is the Molecular Initiating Event (MIE) in the skin sensitization AOP. As an *in chemico* test, the DPRA lacks the ability to predict the activity of chemicals that are metabolically transform to a reactive species.

The DPRA test has been conducted with phthalic anhydride and the results were published in a peer reviewed scientific journal. The study returns a positive result and indicates high reactivity, with a cysteine depletion value of 1.9% and a lysine depletion value of 75%[19,20]. However, the GLP status of the study was not disclosed and despite the detailed description of the method and results, not all information as required by the test guideline, was provided. The study adheres to established test guideline OECD TG 442C[21]. Consequently, due to the high reliability of the experimental method but lack of GLP status and a study report, the data was assigned a reliability score of RS2.

The relevance of the method for predicting a potential of the compound to bind proteins has been well established with the limitations discussed in the guideline[21]. One of the limitations potentially applicable to phthalic anhydride is its low stability in aqueous solution due to a rapid hydrolysis to phthalic acid (non-

sensitizer)[22]. Low stability of the compound in the test conditions can cause false negative results. In the view of a positive result with phthalic anhydride, this reservation did not affect the relevance of the test at the compound level. Further, chemical properties of the compound were evaluated by an expert. Phthalic anhydride is assigned to the acyl transfer mechanism, RS5 (Figure 3). This reaction mechanism is supported by the preferential reactivity of anhydrides with lysine substantiating relevance of the proposed mechanism.

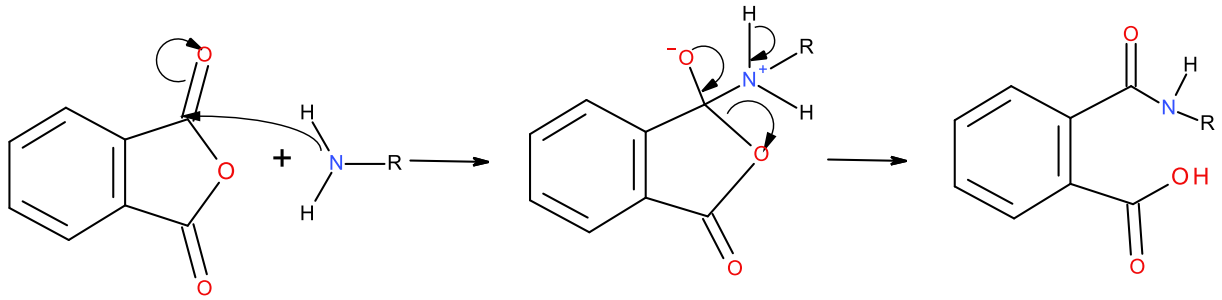


Figure3. Reaction of phthalic anhydride with lysine

Phthalic anhydride has the potential to covalently bind to skin proteins based on the experimental results generated in a DPRA test and expert review of the compound chemical properties. Evaluation of reliability and relevance in this instance lead to a high confidence in the conclusion (as shown in Figure 5).

3.2.3 Events in Keratinocytes

Key Events (KE) within skin sensitization AOP include inflammatory response and changes in gene expression associated with specific cell signaling pathways such as those regulated by binding of the NF-E2-related factor 2 (Nrf2) to antioxidant responsive element (ARE). The KeratinoSens™ assay addresses this mechanism. Experimental data were available for the assessment of the activation of Nrf-2-ARE pathways through the KeratinoSens™ test method[20]. The study adheres to OECD TG 442D[23]. The negative results were assessed and are assigned a reliability score of RS2 due to the sufficient reliability of the study.

While the experimental level relevance is well established for the KeratinoSens™ assay, a review of the compound level relevance is important. The KeratinoSens™ assay is driven by the modification of a cysteine moiety. Chemicals that belong to the acyl transfer reaction domain are hard electrophiles which preferentially bind hard nucleophiles such as lysine[20,24]. Further, any adduct formed via interaction of the

phthalic anhydride and the SH groups of cysteine may be hydrolyzed. Although the KeratinoSens™ assay is applicable to these compounds, the relevance of the test for compounds that react via acyl transfer compounds, especially if they are shown to preferentially bind lysine in the DPRA, is reduced based on the decreased predictivity within this domain[3,20,25]. The decreased compound level relevance of the KeratinoSens™ assay for the assessment of phthalic anhydride leads to a low confidence in the activation of the events in keratinocytes.

3.2.4 Activation of Dendritic Cells

Activation of dendritic cells is another KE in the skin sensitization AOP. Methods developed to address this KE are based on expression of the specific cell surface markers, chemokines and cytokines. These methods include the human cell line activation test (h-CLAT) and the U937 cell line activation test (U-SENS™). Phthalic anhydride has been evaluated in h-CLAT and U-SENS™ tests and the data were published in peer-reviewed journals[26,22]. Both tests provided negative results. The h-CLAT test has been generally conducted as recommended in the validated OECD TG 442E guideline. Adherence to the GLP standards was not addressed in the publication. Further, method and results were missing some details required by the guideline. Consequently, score RS2 was assigned to reliability.

The experimental level relevance of the method for assessing skin sensitization has already been established[27]. Compound level relevance considers whether the appropriate concentrations were tested. This question is particularly pertinent if the result is negative as with the h-CLAT result. A review of the study indicates that phthalic anhydride solubility in DMSO and culture medium was limited and this could have affected the maximal achievable dose[26]. In addition, phthalic anhydride hydrolysis by the aqueous vehicle is suspected to occur in the h-CLAT[28]. The exposure of the THP-1 cells to the anhydride is therefore an unknown parameter that introduces uncertainty around the negative result. Although the study was reliable (RS2) based on adherence to OECD TG 442E, the compound level relevance is reduced based on the above discussion. Information on the available *in vitro* test concentration compared to the potential concentration in the skin could provide additional support for this conclusion.

The U-SENS™ test is the second method recommended in the OECD 442E guideline. Also, for this study GLP status was not addressed in the publication. However, the study was conducted according to the test guideline and the publication contained sufficient details supporting an evaluation of the study conduct

and the validity of the results. Included controls supported evaluation of the method performance. Finally, acceptance criteria were provided. Therefore, a reliability score RS1 has been assigned to the experimental data despite the lack of a GLP study report. The hydrolysis of phthalic anhydride in the culture medium was indicated as the reason for the negative result. Similar to the discussion above for the analysis of phthalic anhydride in the h-CLAT test, the compound level relevance of the U-SENS™ test could be challenged. Additionally, a statistical model was used to predict the activation of the dendritic cells, ((Human Cell Line Activation Test (h-CLAT) (v2.0)). The statistical model returned a positive result with a predicted probability of 0.612.

The studies from which the training set examples are derived adhered to OECD 442E and so the training set examples are reliable. Reliability of the model was strengthened by the details provided in the prediction enabling expert review. The prediction was considered reliable because the compound was within the applicability domain of the model. Consequently, a reliability score of RS3 is assigned to the *in silico* result.

Features of the training set compounds triggering the prediction were reviewed to assess the relevance of the prediction. The oxolane and the anhydride features contributed significantly to the prediction. Note that other contributing features were also identified but are not discussed in detail in the context of this manuscript. The oxolane feature mapped to three training set examples and carried an overall positive weight in the assessment, Figure 4. Propylidene phthalide (LS-933; CAS# 17369-59-4) and Tween 80 (LS-2298; CAS# 9005-65-6) were positive in the h-CLAT[26] and U-SENS™[22] respectively, while Streptomycin sulfate (LS-1247; CAS# 3810-74-0) was negative in both tests[22,26]. These positive results could be explained through characteristics that are not related to the oxolane feature. LS-933 is expected to either react via an acyl transfer mechanism or autoxidize to a hydroperoxide[29]. LS-2298 is negative in the h-CLAT[26] and positive in U-SENS™[22]; however, given that LS-2298 is a surfactant, the U-SENS™ positive result could be due to disruption of cell membranes rather than sensitization related expression of Cluster of Differentiation 86 (CD86)[22]. This brings into question the relevance of the oxolane feature. The anhydride feature maps to trimellitic anhydride (LS-215; CAS# 552-30-7) and maleic anhydride (LS-458; CAS# 108-31-6), both recorded with a positive result. The two training set examples are closely related to phthalic anhydride as they contain the cyclic anhydride moiety through which sensitization may occur; maleic

anhydride may also sensitize through a Michael acceptor mechanism. Overall, the anhydride feature is considered relevant; however, a limitation is realized in that there are only two examples. LS-215 is considered a close analogue of phthalic anhydride and one of particular value, (structure shown in Figure 4). LS-215 differs from phthalic anhydride by the addition of a carboxylic group on the benzene ring. The addition of the carboxylic acid group is not expected to mitigate the sensitization of the anhydride and thus supports the positive prediction of phthalic anhydride. Given the mechanistic similarity between phthalic anhydride and the two anhydrides identified by the model, the positive prediction appears relevant.

'The activation of Dendritic cells' is assessed as positive, with medium confidence. The medium confidence level reflects the uncertainty in the use of an *in silico* prediction compared to reliable and relevant experimental data. In this case, while the experimental data were reliable (the negative assessment could be reproduced), the relevance of the anhydride to the experimental systems was challenged.

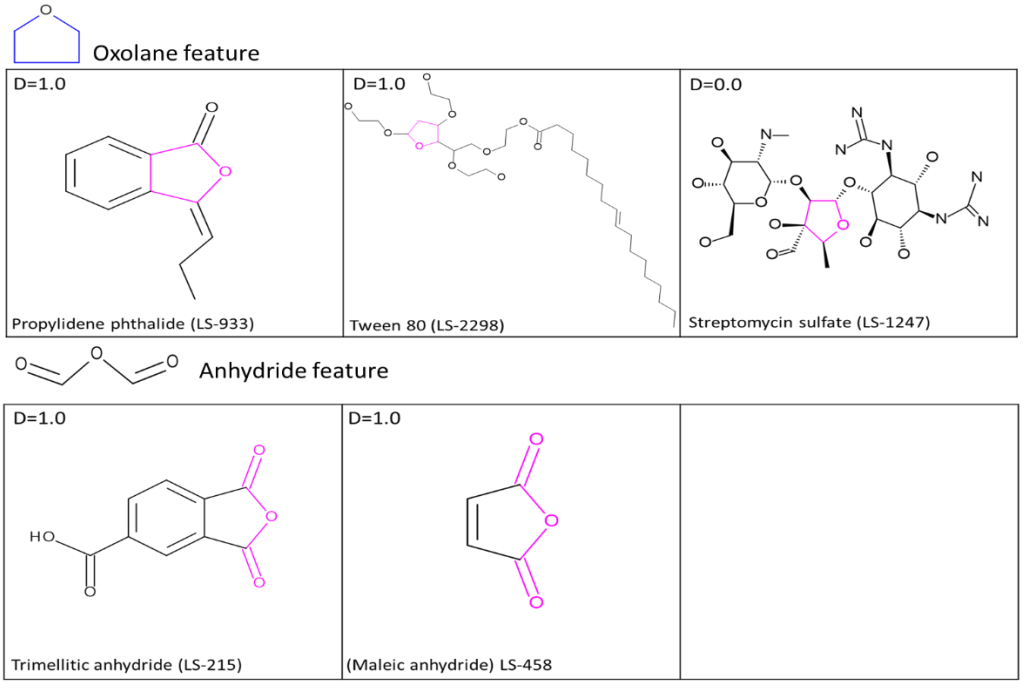
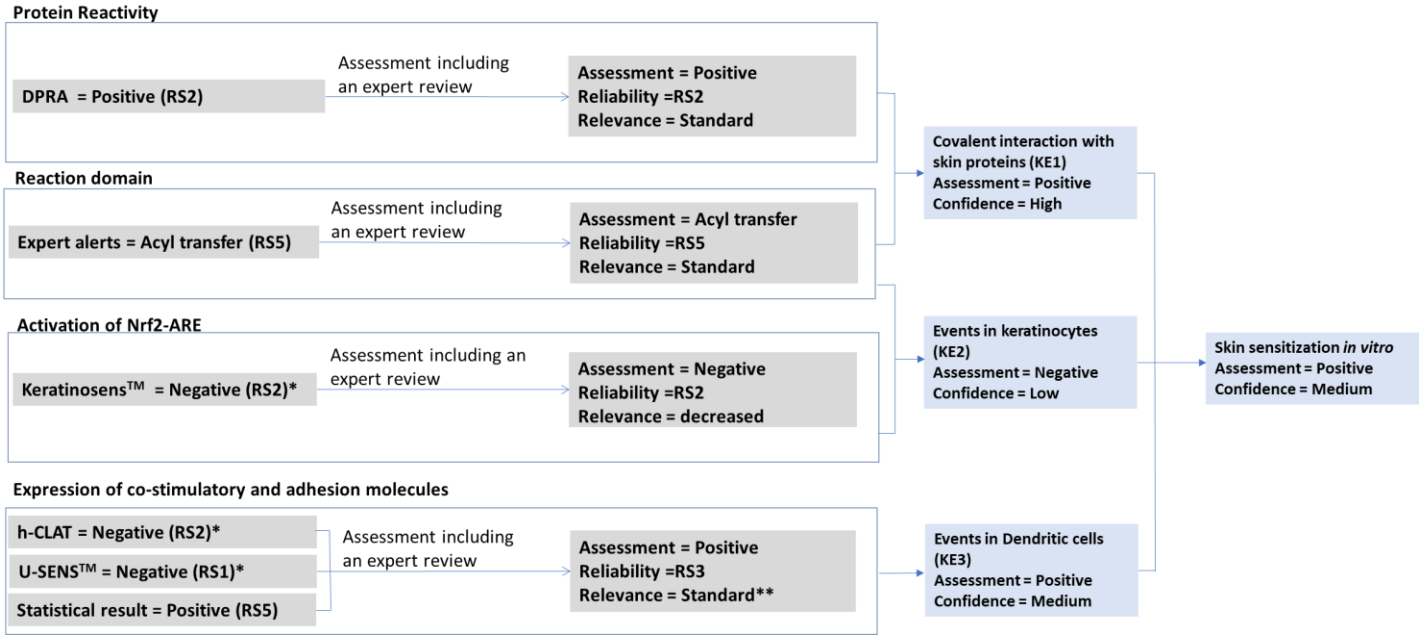


Figure 4. Examples that map to the oxolane and anhydride features for the dendritic cell activation. D=1.0 refers to compounds with a positive response in the experimental test while the result D=0.0 refers to compounds with a negative response in the experimental test.

3.2.5 Endpoint: Skin sensitization in vitro

The skin sensitization *in vitro* endpoint considers the body of evidence presented for KE1 (the molecular initiating event) in addition to KE2 and KE3. Figure 5 summarizes the results for the *in vitro* endpoints. The weight of evidence points to a skin sensitization potential for phthalic anhydride. The lower confidence scores of the two concordant assessments (medium) is adopted as a conservative measure. While a medium confidence score is obtained at the *in vitro* level and reflects the difficulty in assessing unstable (hydrolytic and poorly soluble) substances in experimental systems, the *in silico* tools provide an additional perspective through analysis of similar analogs.



* The relevance of these studies was decreased after the expert review highlighted limitations to testing phthalic anhydride in the experimental systems.

** The standard relevance and RS3 score were assigned to the positive statistical model result.

Figure 5. Derivation of the skin sensitization *in vitro* assessment of phthalic anhydride given the reliability, relevance, and confidence of the supporting assessments

3.2.6 Events in rodent lymphocytes

The last KE in the skin sensitization AOP is T-cells Activation/Proliferation. The effect can be evaluated in the *in vivo* mouse LLNA, which measures primary proliferation of lymphocytes in the auricular lymph nodes following local administration of the test compound to the ear.

Phthalic anhydride (AlogP =1.0) has been tested in the LLNA and has been shown to be a strong sensitizer with reported effective concentrations inducing a stimulation index (SI) of 3 (EC3 values) of 0.16%[30] and 0.36%[31]. These EC3 values are consistent with what would be expected from the well-known high reactivity of anhydrides as acylating agents. The data presented in Dearman et al.[30] were available for an independent review as a publication in a peer-reviewed journal. The study followed general principles included in the OECD TG 429 guideline [30,32]. As discussed in previous section, documentation of the study procedures and results in this form provide a reliability score RS2. Kimber et al.[31] provided an EC3 value but no reference to the original study and thus study detail were not available for review triggering reliability score RS5.

When adequate experimental data are available, *in silico* results may provide information to support the assessment. Statistical and expert alert models support the positive result. An expert review returned two closely related anhydrides, hexahydrophthalic anhydride (AlogP = 0.88, EC3 = 0.84%[30]) and trimellitic anhydride (AlogP = 0.7, EC3 values of 0.6%[30], 0.11%[33] and 9.2%[34]). These both have only the cyclic anhydride entity as a reactive sub-structure and are both strong/moderate sensitizers in the LLNA. Given the comparable AlogP values for the anhydrides and that additional substructures do not support mitigation of the sensitization potential, the positive assessment is supported. Such an analysis could be considered as part of an expert review of any model output.

Consequently, phthalic anhydride was concluded to activate T-cells proliferation and a high confidence was assigned to the assessment of this endpoint based on the reliable and relevant data from an *in vivo* study supported by the concordant result of an *in silico* approach.

3.2.7 Guinea Pig Maximization

Guinea Pig Maximization Tests (GMPT) provide information to support the assessment of the skin sensitization potential of a compound by a direct measurement of this endpoint after epidermal application of the test compound to animals. Phthalic anhydride was subjected to the GMPT performed according to the standard procedures of Magnusson and Kligman[35] and was classified as an extreme/strong sensitizer.[36,37] The studies were published in peer-reviewed journals. The Basketter and Scholes 1992 study reported that phthalic anhydride induced sensitization in 90% of the animals tested at an intracutaneous injection concentration of 0.1%, induction patch concentration of 25%, and a challenge patch concentration of 10%. While this study is similar to published guidelines, data are lacking on the number of animals used as well as the solvent controls and so the reliability of the information is assigned at an RS5 level.

3.2.8 Endpoint: Skin sensitization in rodents

This step considers altogether the results discussed in 3.2.6 and 3.2.7. The LLNA measures the increase in lymph node proliferation associated with application of the test chemical and reports that as an index of induced sensitization. The Guinea Pig Maximization Test (GPMT) assesses, by challenges applied to the skin and subsequent evaluation of the challenge sites, whether skin sensitization has been induced. Phthalic anhydride is positive in both LLNA and GPMT methods. Although there may be more than one biological mechanism at play, involving different pathways and cell sub-populations[30,38], the dermal application in the GPMT challenge indicates that sensitization to dermal tissues occurs as a result of phthalic anhydride exposure. Based on the LLNA and GPMT data, the 'Skin sensitization in rodents' endpoint is assessed as positive with high confidence, Figure 6.

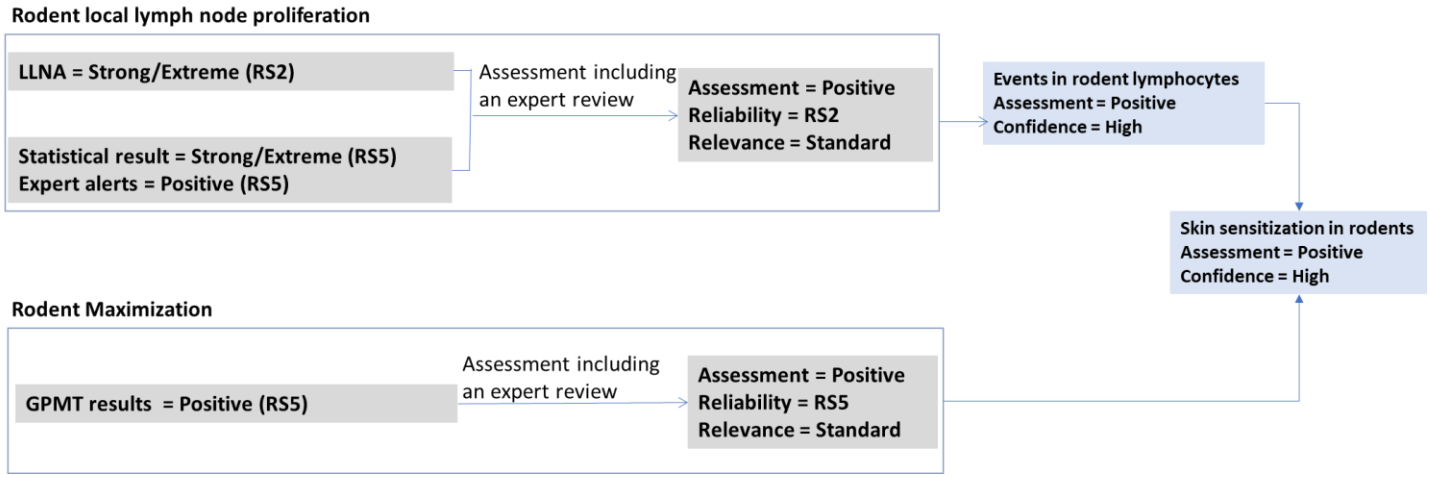


Figure 6. Derivation of the skin sensitization in the rodent assessments of phthalic anhydride given the reliability, relevance, and confidence of the supporting assessments

3.2.9 Human skin sensitization

There is a paucity of Human Maximization test (HMT) and Human Repeat Insult Patch tests (HRIPT) data on the occurrence of sensitization due to phthalic anhydride. ICCVAM (2010) indicates that phthalic anhydride is a skin sensitizer and was assessed either from a HMT, inclusion of the test substance in a human patch test allergen kit, and/or published clinical case studies/reports.[39] The data were not found in the publication referenced. A reliability score of (RS5) was assigned to this data. Allergy to a combination of phthalic anhydride, trimellitic anhydride, and glycol copolymer has been reported in three patients, which were negative to phthalic anhydride alone.[40] However, details on the tested concentrations of phthalic anhydride itself were not provided. Additional studies describe positive reactions to the phthalic anhydride, trimellitic anhydride, and glycol copolymer combined in nail polish without describing results on phthalic anhydride alone[41]. Overall, the results of the human studies are inconclusive, given conflicting pieces of evidence with incomplete information.

3.2.10 Endpoint or overall assessment: Skin sensitization in humans

This apical endpoint takes all available assessments into consideration. The *in vitro* assessments, supported by structure-activity based assessments and the experimental studies in rodents all indicate that phthalic anhydride has the potential to sensitize. The skin sensitization of phthalic anhydride was, therefore, assessed as positive with high confidence, as shown in Figure 7. The different mechanisms involved in the assessment are well covered (apart from the human skin sensitization) and reasons for conflicting data (lack of Activation of the Nrf2-ARE pathway) are explained so the confidence is high.

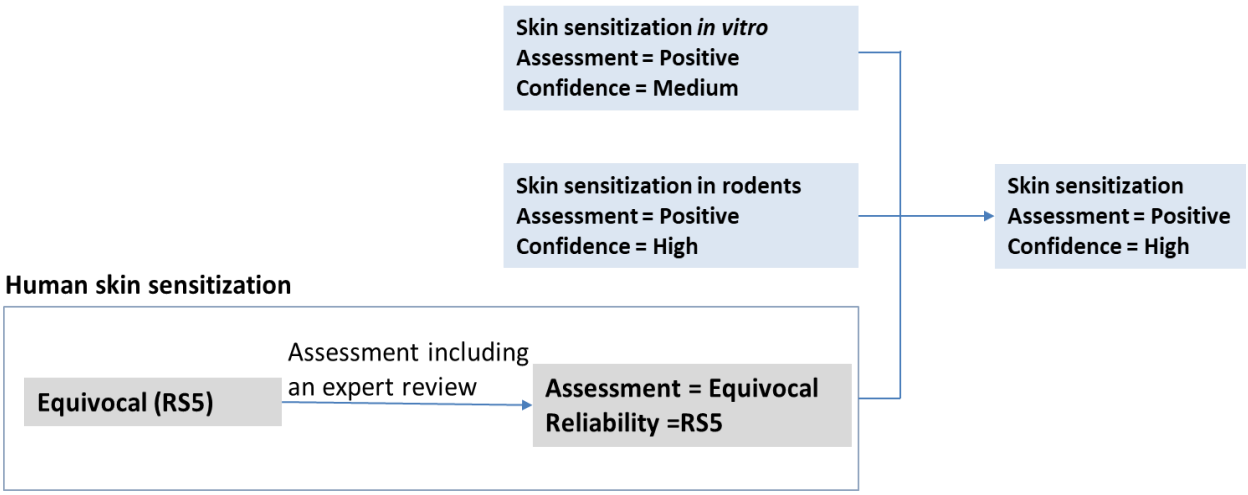
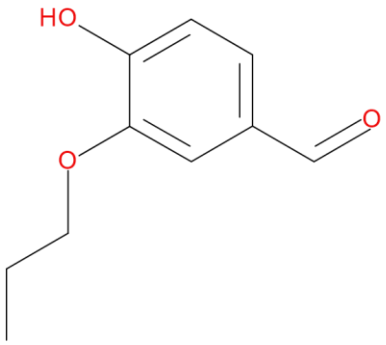


Figure 7. Derivation of the overall skin sensitization assessment of phthalic anhydride given the reliability, relevance, and confidence of the supporting assessments

3.3. 4-hydroxy-3-propoxybenzaldehyde

3.3.1 Chemistry

The chemical structure of4-hydroxy-3-propoxybenzaldehyde (CAS# 110943-74-3) is shown in Figure 8. This second example presents a review of reliability and relevance for model predictions in a data poor situation, and where the data for the closest analogs (vanillin and methyl vanillin) are available. The analogs were selected based on structural similarity and homology with 4-hydroxy-3-propoxybenzaldehyde. Similarity was assessed by Tanimoto scores based on Leadscope fingerprints, and were 0.84 and 0.94 for vanillin and ethyl vanillin respectively. While experimental data are available for the analogs, no data are available for 4-hydroxy-3-propoxybenzaldehyde. Therefore, *in silico* analyses were used to assess the relevant mechanisms and effects.

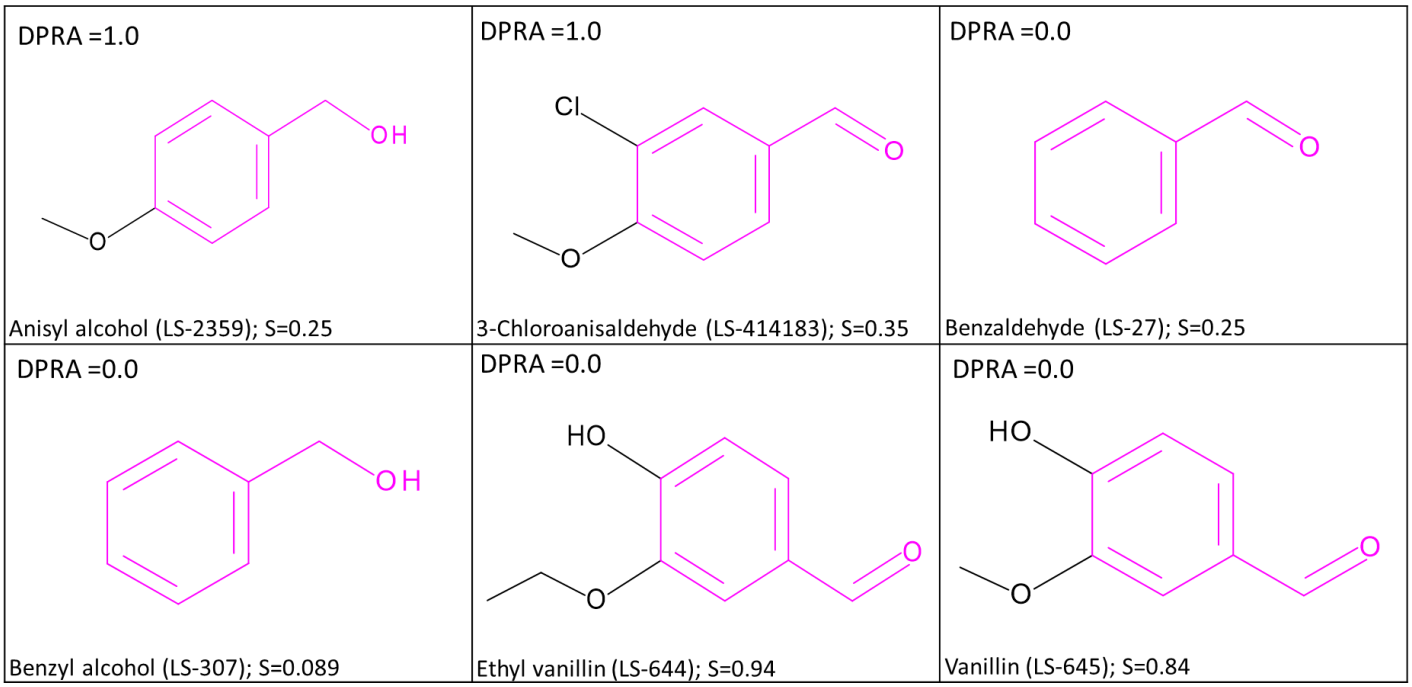


Property	Value
Molecular weight	180.2 g/mol

Figure 8: Chemical structure and properties of 4-hydroxy,3-propoxybenzaldehyde

3.3.2 Covalent Interaction with skin proteins

A statistical model (Direct Peptide Reactivity Assay (v1.0)) predicting the reactivity classes of the DPRA was used to assess potential for covalent binding to proteins. The model returned a result of 'No or minimal reactivity', with a predicted probability value of 0.017. An expert review was conducted to evaluate the reliability and relevance of the prediction. The initial stages of the assessment consider whether the chemical structure is within the applicability domain of the model. A structure is within the applicability domain of Leadscope's statistical model if there is at least one structural feature identified by the model and one analog with a similarity score of 0.3 or greater. The score of 0.3 is based on Leadscope's 27,000 sub-structural features and hence will be lower than similarity scores that use smaller feature sets. There were 2 structural features identified by the statistical model and 11 analogs with similarity scores greater than 0.3, indicating that the compound was within the applicability domain of the model. Note that these analogs indicate that the structure belongs within a chemical neighborhood which is characterized by the model and these analogs are not necessarily used in the prediction. The training set examples are mostly aromatic aldehydes, hydroxybenzene derivatives, and two benzyl alcohols, Figure 9.



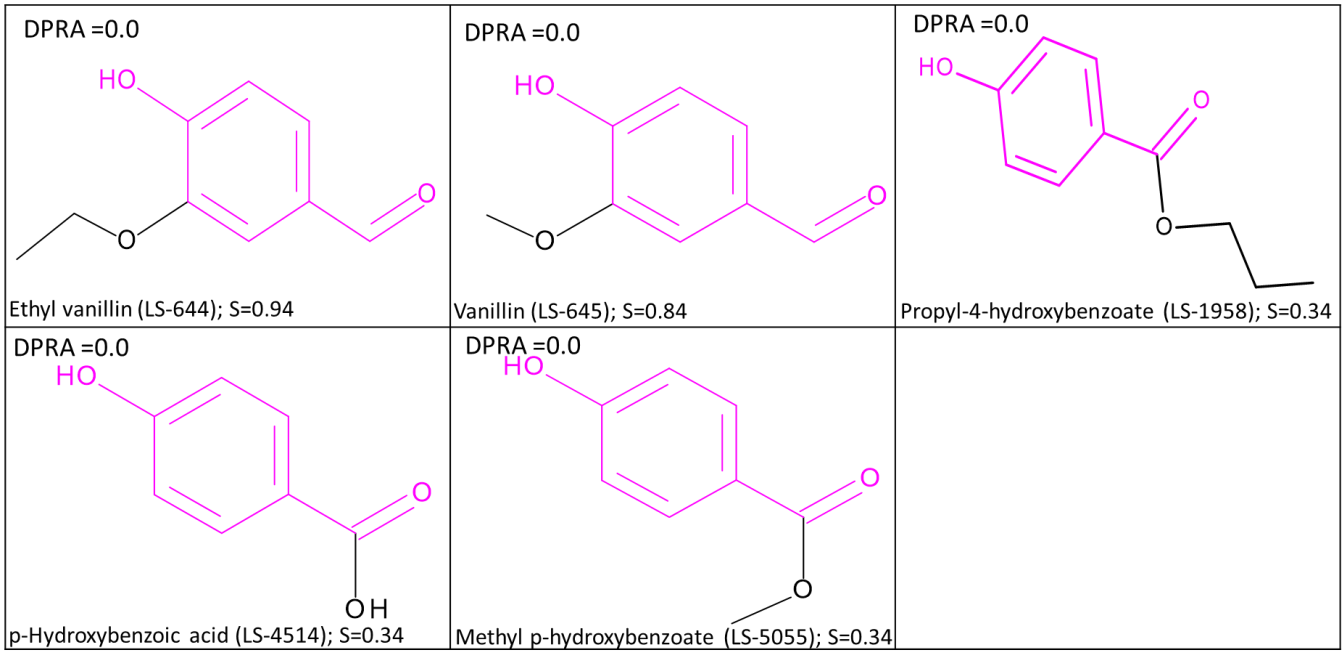
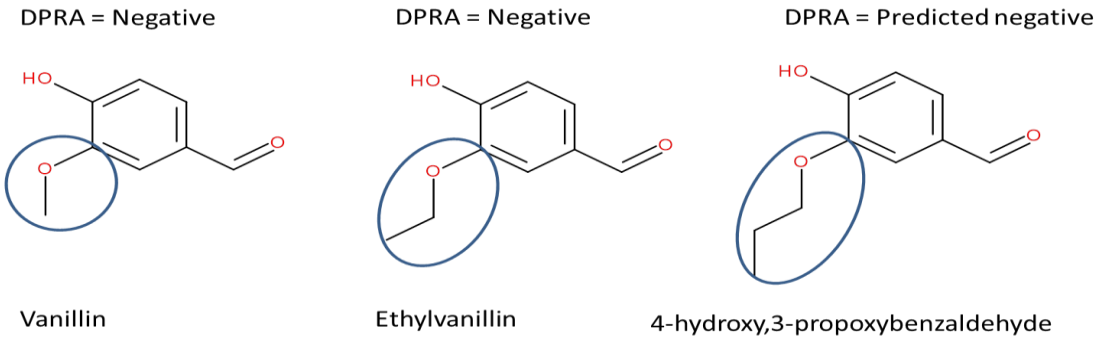


Figure 9. Examples that map to features identified by the DPRA model. DPRA =1.0 refers to compounds with a positive response in the experimental test, while DPRA =0.0 refers to compounds with a negative response in the experimental test. S is the similarity score with the query compound.

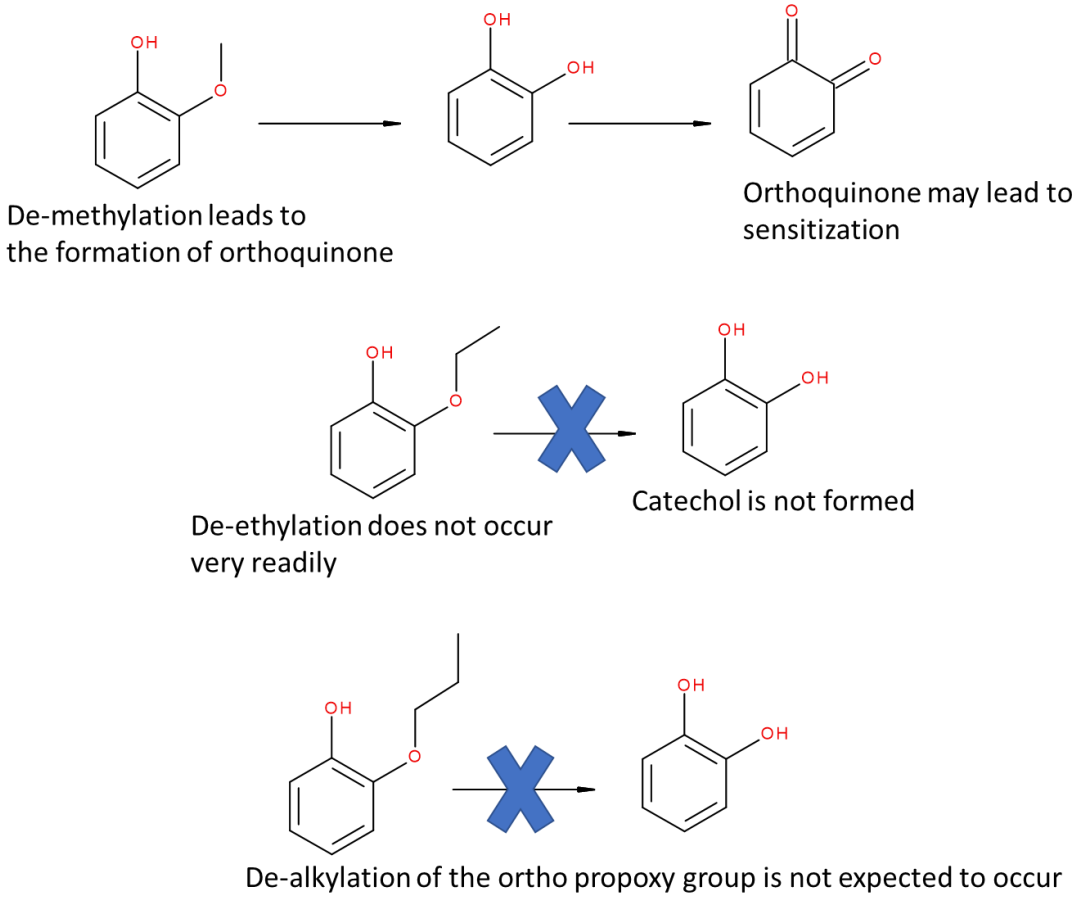
The test structure and two training set examples, vanillin (LS-645; CAS# 121-33-5) and ethyl vanillin, (LS-644; CAS# 121-32-4), form a homologous series with increasing chain length at the o-alkyl group, Figure 10. Vanillin and ethyl vanillin were both assessed as having 'minimal reactivity', in cysteine, and lysine peptide depletion assays[19]. Vanillin, however, may be implicated in sensitization through metabolism to a reactive ortho-quinone[42]. The DPRA lacks metabolic capability and may 'miss' reactivity that could be associated with vanillin metabolite. Ethyl vanillin is a closer analog to the test structure and since de-ethylation is expected to occur less readily than de-methylation, metabolism is not expected to occur in the case of ethyl vanillin.[42] While the relevance of vanillin as an analog may be questioned on the basis of metabolism, the argument is not extended to ethyl vanillin , Figure 11. Since it is unlikely that the addition of the methyl group would confer reactivity to ethyl vanillin, the analogs support the 'no or minimal reactivity' conclusion.



615

616 Figure 10. Examination of the close analogs vanillin (LS-645; similarity = 0.84) and ethyl vanillin (LS-644;

617 similarity = 0.94) highlighting the differences in their structure



618

619 Figure 11. Formation of reactive orthoquinone by de-methylation

Two compounds, chloro-p-anisaldehyde (LS-414183; CAS# 4903-09-7) and anisyl alcohol (LS-2359; CAS# 105-13-5), were positive in the DPRA. Reviewing the training set examples that match the structural descriptors and identifying if other moieties are potentially responsible for biological activity is also useful. Chloro-p-anisaldehyde (LS-414183) is negative in the LLNA[20] and could be considered a DPRA false positive (FP) when compared to the LLNA. Anisyl alcohol (LS-2359) is positive in the LLNA; however, it has been postulated that metabolic transformation (sulphation of the benzylic OH to Ar-$CH_2OSO_3^-$, which is an SN2 electrophile) or abiotic transformation are needed to convert this compound to an active sensitizer[43]. Neither of these mechanisms are expected to occur for 4-hydroxy,3-propoxybenzaldehyde. Therefore these mechanisms are not relevant for the test structure. The questionable relevance of LS-414183 (possible FP based on LLNA) and LS-2359 (mechanistic relevance) supports the negative prediction, since any positive contribution to the feature weight by these examples, could be refuted. It is also worth noting that the similarity of LS-2359, and LS-414183 to the test compound was low (≤ 0.35). The negative prediction for 4-hydroxy-3-propoxybenzaldehydeappears valid and the reliability score is increased to an RS3 level.

3.3.3 Events in Keratinocytes

The KeratinoSens™ (v2.0) statistical model has been used to predict the test compound's potential to activate keratinocytes. The model predicted a negative result with a probability value of 0.078. The compound was within the applicability domain of the model. There were 3 features which were identified and there are 14 analogs which share >30% similarity with the test structure. The training set examples were mainly benzaldehyde and aromatic alkoxy derivatives. The test structure is a benzaldehyde derivative that contains the methoxyaryl feature. Figure 11 shows the coverage of 4-hydroxy-3-propoxybenzaldehydeby the model features. An initial assessment indicates that any uncertainty in the negative prediction most likely will result from the methoxyaryl feature.

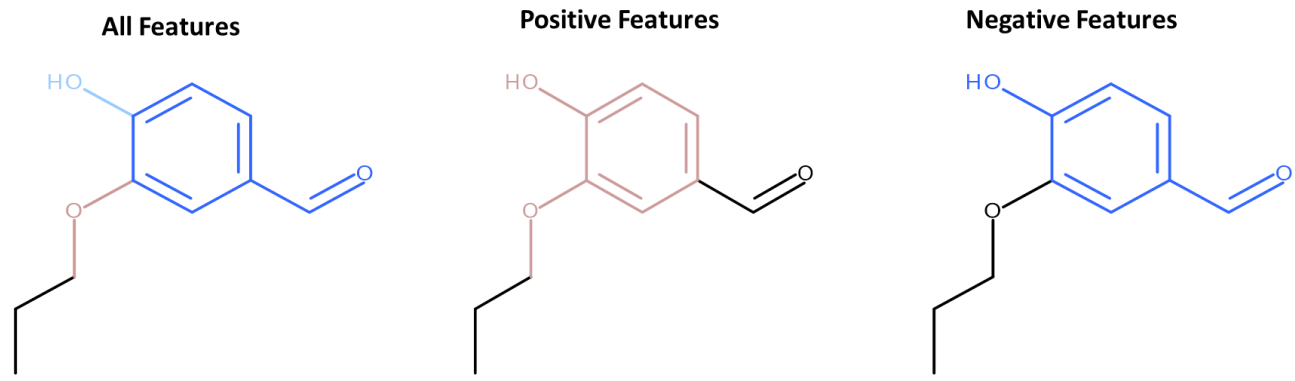


Figure 11. Coverage of 4-hydroxy-3-propoxybenzaldehyde by the KeratinoSens™ model features. Features which contribute to a negative prediction are highlighted in a blue color and those which contribute positively are highlighted in red.

The training set examples that map to the methoxy phenol feature are shown in Figure 12. As discussed previously, structures that contain the methoxyaryl feature could potentially cause sensitization following a metabolic conversion. The positive experimental calls for training set examples LS-2028; CAS# 97-54-1, LS-2674; CAS# 91-10-1, and LS-2898; CAS# 2785-87-7[20,44] reflect this mechanism. LS-645 (vanillin) is negative[20]. This negative result indicates that the aldehyde group may play a role in the lack of a response in the KeratinoSens™ test. Figure 13 shows the examples that map to the benzaldehyde feature. It is worth noting that para-hydroxybenzaldehydes and para-methoxybenzaldehydes are negative in the KeratinoSens™ test. Natsch et al. [45] explains that the p-methoxy and p-hydroxy benzaldehydes have a low propensity to form stable Schiff bases in aqueous solutions compared to unsubstituted benzaldehyde. Ethyl vanillin is, however, not included in the training set but experimental data for this compound was published[20]. The positive result of ethyl vanillin introduces some uncertainty in the assessment. Compared to the LLNA result, this prediction would be considered a false positive result; however, there is no mechanistic rationale for this prediction. In light of the positive result for a close structural analog, a reliability level of RS5 was assigned to the negative prediction.

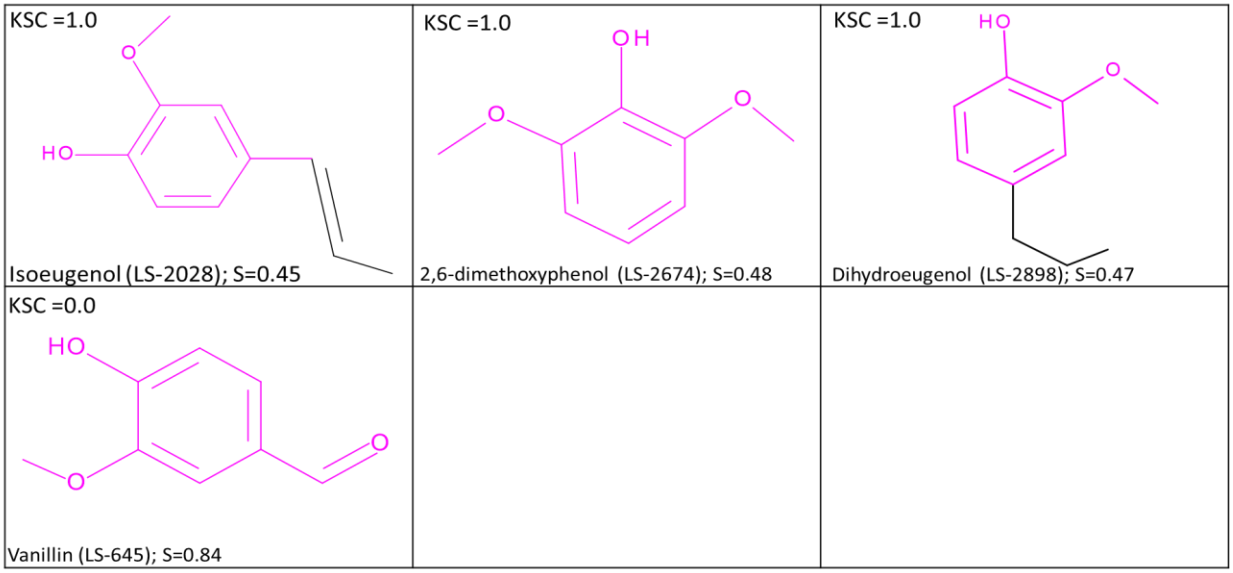


Figure 12. Examples which map to the methoxyaryl feature. KSC =1.0 refers to compounds with a positive response in the experimental test while KSC =0.0 refers to compounds with a negative response in the experimental test. S is the similarity score with the target.

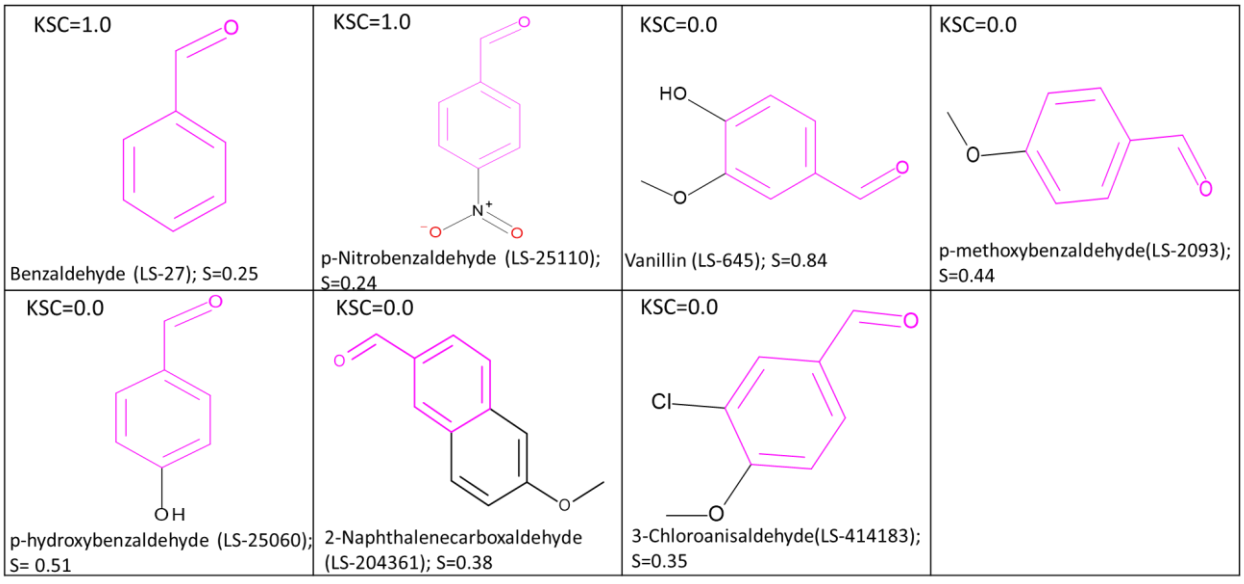


Figure 13. Examples which map to the benzaldehyde feature. KSC =1.0 refers to compounds with a positive response in the KeratinoSens™ test while KSC =0.0 refers to compounds with a negative response in the KeratinoSens™ test. S is the similarity score with the target.

3.3.4 Activation of Dendritic Cells

The statistical model predicting the events in dendritic cells (Human Cell Line Activation Test (h-CLAT) (v2.0)) model returned a negative result and much of the same arguments above could be applied to the review of the predictions; however, the context in which they are applied are slightly different. In this case, the model returns a negative prediction with a predicted probability value of 0.49. This predictive value is close to the predictive threshold (0.5) and as expected for a higher predictive value, the positive features are more apparent in the structure's coverage, compared to other assessments, Figure 14.

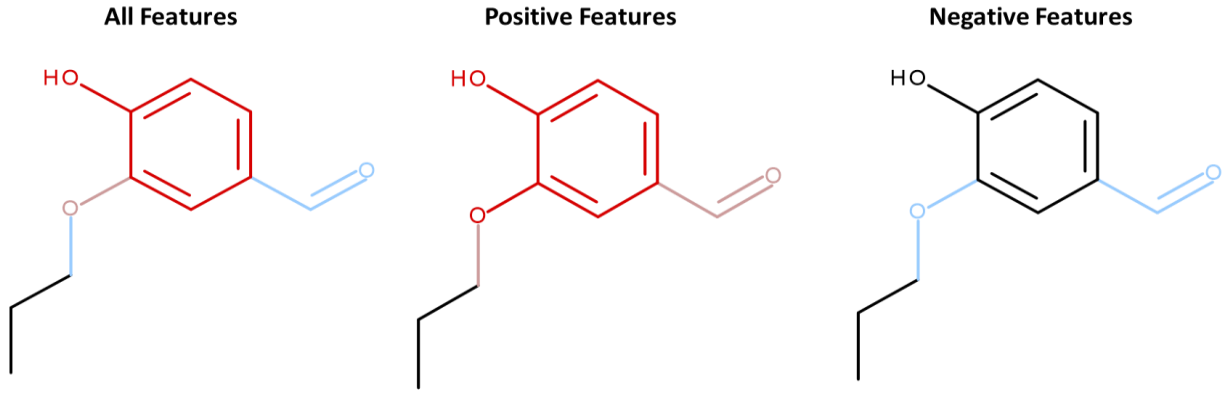


Figure 14. Coverage of 4-hydroxy-3-propoxybenzaldehyde by the dendritic cell activation model features. Features which contribute to a negative prediction are highlighted in a blue color and those which contribute positively are highlighted in red.

The most positively contributing features include the methoxyaryl and di-substituted benzenes. Arguments related to the methoxyaryl feature are similar to those discussed above. The di-substituted benzene feature maps to examples which are pro-haptens such as aminophenol, propyl gallate, dihydroeugenol, and in addition to vanillin and ethyl vanillin. These examples (except ethyl vanillin) are assessed as positive and have some intrinsic potential to metabolize to a reactive quinone, similarly to compounds containing the methoxyaryl feature. While vanillin has a negative assessment in the h-CLAT method[46], it is assessed as positive in the U-SENS™[22]. Further, ethyl vanillin, which is postulated to have a lower sensitization potential than vanillin based on the unfavorable de-ethylation[42], is negative in the h-CLAT and a U-937 test.[20,44,46] The negative features include ether and aryl carbonyl, highlighted in blue in Figure 14. The examples which map to the ether feature are diverse (terminal, aromatic and non-aromatic

ethers are represented); the examples are predominantly negative and contain no obvious reactive features. The aryl carbonyl feature contains three positive examples, the reactivity of which could be explained by moieties other than a single carbonyl group (for example, anhydrides and diketones). The negative examples include carboxylic acids, aromatic esters, and ketones. Given the weight of evidence presented in this case, it is reasonable to consider this negative prediction to be reliable and an RS3 score is assigned.

3.3.5 Endpoint: skin sensitization in vitro

No *in vitro* tests were conducted in this assessment. The *in silico* assessments based on *in vitro* findings agree on a negative result. Two results were assigned a medium confidence level (Covalent interaction with skin proteins, Events in dendritic cells), and the third result (Events in Keratinocytes) was assigned a low confidence. The overall *in vitro* result is considered to be negative with medium confidence based on the two results of medium confidence, Figure 15.

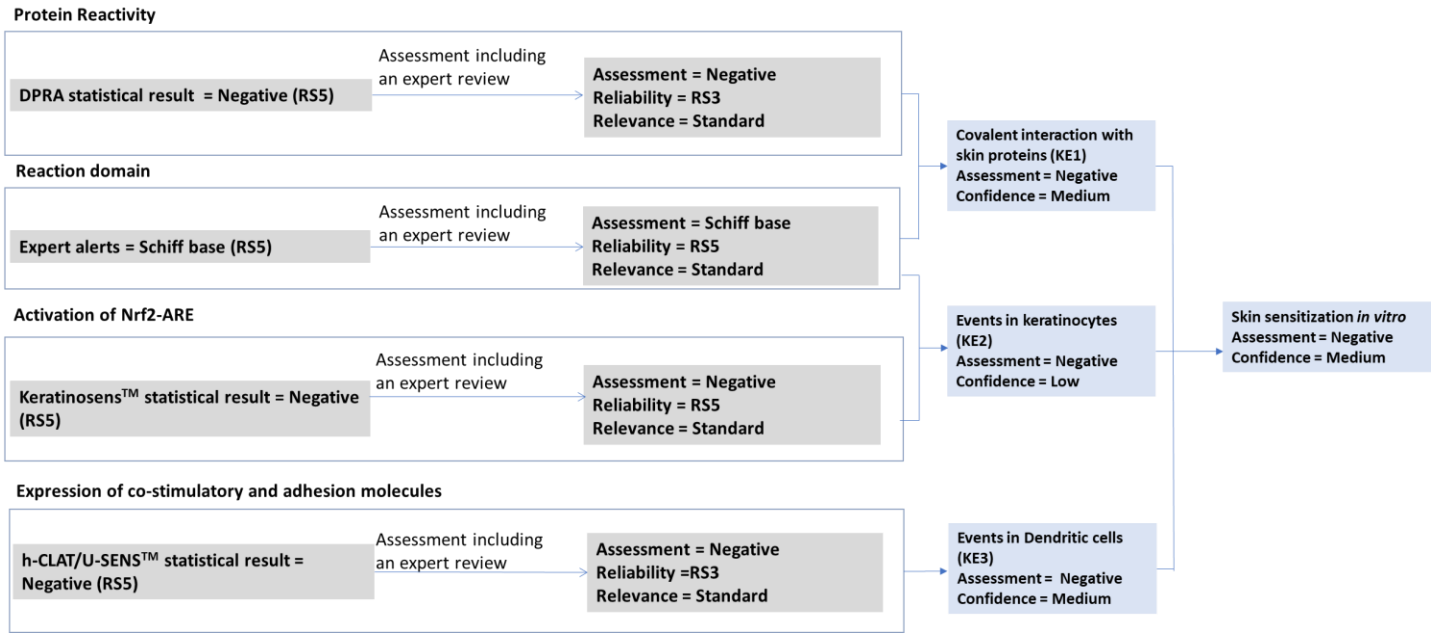


Figure 15. Derivation of the skin sensitization *in vitro* assessment of 4-hydroxy-3-propoxybenzaldehyde given the reliability, relevance, and confidence of the supporting assessments

3.3.6 Events in rodent lymphocytes

No experimental data are available for the assessment of the events in rodent lymphocytes. Expert alerts (Local Lymph Node Assay Expert Alerts (v2.0)) and statistical models (Local Lymph Node, (v2.0)) were used to predict the LLNA responses. No alerts were identified in 4-hydroxy-3-propoxybenzaldehydeand the statistical model predicted a negative result. The compounds were within the applicability domain of the models. For the statistical model, 5 structural features and 30 analogs with similarity scores greater than 0.3 were identified. The feature coverage presents an analysis of the entire test structure and no positive features were identified, Figure 16.

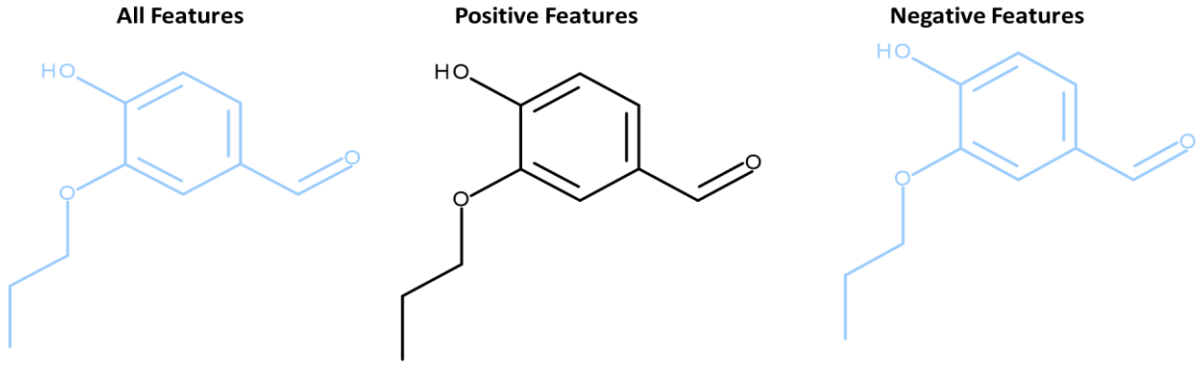


Figure 16. Coverage of 4-hydroxy-3-propoxybenzaldehyde by the LLNA model features. Features which contribute to a negative prediction are highlighted in a blue color and those which contribute positively are highlighted in red. No features expected to contribute to a positive prediction were identified.

The training set examples are predominantly negative and are diverse. Analogs discussed in previous sections (vanillin and ethyl vanillin), in addition to isovanillin are included amongst the training set examples and are assessed as negative in the LLNA[47]. Concomitant predictions supported by an expert review triggered a reliability score of RS3.

A Quantitative Mechanistic Model (QMM) has been developed for LLNA potency of aldehydes and ketones[48]. This QSAR performs well for aliphatic aldehydes and ketones, but substantially overpredicts the potency of most aromatic aldehydes. Apart from a few cases with special features (notably ortho-hydroxybenzaldehydes, but not para-hydroxy), aromatic aldehydes, although predicted by the QSAR to have single figure EC3 values, are weak or non-sensitizing in the LLNA. For example, benzaldehyde is

predicted to have an EC3 value of 4.2% but gives SI values <3 up to 25% (highest concentration tested). However, since the aldehyde is aromatic and has no special features, this is an overestimate of potency. By analogy with benzaldehyde, if it can exhibit an EC3 value, this value is expected to be >25%. A similar calculation could be made for ethyl vanillin. From the $\Sigma\sigma^*$ value of 0.97 and the logP value of 1.74, an EC3 value of 10.5% is calculated from the QSAR. However, since the aldehyde is aromatic and has no special features, this is an overestimate of potency and ethyl vanillin has been assessed as negative in the LLNA[47]. The Events in rodent lymphocytes endpoint is predicted as negative with medium confidence, based on a lack of alerting fragments, and concurring reliable negative statistical results, as shown in Figure 17. However, given the rough estimate of potency from the QMM (EC3 >25%) and the medium level confidence, if any sensitization occurs as a result of exposure to 4-hydroxy,3-propoxybenzaldehyde, it would be expected to be a weak sensitizer.

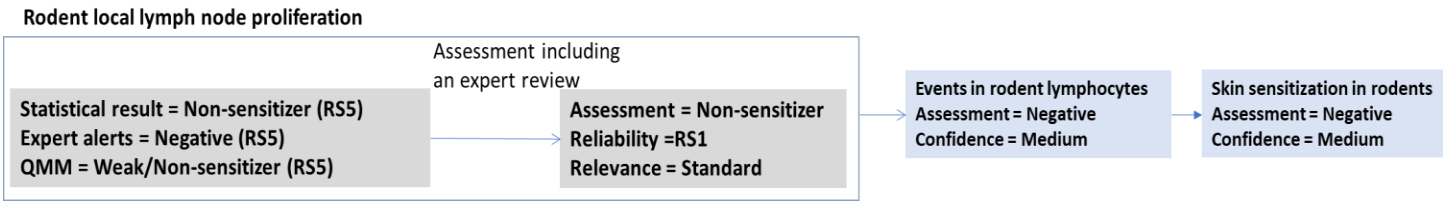


Figure 17. Derivation of the skin sensitization in rodents assessment of 4-hydroxy-3-propoxybenzaldehyde given the reliability, relevance, and confidence of the supporting assessments

3.3.7 Human skin sensitization

A QMM has also been developed for human potency (NOEL values)[49]. Similarly, to the LLNA QMM, this model substantially overpredicts the potency of aromatic aldehydes. For 4 aromatic aldehydes with no observed effect level (NOEL) data (benzaldehyde, cuminaldehyde, piperonal, and p-methoxybenzaldehyde), the NOEL was underpredicted (that is, potency overpredicted) by a factor ranging from 20 to 50[49]. Bearing the above in mind, a rough prediction of the NOEL for 4-hydroxy-3-propoxybenzaldehyde of 127 $\mu g/cm^2$ is calculated. By analogy with other aromatic aldehydes, the true NOEL is expected to be 20-50 times higher. Applying a conservative factor of 20, the NOEL is expected to be ≥2500 $\mu g/cm^2$. Given that the aromatic aldehydes are outside the applicability domain of the QMM[49], it is challenging to assess the reliability and relevance. An RS5 is conservatively assigned, with unknown

relevance. However, this information is useful as it reflects that under the most conservative circumstances,4-hydroxy-3-propoxybenzaldehyde would be expected to be a weak sensitizer, based on the predicted NOEL and according to the classification scheme presented by Api et al.[50], Figure 18.

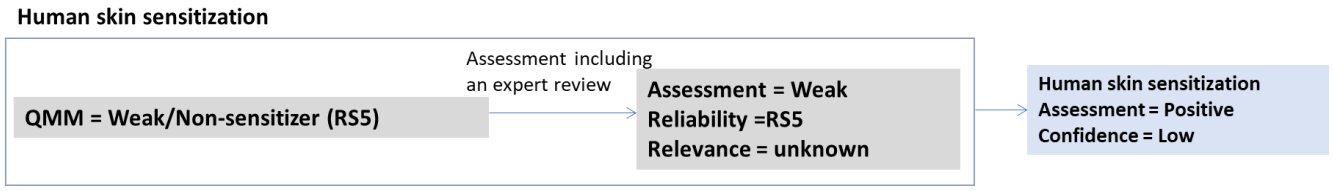
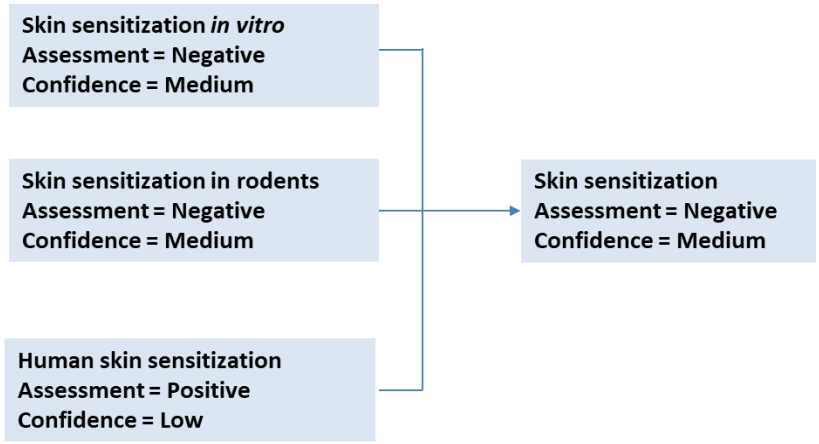


Figure 18. Derivation of the human skin sensitization assessment of 4-hydroxy-3-propoxybenzaldehyde given the reliability and relevance of the supporting assessments

3.3.8 Endpoint or overall assessment: skin sensitization in humans

The overall assessment of the endpoint takes all components of the framework into consideration. The confidence score of each non-apical endpoint incorporates an evaluation of the reliability and relevance of the information presented. Non-apical endpoints with higher confidence scores (more reliable and/or relevant information) have greater weights in the final assessment, particularly when the information adequately covers the pathways leading to the adverse outcome. The *in silico* prediction of LLNA and *in vitro* endpoints are aligned on a negative assessment with a medium confidence level. The uncertainties in the assessment around potential metabolism to a reactive species could be rationalized in different systems. The overall medium confidence adequately reflects the degree of certainty in the conclusion of a negative skin sensitization in humans and the lack of experimental data, Figure 19.



773 Figure 19. Derivation of the overall skin sensitization assessment of 4-hydroxy-3-propoxybenzaldehyde
774 given the confidence in the supporting assessments

775 4. Discussion

776 The above case studies demonstrate how the concepts of reliability, relevance, and coverage could be
777 applied to evaluate multiple lines of evidence. As toxicology moves towards new approach
778 methodologies, using standardized language becomes an important part of evaluating, integrating, and
779 communicating the confidence in new methods and their results. Here, we demonstrate that the concepts
780 of reliability, relevance, and coverage could be applied to *in silico* methods combined with experimental
781 data and across multiple endpoints to derive an overall assessment and confidence. Such weight of
782 evidence approaches were previously described[51,52]. In fact, an evaluation of reliability, relevance, and
783 coverage are fundamental to the application of IATAs. One of the more obscure principles, however, has
784 been the evaluation of *in silico* results within these contexts. The use of controlled vocabulary, along with
785 transparent tools, allow the assessor to interrogate the predictions and allows for application of the
786 principles discussed. The overall impact is the mitigation of black box concerns, effective communication,
787 and reproducibility of *in silico* and experimental results combined.

788 The *in vitro* and *in chemico* analysis of phthalic anhydride presents a case in which experimental systems
789 indicate mixed results with a majority consensus negative call. Depending on the defined approach used,
790 and in the absence of a review of reliability and relevance, varying final assessments may be made.

791 However, once the compound level relevance of the systems for the analysis of phthalic anhydride are
792 examined, the uncertainties around the discordant results become communicable. Further, the added
793 advantage of a reliable and relevant statistical model result predicting the expression of co-stimulatory
794 adhesion molecules, which is concordant with the protein reactivity assessment supports the final
795 assessment of a positive call. The final assessment is made considering all lines of evidence and at this
796 point it is important to communicate the confidence in the result and the principles involved in deriving
797 that confidence. Within the IATA employed[3] and evaluating other lines of evidence including reactivity
798 domains, aspects of reliability, and relevance at the various discussion levels and utilizing structure activity
799 relationships from known examples, the positive assessment can be rationalized.

The second case study of 4-hydroxy-3-propoxybenzaldehyde is an example in which the *in silico* analysis predictions are predominantly used to derive an assessment. In this case, the potential metabolism within *in chemico*, and *in vitro* systems are addressed. Experimental results from close structural analogs, vanillin and ethyl vanillin offered some degree of reliability and supported relevance to the negative prediction. Vanillin has a low incidence of sensitization (Diagnostic Patch Testing (DPT) data % incidence ranging from 0 – 0.19%)[53,54] despite its wide use and has been classified as a category 5 sensitizer (very weak; not GHS classified) by Basketter et al. (2014)[55]. Data are lacking on the human sensitization potential of ethyl vanillin; however, the LLNA assesses both vanillin and ethyl vanillin as non-sensitizers. While in this case, analysis of these analogs along with other lines of evidence lead to a medium level confidence in the assessment, such relevant analogs may not be available for a test compound for which metabolic or abiotic transformation is suspected. In such cases, the relevance of the test system for the particular test compound will bring uncertainty to the overall assessment, and a low confidence rating may be appropriate.

5. Conclusions

As we continue to explore the role of *in silico* models in regulatory settings, it is important to discuss how we could consistently and transparently review model predictions and combine different lines of evidence to derive an overall assessment. In experimental systems, the concept of reliability and relevance are well defined and the degree of uncertainty in an experimental system is reviewed by analyzing various experimental parameters and through a mechanistic understanding of how different chemistries interact with the biological systems. *In silico* methods are built on computer-derived relationships between the chemical structure and biological systems, which should be explored in a manner that allows an assessment of reliability and relevance. Such analyses are important to better understand how much emphasis could be placed on an *in silico* model's result in a weight of evidence scenario. The assessment framework originally presented by Myatt et al.[2] and exemplified here, should find use across various toxicological endpoints.

Acknowledgements

Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number R44ES026909. The content is solely the

828 responsibility of the authors and does not necessarily represent the official views of the National Institutes
829 of Health.

830

831 ## References

832 1. Hardy A, Benford D, Halldorsson T, et al. Guidance on the use of the weight of evidence approach
833 in scientific assessments. *EFSA J*. 2017. doi:10.2903/j.efsa.2017.4971

834 2. Myatt GJGJ, Ahlberg E, Akahori Y, et al. In silico toxicology protocols. *Regul Toxicol Pharmacol*.
835 2018;96:1-17. doi:10.1016/j.yrtph.2018.04.014

836 3. Johnson C, Ahlberg E, Anger LT, et al. Skin sensitization in silico protocol. *Regul Toxicol*
837 *Pharmacol*. 2020;116:104688. doi:https://doi.org/10.1016/j.yrtph.2020.104688

838 4. Hasselgren C, Ahlberg E, Akahori Y, et al. Genetic toxicology in silico protocol. *Regul Toxicol*
839 *Pharmacol*. 2019;107. doi:10.1016/j.yrtph.2019.104403

840 5. OECD. Guidance document on the validation and international acceptance of new or updated
841 test methods for hazard assessment. In: *Series on Testing and Assessment*. ; 2005.
842 doi:ENV/JM/MONO(2005)14

843 6. OECD. *Guidance Document on Good In Vitro Method Practices (GIVIMP), OECD Series on Testing*
844 *and Assessment*.; 2018.

845 7. Myatt GJ et al. Increasing the acceptance of in silico toxicology through development of protocols
846 and position papers. *J Comput Toxicol*. 2021;To be subm.

847 8. Myatt GJ, Ahlberg E, Akahori Y, et al. In silico toxicology protocols. *Regul Toxicol Pharmacol*.
848 2018;96. doi:10.1016/j.yrtph.2018.04.014

849 9. OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship
850 [(Q)Sar] Models. *Transport*. 2007. doi:10.1787/9789264085442-en

851 10. OECD. Guideline No. 497 Guideline on Defined Approaches for Skin Sensitisation Section 4 Health
852 effects. *OECD Guidel Test Chem Sect 4, OECD Publ Paris*. 2021.

https://doi.org/10.1787/b92879a4-en.

11. Riniker S, Landrum GA. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J Cheminform*. 2013. doi:10.1186/1758-2946-5-26

12. Gobbi A, Giannetti AM, Chen H, Lee ML. Atom-Atom-Path similarity and Sphere Exclusion clustering: Tools for prioritizing fragment hits. *J Cheminform*. 2015. doi:10.1186/s13321-015-0056-8

13. Smith DH, Carhart RE, Venkataraghavan R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J Chem Inf Comput Sci*. 1985. doi:10.1021/ci00046a002

14. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010. doi:10.1021/ci100050t

15. Rogers DJ, Tanimoto TT. A computer program for classifying plants. *Science (80-)*. 1960. doi:10.1126/science.132.3434.1115

16. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945. doi:10.2307/1932409

17. *Guidance on Grouping of Chemicals*. OECD; 2014. doi:10.1787/9789264085831-en

18. United States Environmental Protection Agency. OPPT Chemical Fact Sheets (Phthalic Anhydride) Fact Sheet: Support Document (CAS No. 85-44-9). *OPPT Chem Fact Sheets*. 1994.

19. Gerberick GF, Vassallo JD, Foertsch LM, Price BB, Chaney JG, Lepoittevin JP. Quantification of chemical peptide reactivity for screening contact allergens: A classification tree model approach. *Toxicol Sci*. 2007. doi:10.1093/toxsci/kfm064

20. Natsch A, Ryan CA, Foertsch L, et al. A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. *J Appl Toxicol*. 2013. doi:10.1002/jat.2868

21. OECD. Test No. 442C: In Chemico Skin Sensitisation: Assays addressing the Adverse Outcome Pathway key event on covalent binding to proteins, OECD Guidelines for the Testing of

Chemicals, Section 4. *OECD Publ Paris*. June 2019. doi:10.1787/9789264229709-en

22. Piroird C, Ovigne J-M, Rousset F, et al. The Myeloid U937 Skin Sensitization Test (U-SENS) addresses the activation of dendritic cell event in the adverse outcome pathway for skin sensitization. *Toxicol Vitr*. 2015;29(5):901-916. doi:10.1016/j.tiv.2015.03.009

23. OECD. Test No. 442D: In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method, OECD Guidelines for the Testing of Chemicals, Section 4. *OECD Publ Paris*. June 2018. doi:10.1787/9789264229822-en

24. Aptula AO, Roberts DW. Mechanistic Applicability Domains for Nonanimal-Based Prediction of Toxicological End Points: General Principles and Application to Reactive Toxicity. *Chem Res Toxicol*. 2006;19(8):1097-1105. doi:10.1021/tx0601004

25. Urbisch D, Mehling A, Guth K, et al. Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul Toxicol Pharmacol*. 2015;71(2):337-351. doi:10.1016/j.yrtph.2014.12.008

26. Takenouchi O, Miyazawa M, Saito K, Ashikaga T, Sakaguchi H. Predictive performance of the human cell line activation test (h-CLAT) for lipophilic chemicals with high octanol-water partition coefficients. *J Toxicol Sci*. 2013. doi:10.2131/jts.38.599

27. OECD. Test No. 442E: In Vitro Skin Sensitisation: In Vitro Skin Sensitisation assays addressing the Key Event on activation of dendritic cells on the Adverse Outcome Pathway for Skin Sensitisation, OECD Guidelines for the Testing of Chemicals, Section 4. *OECD Publ Paris*. June 2018. doi:10.1787/9789264264359-en

28. Narita K, Ishii Y, Vo PTH, et al. Improvement of human cell line activation test (h-CLAT) using short-time exposure methods for prevention of false-negative results. *J Toxicol Sci*. 2018. doi:10.2131/jts.43.229

29. Casati S, Aschberger K, Asturiol D, et al. Ability of non-animal methods for skin sensitisation to detect pre- and pro-haptens: Report and recommendations of an EURL ECVAM expert meeting. *EUR 27752 EN*. 2016. doi:10.2788/01803.

30. Dearman RJ, Warbrick EV, Humphreys IR, Kimber I. Characterization in mice of the immunological properties of five allergenic acid anhydrides. *J Appl Toxicol*. 2000. doi:10.1002/(SICI)1099-1263(200005/06)20:3<221::AID-JAT651>3.3.CO;2-R

31. Kimber I, Basketter DA, Butler M, et al. Classification of contact allergens according to potency: Proposals. *Food Chem Toxicol*. 2003. doi:10.1016/S0278-6915(03)00223-0

32. OECD. Test No. 429: Skin Sensitisation: Local Lymph Node Assay, OECD Guidelines for the Testing of Chemicals, Section 4. *OECD Publ Paris*. July 2010. doi:10.1787/9789264071100-en

33. Boverhof DR, Gollapudi BB, Hotchkiss JA, Osterloh-Quiroz M, Woolhiser MR. Evaluation of a toxicogenomic approach to the local lymph node assay (LLNA). *Toxicol Sci*. 2009. doi:10.1093/toxsci/kfn247

34. Estrada E, Patlewicz G, Chamberlain M, Basketter D, Larbey S. Computer-Aided Knowledge Generation for Understanding Skin Sensitization Mechanisms: The TOPS-MODE Approach. *Chem Res Toxicol*. 2003. doi:10.1021/tx034093k

35. MAGNUSSON B, KLIGMAN AM. The identification of contact allergens by animal assay. The guinea pig maximization test. *J Invest Dermatol*. 1969;52(3):268-276. doi:10.1038/jid.1969.42

36. Basketter DA, Scholes EW. Comparison of the local lymph node assay with the guinea-pig maximization test for the detection of a range of contact allergens. *Food Chem Toxicol*. 1992. doi:10.1016/0278-6915(92)90138-B

37. Cronin MTD, Basketter DA. Multivariate Qsar Analysis of a Skin Sensitization Database. *SAR QSAR Environ Res*. 1994;2(3):159-179. doi:10.1080/10629369408029901

38. Dearman RJ, Basketter DA, Kimber I. Inter-relationships between different classes of chemical allergens. *J Appl Toxicol*. 2013. doi:10.1002/jat.1758

39. ICCVAM. ICCVAM Test Method Evaluation Report on the Murine Local Lymph Node Assay: DA A Nonradioactive Alternative Test Method to Assess the Allergic Contact Dermatitis Potential of Chemicals and Products. *NIH Publ Number 10-7551 Res Triangle Park NCNational Inst Environ*

928 *Heal Sci*. 2010.

929 40. Nassif AS, Le Coz CJ, Collet É. A rare nail polish allergen: Phthalic anhydride, trimellitic anhydride
930 and glycols copolymer. *Contact Dermatitis*. 2007. doi:10.1111/j.1600-0536.2007.01034.x

931 41. Gach JE, Stone NM, Finch TM. A series of four cases of allergic contact dermatitis to phthalic
932 anhydride/trimellitic anhydride/glycols copolymer in nail varnish. *Contact Dermatitis*. 2005.
933 doi:10.1111/j.0105-1873.2005.00456h.x

934 42. Patlewicz G, Basketter DA, Smith CK, Hotchkiss SAM, Roberts DW. Skin-sensitization structure-
935 activity relationships for aldehydes. *Contact Dermatitis*. 2001. doi:10.1034/j.1600-
936 0536.2001.044006331.x

937 43. Nishijo T, Miyazawa M, Saito K, Otsubo Y, Mizumachi H, Sakaguchi H. Sensitivity of
938 keratinosens™ and h-CLAT for detecting minute amounts of sensitizers to evaluate botanical
939 extract. *J Toxicol Sci*. 2019. doi:10.2131/jts.44.13

940 44. Asturiol D, Casati S, Worth A. Consensus of classification trees for skin sensitisation hazard
941 prediction. *Toxicol Vitr*. 2016;36:197-209. doi:https://doi.org/10.1016/j.tiv.2016.07.014

942 45. Natsch A, Gfeller H, Haupt T, Brunner G. Chemical Reactivity and Skin Sensitization Potential for
943 Benzaldehydes: Can Schiff Base Formation Explain Everything? *Chem Res Toxicol*.
944 2012;25(10):2203-2215. doi:10.1021/tx300278t

945 46. Nukada Y, Ashikaga T, Miyazawa M, et al. Prediction of skin sensitization potency of chemicals by
946 human Cell Line Activation Test (h-CLAT) and an attempt at classifying skin sensitization potency.
947 *Toxicol Vitr*. 2012. doi:10.1016/j.tiv.2012.07.001

948 47. ICCVAM. ICCVAM Evaluations of the Murine Local Lymph Node Assay (LLNA), NICEATM LLNA
949 database. 2013. https://ntp.niehs.nih.gov/whatwestudy/niceatm/test-method-
950 evaluations/immunotoxicity/llna/index.html.

951 48. Roberts DW, Aptula AO, Patlewicz G. Mechanistic Applicability Domains for Non-Animal Based
952 Prediction of Toxicological Endpoints. QSAR Analysis of the Schiff Base Applicability Domain for

Skin Sensitization. *Chem Res Toxicol*. 2006;19(9):1228-1233. doi:10.1021/tx060102o

49. Roberts DW, Schultz TW, Api AM. Skin Sensitization QMM for HRIPT NOEL Data: Aldehyde Schiff-Base Domain. *Chem Res Toxicol*. 2017. doi:10.1021/acs.chemrestox.7b00050

50. Api AM, Parakhia R, O'Brien D, Basketter DA. Fragrances Categorized According to Relative Human Skin Sensitization Potency. *Dermat contact, atopic, Occup drug*. 2017;28(5):299-307. doi:10.1097/DER.0000000000000304

51. OECD. Overview of Concepts and Available Guidance related to Integrated Approaches to Testing and Assessment (IATA), Series on Testing and Assesment No. 329. *Environ Heal Safety, Environ Dir OECD*. 2020.

52. OECD. *Guiding Principles an Key Elements For Establishing A Weight of Evidence for Chemical Assessment No. 311*.; 2019.

53. Uter W, Geier J, Frosch P, Schnuch A. Contact allergy to fragrances: Current patch test results (2005-2008) from the Information Network of Departments of Dermatology. *Contact Dermatitis*. 2010. doi:10.1111/j.1600-0536.2010.01759.x

54. Hausen BM. Contact allergy to balsam of Peru. II. Patch test results in 102 patients with selected balsam of Peru constituents. *Am J Contact Dermat*. 2001. doi:10.1053/ajcd.2001.19314

55. Basketter DA, Alépée N, Ashikaga T, et al. Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis*. 2014. doi:10.1097/DER.0000000000000003