Journal Pre-proof

Skin sensitization *in silico* protocol

Candice Johnson, Ernst Ahlberg, Lennart T. Anger, Lisa Beilke, Romualdo Benigni, Joel Bercu, Sol Bobst, David Bower, Alessandro Brigo, Sarah Campbell, Mark T.D. Cronin, Ian Crooks, Kevin P. Cross, Tatyana Doktorova, Thomas Exner, David Faulkner, Ian M. Fearon, Markus Fehr, Shayne C. Gad, Véronique Gervais, Amanda Giddings, Susanne Glowienke, Barry Hardy, Catrin Hasselgren, Jedd Hillegass, Robert Jolly, Eckart Krupp, Liat Lomnitski, Jason Magby, Jordi Mestres, Lawrence Milchak, Scott Miller, Wolfgang Muster, Louise Neilson, Rahul Parakhia, Alexis Parenty, Patricia Parris, Alexandre Paulino, Ana Theresa Paulino, David W. Roberts, Harald Schlecker, Reinhard Stidl, Diana Suarez-Rodrigez, David T. Szabo, Raymond R. Tice, Daniel Urbisch, Anna Vuorinen, Brian Wall, Thibaud Weiler, Angela T. White, Jessica Whritenour, Joerg Wichard, David Woolley, Craig Zwickl, Glenn J. Myatt

PII: S0273-2300(20)30114-8

DOI: https://doi.org/10.1016/j.yrtph.2020.104688

Reference: YRTPH 104688

To appear in: *Regulatory Toxicology and Pharmacology*

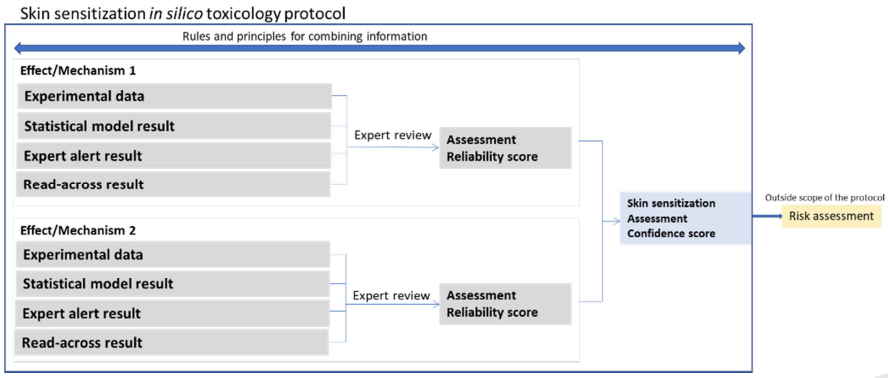
Received Date: 24 February 2020

Revised Date: 18 May 2020

Accepted Date: 21 May 2020

Please cite this article as: Johnson, C., Ahlberg, E., Anger, L.T., Beilke, L., Benigni, R., Bercu, J., Bobst, S., Bower, D., Brigo, A., Campbell, S., Cronin, M.T.D., Crooks, I., Cross, K.P., Doktorova, T., Exner, T., Faulkner, D., Fearon, I.M., Fehr, M., Gad, S.C., Gervais, Vé., Giddings, A., Glowienke, S., Hardy, B., Hasselgren, C., Hillegass, J., Jolly, R., Krupp, E., Lomnitski, L., Magby, J., Mestres, J., Milchak, L., Miller, S., Muster, W., Neilson, L., Parakhia, R., Parenty, A., Parris, P., Paulino, A., Paulino, A.T., Roberts, D.W., Schlecker, H., Stidl, R., Suarez-Rodrigez, D., Szabo, D.T., Tice, R.R., Urbisch, D., Vuorinen, A., Wall, B., Weiler, T., White, A.T., Whritenour, J., Wichard, J., Woolley, D., Zwickl, C., Myatt, G.J., Skin sensitization *in silico* protocol, *Regulatory Toxicology and Pharmacology* (2020), doi: https://doi.org/10.1016/j.yrtph.2020.104688.

Graphical abstract



Skin sensitization *in silico* protocol

Candice Johnson[a*], Ernst Ahlberg[b], Lennart T. Anger[c], Lisa Beilke[d], Romualdo Benigni[e], Joel Bercu[f], Sol Bobst[g], David Bower[a], Alessandro Brigo[h], Sarah Campbell[i], Mark T.D. Cronin[j], Ian Crooks[k], Kevin P. Cross[a], Tatyana Doktorova[l], Thomas Exner[l], David Faulkner[m], Ian M. Fearon[n], Markus Fehr[o], Shayne C Gad[p], Véronique Gervais[q], Amanda Giddings[r], Susanne Glowienke[s], Barry Hardy[l], Catrin Hasselgren[c], Jedd Hillegass[t], Robert Jolly[u], Eckart Krupp[v], Liat Lomnitski[w], Jason Magby[x], Jordi Mestres[y], Lawrence Milchak[z], Scott Miller[a], Wolfgang Muster[h], Louise Neilson[aa], Rahul Parakhia[bb], Alexis Parenty[s], Patricia Parris[cc], Alexandre Paulino[dd], Ana Theresa Paulino[dd], David W. Roberts[j], Harald Schlecker[ee] , Reinhard Stidl[ff], Diana Suarez-Rodrigez[gg], David T. Szabo[hh], Raymond R. Tice[ii], Daniel Urbisch[jj], Anna Vuorinen[o], Brian Wall[x], Thibaud Weiler[q], Angela T. White[r], Jessica Whritenour[kk], Joerg Wichard[ee], David Woolley[ll], Craig Zwickl[mm], Glenn J. Myatt[a]

a) Leadscope, Inc. 1393 Dublin Rd, Columbus, OH 43215, USA
b) Bioinformatics Department, The University of Uppsala, 752 36 Uppsala, Sweden
c) Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA
d) Toxicology Solutions Inc., San Diego, CA, USA
e) Alpha-PreTox, via G.Pascoli 1, 00184 Roma, Italy
f) Gilead Sciences, 333 Lakeside Drive, Foster City, CA, USA
g) Toxsci Advisors LLC, 2016 Main Suite 1901 Houston TX, USA
h) Roche Pharmaceutical Research & Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland
i) Nelson Laboratories, LLC, 6280 South Redwood Road, Salt Lake City, UT 84123
j) School of Pharmacy and Biomolecular Sciences ,Liverpool John Moores University, Liverpool, L3 3AF, UK
k) British American Tobacco, Research and Development, Regents Park Road, Southampton, Hampshire SO15 8TL, UK
l) Edelweiss Connect GmbH, Technology Park Basel, Hochbergerstrasse 60C, CH-4057 Basel / Basel-Stadt, Switzerland
m) Chemical Sciences Division, Lawrence Berkeley National Lab
n) whatIF? Consulting Ltd. The Crispin Burr Street, Harwell OX11 0DT, U.K.
o) DSM Nutritional Products, Kaiseraugst, Switzerland

p) Gad Consulting Services, 4008 Barrett Drive, Suite 201, Raleigh, NC 27609, USA
q) Servier Group, 905 route de Saran, 45520 Gidy, France
r) GlaxoSmithKline, Park Road, Ware, Hertfordshire, SG12 0DP, United Kingdom
s) Novartis Pharma AG, Pre-Clinical Safety, Werk Klybeck, CH-4057, Basel, Switzerland
t) Bristol-Myers Squibb, Drug Safety Evaluation, 1 Squibb Dr, New Brunswick, NJ 08903, USA
u) Toxicology Division, Eli Lilly and Company, Indianapolis, IN, USA
v) Sanofi, Corporate HSE, Global Product Steweardship, Industriepark Hoechst, D-65926 Frankfurt am Main, Germany
w) Perrigo Israel Pharmaceuticals Ltd. Shoham Israel
x) Colgate-Palmolive Technology Center, 909 River Road, Piscataway NJ 08855 USA
y) IMIM Hospital del Mar Institute of Medical Research and University Pompeu Fabra, Doctor Aiguader 88, Parc de Recerca Biomèdica, 08003 Barcelona, Spain; and Chemotargets SL, Baldiri Reixac 4, Parc Científic de Barcelona, 08028 Barcelona, Spain
z) 3M Company, St. Paul, MN
aa) Broughton Nicotine Services, Oak Tree House, West Craven Drive, Earby, Lancashire. BB18 6JZ UK
bb) Church & Dwight Co., Inc. 469 North Harrison Street, Princeton, NJ 08543
cc) Pfizer Worldwide Research and Development, Sandwich, UK
dd) ORO AGRI Europe, S.A. (Palmela - Portugal)
ee) Bayer AG, Research & Development, Pharmaceuticals, Industrial Chemicals Toxicology & Genetic Toxicology, 42096 Wuppertal, Germany
ff) Safetree Consulting e.U., Vienna, Austria
gg) FStox consulting LTD, 2 Brooks Road Raunds Wellingborough NN9 6NS
hh) PPG Industries, Pittsburgh, PA 15146, USA
ii) RTice Consulting, Hillsborough, NC 27278, USA
jj) BASF SE , product safety, Carl-Bosch-Strasse 38, 67056 Ludwigshafen am Rhein, Germany
kk) Pfizer Inc., Drug Safety Research and Development, Eastern Point Road, Groton, CT 06340
ll) ForthTox Limited, PO Box 13550, Linlithgow, EH49 7YU, UK
mm) Transendix LLC, 1407 Moores Manor, Indianapolis, IN 46229, USA

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

Glossary of acronyms

Ac	Acylation
ACD	Allergic Contact Dermatitis
ADRA	Amino Acid Derivative Reactivity Assay
AOP	Adverse Outcome Pathway
ARE	Antioxidant/electrophile response element
BT	Buehler test
CD54	Cluster of Differentiation 54, a co-stimulatory adhesion molecule that is expressed in dendritic cells
CD86	Cluster of Differentiation 86, a co-stimulatory adhesion molecule that is expressed in dendritic cells
CV_{70}	Concentration of test chemical yielding a cell viability of 70% in the U-SENS™ method
DA	Defined Approach
DC	Dendritic cells
DIP	Data interpretation procedure
DPRA	Direct Peptide Reactivity Assay
DSA_{05}	Dose per skin area that produced a positive response in 5% of the tested population
$EC_{1.5}$	Lowest concentration inducing a 1.5-fold change in luciferase activity in the assays measuring KE2
EC_{150}	Effective concentrations yielding a relative fluorescence intensity [RFI] of 150% for CD86 in the h-CLAT test
EC_{200}	Effective concentrations yielding a relative fluorescence intensity [RFI] of 200% for CD54 in the h-CLAT test
EC3	Effective concentration of a test chemical that gives a stimulation index with a three-fold increase over the vehicle control in the LLNA
EC_3	Concentration with 3 fold luciferase induction in the KeratinoSens™ test
GARD	Genomic allergen rapid detection
GPMT	Guinea Pig Maximization test
GST	Glutathione S-transferase
HAF	Hazard assessment framework
h-CLAT	Human Cell Line Activation test
HMT	Human Maximization Test
HRIPT	Human Repeat Insult Patch Test
hTCPA	human T cell priming assay
IATA	Integrated approach to testing and assessment
IC	Induction concentration in GPMT
IC_{50}	Concentration for 50% reduction of viability in KeratinoSens™ test
IL-18	Interleukin-18
IL-8	Interleukin-8
IL-8 Luc	Interleukin-8 Reporter Gene Assay
KE	Key Event
KE1	Key event 1: Covalent interaction with skin proteins
KE2	Key event 2: Events in keratinocytes

KE3	Key event 3: Events in dendritic cells
KE4	Key event 4: Events in lymphocytes
Keap1	Kelch-like ECH-associated protein 1
LLNA	Local Lymph Node Assay
LOEL	Lowest observed effect level
Log K_{ow}	n-octanol/water partition coefficient
MA	Michael addition
MHC	Major histo-compatibility complex
MIE	Molecular initiating event
NOEL	No Observed Effect Level
NQ01	NADPH-quinone oxidoreductase 1
Nrf2	Nuclear factor (erythroid-derived 2)-like 2
OECD	Organization for Economic Co-operation and Development
QMM	Quantitative Mechanistic Models
(Q)SAR	(Quantitative) Structure-Activity Relationship
RFI	Relative fluorescence intensity
SB	Schiff base formation
SI	Stimulation index
SLS	Sodium lauryl sulfate
SM	Supplementary material
SN1	Unimolecular nucleophilic substitution
SN2	Bimolecular nucleophilic substitution
SNAr	Nucleophilic aromatic substitution
STS	Sequential Testing Strategy
U-SENS™	U937 cell line activation Test

64

65

66

67

68

69

70

Abstract

The assessment of skin sensitization has evolved over the past few years to include *in vitro* assessments of key events along the adverse outcome pathway and opportunistically capitalize on the strengths of *in silico* methods to support a weight of evidence assessment without conducting a test in animals. While *in silico* methods vary greatly in their purpose and format; there is a need to standardize the underlying principles on which such models are developed and to make transparent the implications for the uncertainty in the overall assessment. In this contribution, the relationship of skin sensitization relevant effects, mechanisms, and endpoints are built into a hazard assessment framework. Based on the relevance of the mechanisms and effects as well as the strengths and limitations of the experimental systems used to identify them, rules and principles are defined for deriving skin sensitization *in silico* assessments. Further, the assignments of reliability and confidence scores that reflect the overall strength of the assessment are discussed. This skin sensitization protocol supports the implementation and acceptance of *in silico* approaches for the prediction of skin sensitization.

Keywords: *In silico*, *in silico* toxicology, computational toxicology, computational toxicology protocols, (Q)SAR, expert alerts, expert review, skin sensitization, defined approach, integrated approaches to testing and assessment (IATA), extractables and leachables.

Contents

1. Introduction

Allergic contact dermatitis (ACD) is a common skin condition that results from the induction of a dermal immunological response after repeated exposure to a skin-sensitizing substance. ACD poses a significant public and occupational health concern, and much effort has been dedicated to the identification and classification of skin sensitizers. Historically, assessors have relied on human (Human repeat insult patch tests (HRIPT) and Human maximization tests (HMT)) or animal testing, the latter commonly using guinea pig (Guinea pig maximization (GPMT) and Buehler tests(BT))(Organisation for Economic Co-operation and Development (OECD), 1992) and mouse models (Local lymph node assay (LLNA))(OECD 2010a) to identify potential skin sensitizers. The guiding principles of the "3Rs" (replacement, reduction, and refinement) as applied to animal research(RUSSELL and BURCH 1959) have influenced the implementation of regulations, such as the 7th amendment of the Cosmetic Directive (Council Directive 76/768/EEC of 1976-07-27; Cosmetics Regulation: REGULATION (EC) No. 1223/2009), European substances legislation No. 1907/2006 (Registration, Evaluation, Authorization and Restriction of Chemicals (REACH)) in the European Union; and Section 4(h) (Reduction of Testing in Vertebrates) of the Toxic Substances Control Act (TSCA) in the United States. These regulations either prohibit the use of animal testing or only allow animal testing if results obtained by alternative methods are not sufficient to assess the sensitizing potential of a chemical. The "3Rs" together with the need for higher throughput and more mechanistically informative methods, continue to drive the development of non-animal methods. In this regard, *in silico*, *in chemico*, and *in vitro* methods in concert play an integral role in the hazard assessment of skin sensitization.

In silico models, along with *in vitro* tests, have been and continue to be developed for predicting the outcome of the four key events (KEs) described in the OECD adverse outcome pathway (AOP) for skin sensitization (OECD 2014). It is generally accepted that the skin sensitizing hazard of a chemical can be effectively assessed through the integration of non-animal approaches (Kleinstreuer et al., 2018; OECD, 2017). However, there may be data gaps that are generated through the exclusion of chemicals that do not meet the physicochemical property requirements for the *in vitro* tests, and *in silico* methods that could be used to fill such gaps may lack transparency as they are sometimes viewed as "black box" tools. There is also no consensus on how to integrate *in vitro* data and/or *in silico* predictions for these events with existing *in vivo* data.

The protocol detailed in this publication outlines a framework in which *in silico* methods could be applied and integrated with existing *in vivo* and *in vitro* experimental data to identify potential skin

sensitizers, and to provide consensus on the development of models and the interpretation of model results. *In silico* methods are likely to play an important role in understanding the hazard and risk associated with chemicals (Myatt et al. 2018). Assessing sensitization is a necessary component of classification and labelling, workers' safety and occupational health (where ~20-30% of compounds may be sensitizers), regulation of cosmetics and other industrial chemicals as well as product discovery. Previous studies have evaluated the potential use of *in silico* tools to predict sensitization hazard or potential (Roberts and Aptula 2014; Roberts, Aptula, and Patlewicz 2006). However, there remains a need for *in silico* guidelines and the definition of principles and procedures that are specific to the prediction of skin sensitization relevant mechanisms. To this end, this skin sensitization protocol has been developed based on the experience of a cross-industry consortium comprising 39 different organizations and represents a consensus of how to use *in silico* methods to predict skin sensitization.

1.1 Hazard Assessment Framework (HAF)

Figure 1 provides a representation of a generic hazard assessment framework. The hazard assessment framework defines the relationship between mechanisms and effects that are relevant for the prediction of skin sensitization. The mechanisms and effects are molecular perturbations and manifestations, respectively, that lead to the adverse outcome and are reflected in the AOP for skin sensitization (Myatt et al. 2018). The mechanisms and effects are assessed based on *in silico* or existing experimental data. Each mechanism/effect assessment is assigned a reliability score which reflects the inherent quality of the assessment (Section 4). The relevance (scientific predictivity) of the effect/mechanism is also assessed. Rules and principles are used to combine the mechanisms/effects to derive an assessment of non-apical endpoints (i.e., endpoint 1 and 2 in figure 1) that are relevant for sensitization. The non-apical endpoint assessment is assigned a confidence score, which is a reflection of the reliability, relevance, and completeness of the assessment. Non-apical endpoints are combined via rules and principles to derive an overall assessment for skin sensitization (the apical endpoint) with an associated confidence score. The framework is designed to derive an assessment for hazard, with risk being outside the scope of the protocol. Figure 2 shows the hazard assessment framework for sensitization and the relationships between the following endpoints:

- Covalent interaction with skin proteins
- Events in keratinocytes
- Events in dendritic cells
- Skin sensitization *in vitro* (defined approach)

- Skin sensitization in rodent lymphocytes
- Skin sensitization in rodents
- Skin sensitization in humans (weight of evidence)

A comprehensive and mechanistic assessment for skin sensitization includes the four KEs described in the AOP as well as available *in vivo* data and other supporting elements (OECD 2014). A mechanistic understanding of the sensitizing process is detailed within the AOP for skin sensitization and becomes necessary in the development of this framework. In order for a chemical to exert a sensitizing effect, a series of well-defined stages/events occur that lead to the development of effector T cells (as opposed to regulatory T cells, which lead to tolerance(OECD 2014). A chemical's ability to induce each KE is critical information that is used in the development of the HAF. Sensitization is acquired through two distinct phases. During the initial induction phase, the immune system is primed through dendritic cell presentation of the sensitizing chemical to naïve T-cells. The induction phase occurs upon first contact with the sensitizer and a physiological response is typically mild or absent. Upon re-exposure to the same sensitizer, the primed immune system is activated and an inflammatory response occurs. This phase is called the elicitation or challenge phase and results in the manifestation of the symptoms associated with ACD: the appearance of rashes, blisters, and welts. A comprehensive assessment of the skin sensitization potential of a chemical includes the four KEs that are described in the induction phase (OECD 2014).

1.1.1 Key Event (KE) 1: Molecular Initiating Event (MIE) – covalent interaction with skin proteins

The MIE for acquiring skin sensitization is the covalent binding of an electrophilic chemical to a nucleophilic protein, typically the thiol group of cysteine or the primary amine group of lysine (Figure 3). The interaction of the sensitizer (hapten) with the protein leads to the formation of a stable hapten-protein conjugate. While a hapten-bound protein may result from direct interaction of the protein with an electrophile, some chemicals require either metabolic (pro-haptens), or abiotic transformation through oxidation (pre-haptens) prior to complexing with dermal proteins. The hapten-protein interaction depends on the number of available nucleophilic target residues, steric considerations (targets on the surface of a protein are more easily accessible than those in folds), and the microenvironment (hydrophilic or hydrophobic)(OECD 2014). The formation of this complex is critical for the activation of the immunological cells that are responsible for sensitization.

1.1.2 Key Event (KE) 2: Events in keratinocytes

It is accepted that interactions with the hapten lead to the modulation of inflammation-related pathways and oxidative stress response pathways in keratinocytes (OECD 2014)(Figure 3).

Nuclear factor (erythroid-derived 2)-like 2 (Nrf2) is a transcription factor that trans-locates into the nucleus of keratinocytes and binds to antioxidant/electrophile response elements (ARE). This in turn, initiates the transcription of genes related to oxidative stress responses, such as NADPH-quinone oxidoreductase 1 (NQ01) and glutathione S-transferase (GST). Nrf2 is repressed and controlled by the Kelch-like ECH-associated protein 1 (Keap1), which facilitates the ubiquitination and degradation of Nrf2. Keap1 is a cysteine (thiol) rich protein which can be modified by electrophiles (haptens) and oxidants. This modification to Keap1 induces conformational changes in the protein that releases bound Nrf2, allowing it to bind AREs and promote the expression of cyto-protective mechanisms (OECD, 2012).

In addition, interaction of the hapten with keratinocytes stimulates the production of pro-inflammatory cytokines such as IL-18 (Natsch 2010). The release of cytokines by keratinocytes (among other factors) plays a role in stimulating the maturation of dendritic cells (Sumpter, Balmert, and Kaplan 2019)

1.1.3 Key Event 3: Events in dendritic cells

Langerhans cells and dermal DCs are responsible for the presentation of the protein-hapten complex to naïve T-cells in the lymph node during the induction phase (Figure 3). Following the uptake of the protein-hapten conjugate, DCs process and present these peptide fragments in the context of major histocompatibity complex (MHC) molecules to naïve T cells. Matured DCs migrate to the dermis and to the lymph node under the influence of cytokines and chemokines that are secreted by keratinocytes and fibroblast in the dermis (OECD 2014; Sumpter et al. 2019). During maturation, cell surface markers, adhesion molecules, cytokines, and chemokines are upregulated. The upregulation of co-stimulatory adhesion molecules (e.g., CD54, CD86) ensures that professional antigen presenting cells develop and initiate an immune response. When there is a lack of co-stimulation, T-cell anergy (a state in which the lymphocytes remain hypo-responsive after encounter with antigen) and a lack of sensitization may result (OECD 2014; Vocanson et al. 2009)

1.1.4 Key Event 4: Events in lymphocytes

Presentation of the fragmented peptide complex within the MHC to naïve T-cells results in their activation. This leads to the differentiation and proliferation of memory T-cells. Memory T-cells migrate to the dermis and also circulate throughout the body. Upon re-exposure to the same hapten, the

memory T-cells are activated (elicitation phase) and the immune response is triggered; the result is the manifestation of ACD, an irreversible immunologic response (OECD 2014).

KE 1-4 can be used to assess the 'skin sensitization *in vitro* endpoint', which in turn can be extrapolated to the 'skin sensitization in humans' endpoint as shown in Figure 2. These *in vitro* endpoints can also be predicted by *in silico* models as outlined in the HAF (Figure 2) and described in Section 2.

The availability of *in vivo* (usually rodent) data is relevant to the overall assessment of 'skin sensitization in humans' and facilitates the development of *in silico* methods to predict the results. KE 4 (lymphocyte activation and proliferation) can be measured with an *in vivo* mouse model and the adverse outcome (e.g., erythema) can be assessed in guinea pigs. The events in lymphocytes (when assessed in mice) and the guinea pig assessments can be combined to provide an overall assessment of 'Skin sensitization in rodents'. Skin irritation may be a confounding factor and so is also considered at this point. An overall assessment of 'skin sensitization in humans' can be determined through the integration of the 'skin sensitization *in vitro*' and 'skin sensitization in rodents' endpoints. Historical human test data may also be available and *in silico* models can be developed to facilitate its prediction. This information also propagates into the 'skin sensitization in humans' endpoint.

The HAF consists of evaluation of KE1-4 via *in vitro* or *in vivo* testing, physio-chemical properties, and human data (Figure 2). The assumption is made that all chemicals are capable of dermal penetration as a conservative measure (Fitzpatrick, Roberts, and Patlewicz 2017). The endpoints in the framework may be informed through available data, *in silico* predictions, or data acquired through conducting a test. The protocol defines general rules and principles for integrating data towards an overall prediction of the adverse outcome in humans. The incorporation of lines of evidence that may not directly relate to sensitization; such as skin irritation, means that the protocol takes the form of an integrated approach to testing and assessment (IATA).

1.2 Integrated approach to testing and assessment (IATA)

Given the definition of an AOP for skin sensitization and the availability of historical data, the endpoint is effectively predicted using an IATA. Limited data for the KEs along the AOP have restricted the development and applicability of *in silico* models to predict these endpoints while *in vitro* testing is mainly used to derive an assessment of the activation of KEs along the AOP pathways. This may change in the future, as more data become available and more robust *in silico* models can be developed. Nonetheless, through an integrated scheme, the overall endpoint of 'skin sensitization in humans' is

assessed as a function of the activity at each KE, with additional evidence from either existing data or *in silico* predictions of *in vivo* responses and metabolic biotransformation. Previous research has focused on developing such schemes and these non-animal integrated strategies are receiving interest from regulatory authorities. The publication of the 'Interim Science Policy: Use of Alternative Approaches for Skin Sensitization as a Replacement for Laboratory Animal Testing' is an example of regulators adopting this more integrated approach(EPA 2018). Additional non-animal assessment strategies are currently being developed and validated, and more approaches may be adopted for regulatory purposes in the future (Kleinstreuer et al. 2018). While several integrated approaches invoke the AOP and integrate the KEs to derive an overall assessment of skin sensitization, it has been argued that failure or ability to sensitize could be explained by (in)sufficient activity in the 'covalent interaction with skin proteins' endpoint, and the evaluation of subsequent KEs is less important (Roberts and Aptula 2008). To this end, the authors believe that a HAF that can facilitate multiple approaches is necessary. The ideal framework should be generic enough to facilitate possible variations in analysis while maintaining a high level of reproducibility and transparency. Rules and principles for combining results for each endpoint are defined in this protocol. These rules will set the foundation for the reproducibility and flexibility of the framework presented here.

1.3 Defined Approaches

Previous approaches have incorporated rules that connect various aspects of the toxicological pathway to skin sensitization. The "2 out of 3" integrated testing strategy approach to skin sensitization hazard identification proposed by BASF uses a data interpretation procedure (DIP) that labels a chemical as a sensitizer or non-sensitizer based on the concordant reactivity of the chemical in two *in vitro* tests for KE1 - KE3 (Urbisch et al. 2015). Several other integrated strategies have been developed to assess either hazard or potency (Section 1 of the supplementary materials and described in detail elsewhere (OECD 2017)). Each approach addresses particular elements of the AOP. At the time of this manuscript, no single approach is viewed as being superior to the others and selected approaches vary based on the availability of computational tools and data.

1. *In silico* methodologies and models

Historically, *in silico* models have focused on the prediction of animal data (particularly the LLNA), and few have considered the rest of the mechanisms established in the AOP. Therefore, it is necessary to examine how *in silico* tools could be developed to model mechanisms related to the KEs described earlier.

Depending on the availability of high-quantity data, different types of *in silico* models can be developed. Table 1 provides a list of data sources. Larger amounts of data, preferably with a strong mechanistic understanding of a specific toxicological process, can support many different types of models. Datasets that cover a broad chemical space can support the development of global Quantitative Structure-Activity Relationship ((Q)SAR) models, provided that the descriptors are relevant and mechanistically-related to the endpoint that is being predicted (Roberts et al. 2007). Where data are sparse, generated with different protocols, or generated through multiple mechanistic pathways (as may be the case in human studies), methods such as expert-alerts or read-across may be more appropriate. [1] Statistical models may also be developed; however, these models are potentially limited by a smaller applicability domain. On the other hand, the mechanistic understanding and classification of chemicals into a mechanistic domain means that local QSAR modeling may be a feasible approach for assessing events related to the sensitizing endpoint. One of the earliest attempts to develop a local mechanism-based QSAR model to predict EC3 concentrations in the LLNA, used the Relative Alkylation Index (RAI, a function of electrophilic reactivity, lipophilicity, and dose)(Roberts et al. 1991; Roberts and Williams 1982). Subsequently, several Quantitative Mechanistic Models (QMM) have been developed with the goal of identifying physicochemical and other descriptors that contribute to a mechanistic understanding of an endpoint of interest (Aptula and Roberts 2006; Roberts and Aptula 2014; Roberts, Aptula, and Patlewicz 2011; Roberts and Natsch 2009). The rest of Section 2 discusses the mechanisms or effects that could be predicted and which types of *in silico* methodologies could facilitate the predictions. On a general note, *in silico* methods typically derive structure activity relationships (SAR) for organic salts by using the structure of the freebase. In cases where a metallic fragment will be removed in the generation of the freebase to derive the SAR form of the structure, the potential hazard posed by the metal should be considered. In the area of skin sensitization, removing nickel fragments may lead to an underestimation of hazard for structures that contain them. To more accurately facilitate predictions in these cases, the metal may be attached to the ligand, or the metal may be kept unattached in the training set. The model builder may also decide to remove the salt structure entirely from the training set; thereby, excluding the metal from the applicability domain of the model.
The following sections describe general considerations for building *in silico* models based on the available chemistry, biology, and testing data. Section 1 of the supplementary material provides a detailed description of the experimental data that are relevant for assessing skin sensitization. Methods

[1] The reader is referred to (Myatt et al., 2018) for a more general discussion on these methods

to assess the reliability of the data as well as *in silico* predictions have been previously described by (Myatt et al. 2018) and are summarized in Section 2 of the supplemental material.

2.1 Covalent interaction with skin proteins, KE1

In silico and or experimental assessments for whether a given compound will participate in covalent interactions with skin proteins are primarily generated based on understanding of metabolism, reaction domain assignment and protein reactivity.

2.1.1 Dermal Metabolism

The allergenic potential of a chemical may be increased or decreased through metabolic pathways or abiotic oxidation; these factors are important for predicting a chemical's potential to induce dermal sensitization. Metabolic detoxification takes place in two phases, which may or may not occur simultaneously. Phase II metabolism appears to be more abundant and active in the skin than in the liver, although Phase I enzymes – though not dominant – are inducible in the skin (Dumont et al. 2015). Given differences in expression profiles between the liver and skin, the potential use of liver metabolic data to predict metabolites in the skin will necessitate strategies for accounting for the differences in the expression of isoenzymes between liver and skin (Madden et al. 2017).[2] One strategy for predicting metabolic activation towards sensitization in dermal tissues is to derive alerts to indicate if a chemical may be a pro-hapten. This approach is currently limited by the size of the databases of pro-haptens and a general lack of skin specific data (although knowledge has been gained through experience over the years). Currently, it appears that the range of structural features that are activated towards sensitization via metabolic pathways is small. Given the absence of skin-specific metabolic data, it is challenging to definitively conclude on the topic. (Natsch and Haupt 2013) investigated the activation of pro-haptens by rat liver S9 fractions in the KeratinoSens^{TM} assay, and identified phenolic and alkoxy groups attached to a benzene ring, some aromatic amines, and conjugated dienes in or in conjunction with six-membered ring as structural features that may require pro-activation to behave as haptens in the assay (Natsch & Haupt, 2013; Bergström, et al., 2006; Bergström, et al., 2006). The features identified do not represent a comprehensive and thoroughly defined list of features that undergo metabolic transformation leading to sensitization.

[2] The supplemental material provides a brief summary of the differences between skin and liver metabolic enzymes with relevance to humans

2.1.2 Reaction Domain

Existing mechanistic information on hapten-protein interactions has been used to construct *in silico* models for predicting sensitization potential based on a compound's structure and known – or predicted – reaction chemistry. The mechanisms for forming protein-hapten complexes involve the interaction between an electrophilic chemical (hapten) and the nucleophilic moiety on a skin protein (generally thiol or primary amine groups). Common mechanisms by which the sensitizer (hapten) may bind to the protein are: Michael addition, acylation, Schiff base formation, unimolecular nucleophilic substitution (SN1), bimolecular nucleophilic substitution (SN2), or nucleophilic aromatic substitution (SNAr). Within each of these mechanistic domains, there are mechanistic alerts and structural alerts. Structural alerts are defined as molecular substructures that can activate the toxicological effect or mechanism (Myatt et al. 2018). Structural alerts that are characterized by a common reaction site are defined as mechanistic alerts (Aptula and Roberts 2006; Enoch, Madden, and Cronin 2008; Roberts et al. 2015). Structural and mechanistic data do not always suggest a toxic effect, however – some structural features, such as steric hinderance, have been found to mitigate toxicity by decreasing the ability of the hapten to covalently bind to proteins – and these features may improve an *in silico* model by providing this additional information.

Classification of mechanistic and structural alerts within mechanistic domains allows for local QSAR modelling within each domain (OECD 2011), provided that one has the relevant quantitative information describing the protein-hapten bond. To this end, the following physical-chemical property descriptors are commonly used to predict interactions between haptens and proteins: Molecular weight (MW), Log P, solubility, rotational bonds, electronic and topological descriptors (e.g., quantum mechanics calculations), or chemical structure-based descriptors (e.g., the presence or absence of different functional groups) (OECD 2011). The factors constituting an acceptable and validated model have been described in previous work (Myatt et al. 2018). However, it must be noted that due to the expert nature of deriving structural alerts based on reaction chemistry, existing *in silico* tools can only incorporate our current knowledge of protein-hapten reaction chemistry (rather than the quantification of a physical or biological process), and that future models could be improved as we increase our mechanistic understanding of these processes. QSARs on the other hand, are not limited by current knowledge of mechanistic processes and the combined use of structural alerts and QSARs may add value to the analysis.

2.1.3 Protein Reactivity

Protein reactivity has been studied using model nucleophiles to assess protein-chemical interaction in *in chemico* assays. While the binding mechanism between the protein and the chemical could be described based on reaction chemistry as discussed in the previous section, any *in silico* tools (either statistical or expert rule-based) developed based on *in chemico* assay data will be limited in their ability to predict sensitization due to pro-activation. To overcome this limitation, predictions based on reaction chemistry, protein reactivity, and dermal metabolism should be considered in concert to generate an overall assessment (described in Section 3.1).

While protein reactivity measurement is feasible across all reaction domains described in section 2.1.2, experimental results show that within the domain of Schiff base formers there is a lower correlation between the *in chemico*-based DPRA model and *in vivo* and human data (Urbisch et al. 2015). While Schiff base formation may be theoretically feasible, the abundance of water within the peptide reactivity testing environment may limit some reactions. As such, peptide reactivity was found to correlate poorly with the potency of aldehydes, as Schiff base formation may be limited under testing conditions in the DPRA (Natsch et al. 2015). Further analysis revealed that more potent Schiff base formers (atranol, chloratranol, and salicylaldehyde) are reactive under physiological conditions (Natsch et al. 2012). However, the LLNA EC3 values of Schiff base formers are well correlated (R^2 = 0.95) with a combination of logP and a reactivity parameter based on substituent constants (Roberts et al. 2006). Differential reactivity within a mechanistic domain is an issue that could become relevant in the development of *in silico* models, and particularly in those that use read-across. Such instances may not be unique to the protein reactivity mechanism but may require examination across all toxicological endpoints.

2.2 Events in keratinocytes, KE2

A comprehensive prediction of keratinocyte activation covers events on several levels of biological organization and includes the expression of biochemical, genomic, and proteomic pathways, and quantifies the release of pro-inflammatory mediators that stimulate dendritic cells in KE3 (OECD 2014). Validated protocols are established for assessing the induction of ARE dependent pathways, and, as such, the development of *in silico* models can be considered for this assessment. However, the breadth of information and data describing other pathways could be informative and may drive the development of *in silico* models to predict additional pathways in the future.

Statistical modelling is feasible; however, the availability of data is a critical factor influencing the success of measures to implement models based on AOP *in vitro* tests. Descriptors relating to the covalent modification of the cysteine-rich Keap1 protein could be used to develop mechanistically-relevant QSAR models. There may be limitations in predicting compounds which preferentially bind hard nucleophiles such as lysine since the *in vitro* tests predicting KE2 rely on the cysteine-dependent modification of Keap1. Therefore, false negative predictions may be more common for compounds that react via acyl transfer, within the domain of Schiff base formers, including short chain aldehydes, and longer chain saturated alkanals. Other electrophiles that prefer hard nucleophiles may also produce false negative predictions (Urbisch et al. 2015). This could be a potential issue in read-across analysis and should be addressed during an expert review.

In silico prediction of KeratinoSens™ and LuSens (*in vitro* test methods for assessing ARE activation in keratinocytes) data yields dichotomous (either positive or negative) test results (OECD 2018b). However, integrated assessments of potency may require continuous data input such as $EC_{1.5}$ (the lowest concentration inducing a 1.5-fold change in luciferase activity), IC_{50} (concentration for 50% reduction of viability) and EC_3 values (concentration with 3 fold luciferase induction) (Natsch et al. 2015).

2.3 Events in dendritic cells, KE3

Dendritic cell activation is similar to keratinocyte activation in that predictions can be made on the levels of protein and gene expression. Methods have been validated for measuring the expression of specific cell surface markers which contribute to T cell activation and proliferation. Published databases may contain data for dendritic cell gene expression of co-stimulatory and adhesion molecules (cell surface markers: CD54 and CD86) and Interleukin-8 (IL-8) (Nukada et al. 2011; Urbisch et al. 2015).

As noted for the KE2 endpoint, care must be taken when integrating testing data from the various *in vitro* assays into KE3 *in silico* models due to differences in the types of data that may be produced by different assays. The continuous data outcomes predicted for these assays, such as the EC_{150} and EC_{200} values from the h-CLAT assay; the CV_{70} and the EC_{150} in the U-SENS™ assay could be used in integrated strategies to predict potency. These and other *in silico* predictions of the Ind-IL8LA (induced interleukin-8 luciferase activity) could be used to support the hazard assessment; however, since a statistically-derived experimental variable (confidence interval) is needed to determine a positive call, a more practical approach may be to dichotomize the assay results and make binary predictions.

Often, it is helpful to build models that use threshold values to convert continuous data into dichotomous (yes or no) values. For any of the *in vitro* or *in chemico* test methods that are used to assess a KE along the AOP, using threshold values, *in silico* predictions could generate dichotomous predictions of KE activity using these *in vitro* or *in chemico* test endpoints.

2.4 Events in human lymphocytes, KE4

The lack of standardized data makes *in silico* predictions of *in vitro* T cell activation and proliferation challenging. A paucity of data for this endpoint is not surprising, however, as the value of predicting this key event remains in question, and the significance of an *in vitro* estimate of KE4 can only be speculated at this time. It is possible that the magnitude of the T cell responses at KE4 may be the key event that allows us to make distinctions between different potency classes *in vitro* (OECD 2014), but the issue has not been settled. Consequently, only the *in vivo* Local Lymph Node Assay has been accepted as a standardized method for assessing this endpoint.

2.5 Events in rodent lymphocytes, KE4

The LLNA is the only standardized in *vivo* method used to measure the proliferation of lymphocytes in response to immune system priming by a test chemical as well as the potency of the chemical as a skin sensitizer. The results of the assay are reported as the concentration of the chemical needed to induce T-cell proliferation by a pre-chosen factor (usually 3, 1.6, or 1.8 times the baseline amount as assessed by the stimulation index (SI))(OECD 2010b, 2010a, 2018a). The LLNA has been used extensively, and it is quite feasible to build *in silico* models using statistical and rule-based methods due to the ready availability of data, although, the majority of such data is proprietary. While the publicly-available LLNA data could facilitate statistical modeling, the model coverage may be reduced for industrial applications. However, the combined use of statistical modeling and structural alert definitions could be a strategy to overcome this limitation.

The irritation potential of a chemical could be a confounding factor in the experimental LLNA, and the issue of irritation translates into *in silico* assessments. Training set examples and analogs under consideration for read-across should be examined for their irritation potential. Studies indicate that non-sensitizing irritants (such as surfactants) could be overestimated by the LLNA, leading to false positive results (Ball et al. 2011; OECD 2010a). While this is certainly the case for sodium lauryl sulfate (SLS), chloroform/methanol, Triton X-100, oxalic acid, methyl salicylate, and nonanoic acid, analysis of chemicals known to be skin irritants has not validated this generalization across the entire class of non-sensitizing irritants(Ball et al. 2011). Most non-sensitizing irritants are negative in the LLNA and those

that are positive may produce borderline results (with few exceptions). For example, the sensitization hazard of SLS is derived from a clear dose-response curve that is indicative of a positive LLNA result; however, when a weight-of-evidence (WoE) approach is used, the interpretation of the LLNA results may be reversed. There is no evidence that SLS is a skin sensitizer in humans despite exposure; albeit limited, it lacks a structural alert for sensitization and is a strong irritant (Basketter et al. 2009). Hence, Basketter et al. 2009 have suggested that for the SI results obtained for SLS in the LLNA (SI_{SLS}), a WoE approach could be developed around the false positive result to implement this approach in a general sense. Using SLS as reference for a test chemical with unknown skin sensitization hazard, irritant potential and SI predictions (SI_{test}); if the $SI_{test} < SI_{SLS}$ and no structural alert exists of sensitization, then the LLNA prediction could be a suspected false positive and confidence in a positive prediction of the "skin sensitization in humans" endpoint is low. The reverse may also be considered: If $SI_{test} > SI_{SLS}$ and an alerting structure exists for sensitization; then the chemical may be suspected to be a true positive (Basketter et al. 2009). The confidence could be adjusted accordingly based on the weight of evidence presented. This sort of analysis would be considered with a low reliability LLNA study which may have been conducted at irritant concentrations. Generally, the LLNA test is preceded by dose finding range studies and minimally irritating to not irritating concentrations are tested.

Some LLNA protocols (LLNA-DA, and LLNA-BrdU-ELISA) use non-radioactive methods to quantify lymphocyte proliferation. Results from these protocols could be combined in training sets that would facilitate binary level predictions; however, varying criteria for predicting a positive call may complicate the prediction of a meaningful continuous SI or EC_X value (where x is 3, 1.6, or 1.8 depending on the LLNA protocol used) from such a dataset and would require a valid strategy for integrating the data. Another relevant issue with LLNA datasets that arises in the curation process is the comparison and combination of SI and EC_3 values for tests conducted in different vehicles. While it seems logical that vehicle effects are normalized in the derivation of the SI and EC_3 values, there are mechanisms that could lead to enhanced bioavailability depending on the choice of vehicle. The rapid evaporation of acetone, for example, may result in volatilization of the test chemical and decreased bioavailability; whereas dimethyl sulfoxide (DMSO) could potentially enhance penetration. Differing results may be obtained between two LLNA tests using different vehicles and this could influence hazard assessment (Hoffmann 2015). In some cases, vehicle effects may lead to the assignment of a chemical to two neighboring potency classes (Anderson, Siegel, and Meade 2011; Basketter, Gerberick, and Kimber 2001; Dumont et al. 2016; Hoffmann 2015). This inherent variability in the LLNA data (not exclusively caused by different vehicles) is translated to *in silico* predictions. When combining multiple data sources, the

most conservative SI and EC_x values could be adopted, unless there is compelling evidence that the vehicle is potentiating or attenuating the effect of the test chemical. A less conservative, but valid, approach is to use the mean, or median values, among other valid approaches (Hoffmann et al. 2018).

2.6 Skin sensitization in rodents

The skin sensitization in rodent endpoint is evaluated through the use of the GPMT and the BT method. Guinea pigs were historically used to assess skin sensitization. Similar to the LLNA, while public data are available, much of the GPMT and BT data are proprietary. The data that exists could facilitate statistical modeling, the derivation of expert alerts, and read-across.

2.7 Skin sensitization in humans

Historical data exist for this endpoint and, based on data quantity, expert-alert derivation and read-across may be preferable to statistical methods. *In silico* predictions could be useful for the prediction of dichotomized results of positive/negative. Potency predictions could be challenging based on data availability. Evidence to support human predictions includes clinical data (DPT) and usage/occupational exposure data (Api et al. 2017). Further, the integration of the 'skin sensitization *in vitro*' and the 'skin sensitization in rodents' endpoints, along with any direct human evidence, are considered together as weight of evidence for the prediction of the 'skin sensitization in humans' endpoint.

3. Endpoint assessment and confidence

The protocol details the integration of data with different reliabilities and relevance. Further, there may be cases in which information that is critical to an assessment is missing. This section outlines the rules/principles that could be applied when deriving an assessment and its associated confidence based on the totality of evidence presented. Figure 4 shows the hazard assessment framework annotated with references to where each of the following sections applies.

3.1 Covalent interaction with skin proteins assessment

Assessment of the 'covalent interaction with skin proteins' endpoint includes consideration of metabolic transformation, reaction chemistry, and DPRA/ADRA predictions. Figure 5 shows how rules could be made around the available information to derive an overall prediction of hazard. If an experimental result is positive for the methods assessing KE1 (DPRA/ADRA), then a positive assessment of the 'covalent interaction with skin proteins' is warranted. However, the reliability of the prediction, as assessed by the scheme presented in Table 6 of the supplementary material and described in (Myatt et al. 2018), varies depending on the quality of the information presented and this has an influence on the

574 confidence score. The quality and reliability of an *in silico* DPRA/ADRA prediction could be assessed
575 according to the expert review criteria described in (Myatt et al. 2018). Additional considerations for
576 both experimental (test article) and *in silico* (training set examples and analogues) results include
577 situations in which DPRA/ADRA could lead to a false positive result due to oxidizing properties of the
578 test chemical, which can lead to peptide dimerization. An expert review could inform on whether or not
579 this is likely and if the assessment and confidence score need adjustment. Assessments of negative
580 DPRA/ADRA results vary based on consideration of the metabolic potential of the chemical together
581 with knowledge of reaction chemistry. In general, when the chemical is expected to be out of the
582 metabolic domain of the DPRA/ADRA then precedence is given to clearly-defined knowledge of reaction
583 chemistry (including mitigating factors, such as sterics) in the overall assessment of the 'covalent
584 interaction with skin proteins' endpoint. If the reaction chemistry indicates a mechanism leading to
585 sensitization; particularly if the mechanism requires pro-activation then the overall assessment of the
586 'covalent interaction with skin proteins' is positive based on reaction chemistry knowledge, but the
587 confidence is medium. If the test article is out of the metabolic domain, negative in DPRA/ADRA and no
588 mechanistic alert could be identified in the structure of the test chemical based on reaction chemistry,
589 then the DPRA/ADRA result is inconclusive as it cannot be said that the overall assessment is either
590 negative or positive. However, if metabolism is not predicted to occur and the chemical is considered
591 within the metabolic domain of the DPRA/ADRA, then the negative result should be given consideration
592 in the overall assessment. A negative DPRA/ADRA prediction (within the DPRA/ADRA metabolic domain)
593 and a positive mechanistic alert lead to a negative overall assessment, with a medium confidence level,
594 given that the DPRA/ADRA result is experimental and the positive mechanistic alert introduces some
595 uncertainty. An expert review would consider whether or not the test chemical is within the Schiff base
596 reaction domain. In these cases a negative DPRA/ADRA result may be mechanistically justifiable due to
597 the protein-hapten interaction being unfavorable under the test conditions as a result of the abundance
598 of water; particularly for chemicals that are indicated as less potent sensitizers by other methods. In this
599 case, the overall assessment could be considered positive (after expert review) with a low confidence.
600 This positive result is based on giving greater precedence to the mechanistic alert within this domain,
601 and the decreased relevance of the DPRA/ADRA due to the differential reactivity of chemicals within the
602 Schiff base domain. Further, co-elution of the test article with the model nucleophile may lead to false
603 negative predictions, although this occurs to a lesser extent in the ADRA than in the DPRA (Fujita et al.
604 2019).

In cases where the DPRA/ADRA result is positive, but no mechanistic alert can be assigned, it is worth considering whether mechanistic knowledge could be provided by the protein reactivity results particularly when close analogs point to the same structure-activity relationship. Figure 5 shows the 'covalent interaction with skin proteins' endpoint and the confidence score decision tree based on RS1 data. The confidence scores are expected to vary based on reliability and relevance; as such, there are several possible permutations of the decision tree. These general "rules" are expanded to provide a sense of the confidence assigned to assessments with varying reliabilities and relevance, Supplementary Material, section 4 (SM 4).

3.2 Events in keratinocytes

The confidence score obtained for the activation of the events in keratinocytes towards skin sensitization varies based on the Log K_{ow} of the chemical. If there is a positive prediction (RS1, experimental) and the Log K_{ow} is <5, then the result is assigned a high confidence. If the Log K_{ow} is greater than 5, then the confidence is medium for a positive result and low for a negative prediction, since limited information is available for such chemicals (OECD 2018b). Regardless of Log K_{ow} values, negative results could be further assessed based on the occurrence of metabolism and the chemical mechanism of action.

A metabolic alert (indicative of an expected metabolic transformation) along with a negative RS1/2 experimental or RS3 *in silico* result, could indicate reduced relevance of the *in vitro* assays predicting KE2 in this case – possibly because limited metabolic competency of the cells used in the assay are responsible for a false negative. Therefore, the overall assessment would be negative but with a low confidence score. If there is no biochemical transformation predicted, then the chemical mechanism of action could be considered. A negative assessment for a chemical within the acyl transfer domain and Schiff Base domain is conservatively assigned a low confidence score based on the preference of chemicals within these domains for the lysine instead of the cysteine moiety (representing decreased relevance). It is worth mentioning that some chemicals within these domains are accurately predicted as true negatives and a review of the relevance is necessary to assign a higher confidence. Such a review might include an examination of close analogs (or the test structure if data is available) for their assessment in the DPRA/ADRA and or an animal model. If close analogs are positive in the DPRA/ADRA and the lysine moiety; but not cysteine, is implicated for covalent modification then the relevance of the KE2 assays for predicting the test structure may be challenged. However, if cysteine modification is apparent in the DPRA/ADRA (positive for covalent interaction with skin proteins), it is more difficult to

challenge the relevance of the KE2 assays on that basis and conflicting information is presented by the two KEs. The analogs may be further assessed and screened for existing animal data and/or *in silico* predictions of the LLNA or GPMT. This serves the purpose to assess the likelihood of a false negative prediction of the test structure by the KE2 assays. Where a false negative seems likely, the low confidence is appropriate. In cases where the analogs are true negatives, the confidence score could be increased to a medium level and this reflects that while uncertainty is somewhat reduced, there is not absolute certainty in the assessment. Within any other domain, a negative KE2 prediction is considered with high confidence, given RS1/2 data. Varying reliabilities of the data could change the confidence scores in figures 6A and B (see SM Table 9).

3.3 Events in dendritic cells

An overall assessment of the events in dendritic cells could be made based on the h-CLAT (Figure 7), U-SENS™ or IL-8 Luc assays (Figure 8). A positive response from these assays typically translates to a positive overall call for the events in dendritic cells with high confidence in the activation of the dendritic cells towards sensitization, but an expert reviewer would be needed to adjust overall calls and confidence scores for certain chemical classes, structural features, and physical-chemical properties. For example: some chemical classes, such as surfactants, may lead to false positive results in the U-SENS™, and a negative result for a chemical that has a Log K_{ow} greater than 3.5 is considered inconclusive for the h-CLAT. The pro/pre-hapten status of the test chemical is also relevant in each of the three assays. Negative results for structures in which a site of metabolism leading to sensitization has been identified are accepted with a medium level confidence from the h-CLAT, U-SENS™ and IL-8 assays. In cases where there are no additional parameters confounding the prediction, then the confidence level is high for the negative predictions from the h-CLAT, U-SENS™, and IL-8 Luc assays.

3.4 Skin sensitization *in vitro*

Integrating data to derive an overall assessment for the 'skin sensitization *in vitro*' endpoint that correlates with the *in vivo* endpoint is an active area of research. A number of defined approaches (DA) which use varying DIPs have been developed to determine an overall assessment of skin sensitization using non-animal/*in-vitro*/*in silico* models. Any of the DAs described in Section 1 may be adopted here. There has been regulatory acceptance of the "AOP 2 out of 3" approach and the KE3/1 sequential testing strategy (STS) as alternatives to the LLNA for regulatory submission to the United States Environmental Protection Agency (US EPA) (EPA 2018). Here, we discuss how to derive an overall

assessment and confidence when the "AOP 2 out of 3" approach is used within the framework presented in this protocol.

The "AOP 2 out of 3" uses the outcome of three individual assays that map to three KEs to derive a final assessment; however, within the framework presented the assay results are integrated and propagated to the three endpoints related to each key event. The difference between the "AOP 2 out of 3" and the approach used in the framework is subtle, but is worth mention. The "AOP 2 out of 3" approach considers the outcome of the experimental systems – DPRA, KeratinoSens^{TM}, and h-CLAT – but within the framework presented, the: 'Covalent interaction with skin proteins', 'Events in keratinocytes', and 'Events in dendritic cells' (KEs in the AOP) are considered. These KEs are assessed based on knowledge of reaction chemistry and mechanistic understanding that is not explicitly considered within the "AOP 2 out of 3" approach. Similar to the "AOP 2 out of 3," an overall assessment of hazard for the 'skin sensitization *in vitro'* endpoint is determined based on a 2 out of 3 consensus among the endpoints. If outcomes (*in silico*/experimental) are available for only two endpoints, and they have aligned outcomes, the overall assessment of the endpoint is based on the concordant assessments and the lower confidence score propagates. The adoption of the lower confidence score reflects a conservative view of the assessment at this stage of the analysis. However, if the confidence scores have the same value for non-concordant assessments, then the overall prediction for the 'skin sensitization *in vitro*' endpoint is inconclusive. Where there are two concordant assessments, and the non-concordant assessment occurs with high confidence, then the overall confidence could be lowered by one level. Table 2 provides examples showing the derivation of the overall assessment and the rationale for the final confidence score. An alternative point of view suggests that the assays that predict the 'events in keratinocytes', and 'events in dendritic cells', are dependent on the ability of the test chemical to bind protein and therefore point to the activation of the molecular initiating event, 'covalent interaction with skin proteins'. In this point of view, any improvement in predictive performance that results from integrating the KEs across the AOP is a result of reducing the influence of technical limitations of each of the assays (Roberts 2018; Roberts and Patlewicz 2018).

The discussion thus far has focused on assessing hazard from *in vitro* data, but there are also existing strategies for predicting potency in humans from *in vitro* data based on the DAs described in Section 1 and reviewed in (Kleinstreuer et al. 2018). The Artificial Neural Network Model for Predicting LLNA EC3 (Shiseido); Bayesian Network DIP (BN-ITS-3) for Hazard and Potency Identification of Skin Sensitizers (P&G); Sequential Testing Strategy (STS) for Sensitizing Potency Classification Based on *in Chemico* and *In Vitro* Data (Kao); and ITS for Sensitizing Potency Classification Based on *In Silico*; *In Chemico*, and *In Vitro* Data (Kao) were found to predict potency class equally well, or better than the LLNA. Similar to the earlier discussion on hazard, the DAs for assessing potency use biological assay outcomes (mechanisms/effects assessment within the HAF e.g. DPRA, KeratinoSens™, h-CLAT) as endpoints and may integrate the information with *in silico* methods to determine a potency class. Within the HAF presented, the assay outcomes (*in vitro*/*in silico* effects/mechanisms assessment) are interpreted in the context of their toxicological significance and integrated to determine a toxicological endpoint according to the rules and principles outlined in previous sections. The overall assessments of the KE endpoints may substitute for the outcome of the individual test methods in data interpretation procedures.

3.5 Skin sensitization *in vitro* to skin sensitization in human extrapolation

Extrapolation of *in vitro* skin sensitization results to human skin sensitization predictions is necessary to satisfy the European Union's 7th Amendment of the Cosmetic Directive and REACH regulations which require and prefer the use of non-animal test methods for assessing the human skin sensitization endpoint. The definition of the AOP and the mechanistic information provided by the assays that map to the AOP allow the human hazard identified for the 'skin sensitization *in vitro*' outcome to be propagated to the human endpoint. The relevance of the integrated *in vitro* battery of tests is equally weighted with the *in vivo* studies except in unique cases; for example, when metabolism is thought to influence the outcome. As such, no change in confidence (reliability and relevance of the prediction) is expected due to the extrapolation of *in vitro* hazard.

3.6 Skin sensitization in rodent lymphocytes

A negative result in the LLNA is propagated to the skin sensitization in rodent lymphocytes endpoint with high confidence. A weak sensitizer may require investigation of the skin irritation potential of the chemical, particularly if the result is derived from a lower-reliability study that may not have considered irritation prior to designing the test. The skin irritation potential will be determined through a HAF that will be published in a separate protocol. Positive results due to confounding factors from irritants usually

result in a low-level increase in lymphocytes which could be misinterpreted as a weak sensitizing response. In cases where a chemical is found to have a strong skin irritation potential and is a weak sensitizer and the influence of irritation cannot be ruled out, a positive assessment with low confidence could be assigned to the 'Events in rodent lymphocytes' endpoint (Figure 9).

3.7 Skin sensitization in rodents

This endpoint integrates guinea pig (GPMT and BT) and mouse (LLNA) data. In the absence of LLNA data, the endpoint could be determined through the scheme shown in Figure 10. If guinea pig tests are not conducted according to standard protocols, irritation could become a confounding factor in the interpretation of the guinea pig test results and influence the relevance of the study (OECD 1992). Freund's complete adjuvant (FCA) is used to maximize the guinea pig response; however, FCA may also lower the irritation threshold. The implication is that concentrations that were identified as non-irritating and suitable for the challenge reaction might in fact produce an irritant response. Further, a hyperirritable state may be induced by the test article during the induction phase that is not represented in the control, unless a suitably irritating surrogate is used to induce the hyperirritable state in the controls (Kligman and Basketter 1995; OECD 1992). An irritant effect cannot be distinguished from an allergic response by visual examination. As such, post challenge examination is helpful in distinguishing a sensitization response from an irritant effect. Chemicals that are identified as irritants could be confidently predicted as non-sensitizers if observations of erythema dissipate within one day of challenge and/or there is a negative re-challenge test one week after the initial challenge (Kligman and Basketter 1995). A positive result for a chemical that is irritating but predicted to be a weak sensitizer is afforded a low confidence score if deviations from OECD 1996 result in decreased reliability and relevance of the study as discussed above.

When both guinea pig and mouse data are available and are concordant, then the result is translated to the 'skin sensitization in rodent' endpoint with exact or higher confidence scores being adopted. For example, if the LLNA is positive with medium confidence and the GPMT/BT is positive with low confidence, then the skin sensitization in rodent endpoint is assessed as positive with medium confidence. In cases where the data are discordant, the strategy for deriving an overall assessment may vary case-to-case. A high reliability guinea pig test has an advantage over the LLNA because it includes both induction and challenge phases, and is as such, more representative of the entire sensitization process. However, in contrast to the LLNA, the guinea pig test results are based on a qualitative measure and a subjective endpoint. Potency is better assessed through the LLNA since it is derived from dose-

response relationships and the read-out is quantitative; nonetheless, some chemical classes are over-classified in the LLNA. It is valuable to consider how the challenge reaction affects interpretation of an assessment. It could be argued that the LLNA is an assay and non-specific reactions can occur that may or may not relate to allergenic potential (respiratory sensitizers test positive in the LLNA, for example) while the dermal challenge in the guinea pig tests lends more confidence that any observations of sensitization are specific to the skin. A default principle that could be adopted is to evaluate the 'skin sensitization in rodent' endpoint based on either the LLNA or GPMT/BT assessment with the higher confidence score and conservatively decrease the score by one level to reflect any uncertainty. For example, an LLNA that is assessed as positive with medium confidence, and a GP test that is negative with low confidence, would lead to a 'skin sensitization in rodent' assessment as positive with low confidence. In these circumstances, a review of the predictions is prudent and the assessment and confidence scores may be adjusted based on the review.

3.8 Skin sensitization in rodents to skin sensitization in human extrapolation

There are two schools of thought on rodent-to-human extrapolation that draw from a two different perspectives on risk assessment: one is that LLNA potency categories and EC3 values correlate well with human potency categories and NOEL values, and could therefore be used as a surrogate for the NOEL and for direct prediction of human potency class (Basketter et al. 2005). Alternatively, a safety factor may be incorporated based on the interspecies variation that may occur between the mouse and humans; although, this factor could be lowered in cases where a better correlation may be expected (e.g., based on existing human data for a close analogue) (Roberts and Api 2018). Roberts and Api 2018, have defined alerts for cases where the LLNA is not a good predictor of human potency. Guinea pig tests also provide relevant information on hazard and potency. However, tests that use adjuvant and intradermal routes of exposure (GPMT) present a challenge for interpreting human potency, and in those situations potency estimation via the BT may be more relevant. The data however, could serve in a weight-of-evidence case for potency determination through interpretation and comparison of different test results and also with known benchmark chemicals (Kimber et al. 2001).

3.9 Skin sensitization in humans

The 'skin sensitization in humans' endpoint could be evaluated through several other endpoints such as the 'skin sensitization *in vitro'* endpoint (section 3.5), the 'skin sensitization in rodent endpoints' (section 3.8), or through the integration of the 'skin sensitization *in vitro*', 'skin sensitization in rodents', and human assessments, combined with supporting data from non-standard endpoints such as photoallergy.

A positive HMT/HRIPT is indicative of adverse outcome in humans and can potentially be used to assign a potency class. In the absence of reliable studies, other sources of evidence may be sought. The first line of evidence arises from the toxicological relationships that could be drawn from the chemical's structure. The presence of a structural alert for sensitization in humans provides evidence for the elicitation of the adverse outcome. Structural alerts and diagnostic patch testing with positive incidences in greater than 1% of the population (considered to be high incidence) in relation to low usage volume (a measure of exposure) provides evidence for the skin sensitization potential of a chemical, although it does not provide a definite assessment (Api et al. 2017). If a compound has no structural alerts and diagnostic patch testing data indicate < 1% frequency, the overall evidence may together indicate a negative assessment, especially if the use volume is high. It is important to note that the indication of a 1% incidence rate is based on expert opinion and as such is not meant to represent a rule that requires strict compliance. Many combinations of scenarios are possible.

In cases where human and *in vitro/in vivo* sensitization assessments do not align, additional information could be gathered from the 'skin sensitization *in vitro*' and/or 'skin sensitization in rodents' endpoints to build a weight of evidence case. There are many permutations of assay results at this level but some general guidance can be provided to the evaluator towards an overall assessment. It is generally recommended that the assessments that are assigned more frequently should be propagated to the overall human endpoint. However, if reliable human data (RS1/2) is available, then the assessment of this data is given priority in the decision-making process. Table 13 of the supplementary material expands on the principles to derive an overall assessment given *in vitro* and rodent evidence. Due to ethical concerns, human testing is no longer considered appropriate for most compounds, so much of the human data is older, or based on clinical reports, and may therefore lack information to assess its quality, necessitating the filter of expert opinion. Careful consideration is required in assessing confidence of the HMT and HRIPTs. For the HMT and, especially for the HRIPT as used by the fragrance industry, low doses are often tested as the goal is to corroborate an animal study while trying to avoid sensitizing the subjects. Therefore, there can be quite a bit of uncertainty in a negative result because a higher test concentration could potentially produce a positive result in humans.

Table 3 shows factors to consider in assigning confidence to a human study in general. There are however some specific exceptions to these criteria when assessing the HMT and HRIPTs. The exposure scenarios in the HMT and HRIPT may not represent real-world exposure because the test chemical is applied under occlusive conditions and the outcomes can be viewed as subjective because an observer grades the skin reaction.

4. Case Studies

The case studies demonstrate the interpretation of results when a series of statistical models ((Q)SARs), structural alerts, or read-across are used to fill data gaps for effects and mechanisms that are included in the hazard assessment framework. The studies demonstrate how aspects of the rules and principles are implemented to derive an assessment, reliability score and confidence score, when the assessment is made using either or both existing experimental data or *in silico* methods.

4.1. Case 1a: Compound with conflicting data ("Skin Sensitization *in vitro*" endpoint determination)

An assessor needed to determine the hazard associated with a compound. The compound was predicted to be reactive towards proteins via an Acyl or SN2 reaction, and could be assigned to a reaction domain based on reaction chemistry alerts. Data that was generated based on OECD TG 442C (DPRA) was available for the compound. The data indicated that the compound was negative for protein reactivity. Based on adherence to the test guideline, a reliability score of RS1 was assigned to the study. *In silico* tools (statistical results (QSAR) and alerts) were available for the DPRA prediction, and these predictions were also negative. The statistical model and the alerts both had a reliability score of RS5. *In silico* assessments of dermal metabolism were negative after an expert review. The review increased the reliability of the dermal metabolism alert from RS5 to RS3. The overall assessment for 'covalent interaction with skin proteins' was negative; however, the confidence was assigned as medium, based on the conflicting mechanistic/reaction chemistry alert for protein reactivity, Figure 11a.

There is experimental data for the KeratinoSens™ assay which is afforded a positive assessment with a reliability score of RS1 (the study adhered to OECD TG 442D), so the overall assessment for the 'Events in Keratinocytes' KE is positive with high confidence. Experimental data is not available for the 'Events in Dendritic Cells' KE. The assessor would like to use the "2 out of 3" approach and is faced with two conflicting assessments based on *in vitro* data. A statistical model (QSAR) was used to predict the results of the h-CLAT assay and the assessment is negative with a reliability score of RS3, after an expert review. The overall assessment of the 'Events in Dendritic Cells' KE is negative with a medium confidence. Based on the two concordant assessments with aligned confidence scores (Negative, Medium confidence), and a third assessment that is conflicting with high confidence (Positive, High confidence), the overall assessment of *in vitro* skin sensitization endpoint is negative with low confidence.

4.2 Case 1b: Compound with conflicting data ('Skin Sensitization in Humans' endpoint determination)

A further assessment was completed for the same compound as in Case 1a. This assessor has LLNA and GPMT data with conflicting assessments. The LLNA data is positive with an EC3 (%) value that indicates weak sensitization. The study is assigned the lowest reliability score of 5 based on significant deviations from OECD Test No. 429 that could alter both the reliability and relevance of the study. *In silico* assessments using expert alerts and statistical models are both negative. The weak sensitizing effect and the mis-aligned *in silico* results prompt the assessor to consider the irritation potential of the chemical. Experimental data is available for the *in vitro* skin irritation test using the Reconstructed Human Epidermis (RHE) test method. The assessment of skin irritation is positive with a score of RS1. The assessor conducts an expert review of the LLNA and suspects a false positive LLNA result. The 'Events in Rodent Lymphocytes' endpoint could be assigned as positive with low confidence; however, the negative *in silico* results are more reliable and relevant in this situation and the negative assessment carries over to the 'Events in Rodent Lymphocytes' endpoint with medium confidence. The GPMT data is negative with a reliability of RS1 since the study adhered to OECD 406 and the irritant effect was considered in the study design and interpretation of results. *In silico* models agree with the experimental GPMT result. The overall assessment of 'Skin sensitization in rodents' is negative with a high confidence, Figure 11b.

To further investigate the outcome in humans, the assessor conducted an *in* silico assessment using a set of alerts that were developed using HMT and HRIPT data as a reference database and no alerting structure were found. No human study data were available; however, DPT data were available and consecutive patients showed frequencies of 0% in a study. The absence of positive DPT results are indicative of no sensitization in humans, although a conclusion cannot be made from DPT data alone.

Given the weight of evidence presented in Case 1a and 1b a final determination of the 'skin sensitization in humans' can be made. In this case, a well conducted GPMT carried significant weight towards the negative sensitization assessment with high confidence; reflecting the high reliability and relevance of the information. Other evidence supporting a negative assessment included a negative protein binding test which was reinforced by negative *in silico* models predictions of protein binding; and negative (Q)SARs predicting the 'Events in dendritic cells', and LLNA. The conflicting piece of information presented by the LLNA study was viewed as less reliable and relevant information due primarily to confounding irritant effects in the study. A second piece of conflicting information was presented by the KeratinoSens™ experimental study. While no specific explanation for this false positive was

892 determined, the body of negative evidence for the 'skin sensitization *in vitro*' endpoint supports the
893 negative assessment and the low confidence reflects any uncertainty in the assessment of that
894 endpoint. However, the 'sensitization *in vitro*' assessment does not discredit the 'skin sensitization in
895 rodents' assessment. Since the *in vitro* and rodent endpoints are both equally relevant when the *in vitro*
896 endpoint is derived through a defined approach, the endpoint that contains more reliable information
897 contributes more to the overall confidence. The *in vitro* endpoint does not introduce any uncertainty in
898 the GPMT experimental findings, and taken together with the DPT data, the final confidence score is
899 high in this negative case, Figure 11c. There may be instances where a higher level of conservatism is
900 necessary than presented. In such instances, the confidence score could be reduced to medium,
901 although a change in the assessment might be difficult to justify.

902 **4.3 Case 2a: Pro/pre-hapten assessment**

903 Figure 12a details the assessment for the mechanisms/effects that were considered in the case. A
904 chemical is being screened for possible use in the cosmetics industry. It is expected to undergo
905 metabolic transformation leading to the formation of quinones, which have a high probability to react
906 via Michael addition (MA). There are positive alerts for dermal metabolism and the site of metabolism
907 coincides with a pro-MA reactivity alert. Negative DPRA data are available and the DPRA study is
908 assigned a reliability score of 1 based on adherence to OECD TG 442C. However, based on knowledge
909 that the compound contains a pro-reactive feature that coincides with a site of metabolism, the
910 relevance of the DPRA for testing the compound is challenged since any activity that results from a
911 metabolic transformation may be missed. The DPRA test is considered not relevant for the compound
912 tested, and the assessment of the 'Covalent interaction with skin proteins' is based on the assignment of
913 a pro-reactive domain. Although there may be cases where a pro-reactive domain assignment does not
914 lead to protein interaction due to deactivating features, a conservative approach to assessing the
915 endpoint given a pro-reactive feature is to assign a positive assessment with a lowered confidence. No
916 other *in vitro* data are available for the compound. A (Q)SAR was developed based on proprietary data
917 for the KeratinoSens™ assay. The test compound is assessed as positive in the (Q)SAR, with the pro-MA
918 feature identified as significant by the model. After a review of the (Q)SAR prediction, the 'Events in
919 Keratinocytes' is assessed as positive with medium confidence. No data or models were available for the
920 'Events in dendritic cells' endpoint. Given the positive assessment for 'Covalent interaction with skin
921 proteins' and the 'Events in Keratinocytes', the overall assessment for the 'Skin Sensitization *in vitro*'
922 endpoint is made using the "2 out of 3" approach. The overall assessment of the 'Skin Sensitization *in*
923 *vitro*' endpoint is positive with low confidence based on the two aligned positive assessments and the

lower confidence score propagating to the endpoint, Figure 12a. It is possible to extrapolate the existing hazard information to the 'Skin sensitization in humans' endpoint and assess it as positive with low confidence.

4.4 Case 2b: Pro/pre-hapten assessment Example 2

Consider an extension of the case presented in Section 6.3. LLNA data are not available for the test compound but are available for close analogs. In addition there is a low quality guinea pig test for the test compound that indicates a positive sensitization response. Read-across is performed using the LLNA data for the analogs. The analogs all contained the pro-reactive feature and formed a congeneric series that allowed interpolation of the LLNA EC3 value. The EC3 value was predicted to be 3.2%, indicative of a moderate sensitizer. The 'Events in rodent lymphocytes' endpoint was assessed as positive with medium confidence based on the read-across result. The guinea pig test is assigned a reliability score of RS5 based on deviations from OECD 406. A review of the study showed that for an induction concentration of 1%, the sensitization incidence is 100% suggesting that the compound could be classified as a Category 1A sensitizer. After an expert review of the study, the reliability score is increased to RS3. The overall assessment of the 'Skin Sensitization in Rodents' endpoint is assessed as positive, with medium confidence based on the weight of evidence presented by the LLNA read-across and guinea pig study.

The 'Skin sensitization in rodents' and the 'Skin sensitization *in vitro*' endpoints both support that assignment of a positive hazard for the 'skin sensitization in humans' with medium confidence. Figure 12b shows the flow of information within the hazard assessment framework.

5. Reporting

An important consideration towards *in silico* standardization, reproducibility and transparency is a consistent reporting format (Myatt et al. 2018). The general protocol (Myatt et al. 2018) describes a proposed reporting format that includes the elements that provide completeness of information. The report format is reproduced in Table 4 with a minor modification for the skin sensitization endpoint. In addition to the description of models, databases, and tools that were used, it is also recommended to describe any IATAs, DIPs or DAs that were used in deriving the overall assessment. The details that are suggested should allow another expert to repeat the process and achieve the same results. Further, the standardized report enables streamlined and consistent review of regulatory submissions across industries and endpoints. Section 5 of the Supplementary Material (SM5) provides an example of a report for sensitization hazard.

6. Conclusion

The skin sensitization *in silico* protocol presented here is the first publication to outline a systematic assessment of skin sensitization based on both experimental data and *in silico* predictions. It includes a HAF and provides general rules for the *in silico* toxicological assessment of chemicals within the framework. The framework is transparent and flexible as it does not require the generation of all endpoints to derive an overall assessment of 'skin sensitization in humans' and can accommodate quantitative and qualitative predictions and/or experimental results. There are cases where extrapolation to the human endpoint is possible and this has been described. The corresponding assessment of the confidence for all endpoints allows the protocol to be used in a variety of use cases. For example, assessments with low confidence scores may still have practical usage in screening or prioritization use cases. In addition, the protocol highlights experimental approaches or *in silico* models that could be incorporated into the HAF in the near future. Expert review is a critical element in any such procedure and items to consider as part of this review are listed to support a more consistent assessment. The standardization of the HAF for performing *in silico* methods is designed to support increased use and acceptance of *in silico* tools among regulatory agencies and industries alike.

7. 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 R43ES026909. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Tables

Figures

Figures 6A. Decision trees showing how an overall assessment and confidence score could be derived for the 'events in keratinocytes'. The confidence scores here are based on RS1 experimental data: assuming relevant data and high reliability, and, in practice, confidence scores may need to be adjusted based on reliability scores.

Figures 6B. Decision trees showing how an overall assessment and confidence score could be derived for the 'events in keratinocytes'. The confidence scores here are based on RS1 experimental data: assuming relevant data and high reliability, and, in practice, confidence scores may need to be adjusted based on reliability scores.

Figure 7. Decision tree showing how an overall assessment and confidence score could be derived for the 'events in dendritic cells' based on the h-CLAT assay. The confidence scores here are based on RS1 experimental data: assuming relevant data and high reliability, and, in practice, confidence scores may need to be adjusted based on reliability scores.

Figure 8. Decision tree showing how an overall assessment and confidence score could be derived for the 'events in dendritic cells' based on the U-SENS™ and IL-8 Luc assay data. The confidence scores here are based on RS1 experimental data: assuming relevant data and high reliability, and, in practice, confidence scores may need to be adjusted based on reliability scores, SM Table 10.

Figure 9. Decision tree showing how an overall assessment and confidence score could be derived for the "Events in rodent lymphocytes" based on the LLNA. The confidence scores here are based on RS1/2 experimental data (except in the case of *): assuming relevant data and high reliability, and, in practice, confidence scores may need to be adjusted based on reliability scores. *Concentrations tested in the LLNA are either non-irritating or mildly irritating. The low confidence score reflects the non-specific increase in lymphocyte proliferation that could occur with irritants.

Figure 10. Decision tree showing how an overall assessment and confidence score could be derived for the 'skin sensitization in rodents' endpoint based on guinea pig tests. The confidence scores here are based on RS1 experimental data (except in the case of *): assuming relevant data and high reliability, and, in practice, confidence scores may need to be adjusted based on reliability scores. *GPMT/BT

challenge concentrations are non-irritating; however, deviations from OECD 406 may reduce the relevance of the study and decrease the confidence in the endpoint.

Figure 11a. Derivation of the 'skin sensitization *in vitro*' endpoint using the "AOP 2 out of 3" approach (Case 1a)

Figure 11b. Derivation of the 'Skin Sensitization in Rodents' endpoint

Figure 11c. Derivation of the 'skin Sensitization in Humans' endpoint from the weight of evidence presented from the 'Skin Sensitization skin *in vitro*' and 'Skin Sensitization in Rodents' endpoints. DPT data is also used to support the overall assessment.

Figure 12a. Derivation of the 'skin sensitization *in vitro*' endpoint using the "AOP 2 out of 3" approach (Case 2)

Figure 12b. Derivation of the 'Skin Sensitization in Humans' using the "AOP 2 out of 3" approach (Case 2)

References

Anderson, Stacey E., Paul D. Siegel, and B. J. Meade. 2011. "The LLNA: A Brief Review of Recent Advances and Limitations." *Journal of Allergy* 2011:424203.

Api, Anne Marie, Rahul Parakhia, Devin O'Brien, and David A. Basketter. 2017. "Fragrances Categorized According to Relative Human Skin Sensitization Potency." *Dermatitis : Contact, Atopic, Occupational, Drug* 28(5):299–307.

Aptula, Aynur O. and David W. Roberts. 2006. "Mechanistic Applicability Domains for Nonanimal-Based Prediction of Toxicological End Points: General Principles and Application to Reactive Toxicity." *Chemical Research in Toxicology* 19(8):1097–1105.

Ball, Nicholas, Stuart Cagen, Juan-Carlos Carrillo, Hans Certa, Dorothea Eigler, Roger Emter, Frank Faulhammer, Christine Garcia, Cynthia Graham, Carl Haux, Susanne N. Kolle, Reinhard Kreiling, Andreas Natsch, and Annette Mehling. 2011. "Evaluating the Sensitization Potential of Surfactants: Integrating Data from the Local Lymph Node Assay, Guinea Pig Maximization Test, and in Vitro Methods in a Weight-of-Evidence Approach." *Regulatory Toxicology and Pharmacology* 60(3):389–400.

Basketter, D. A., G. F. Gerberick, and I. Kimber. 2001. "Skin Sensitisation, Vehicle Effects and the Local Lymph Node Assay." *Food and Chemical Toxicology* 39(6):621–27.

Basketter, David A., Catherine Clapp, Donna Jefferies, Bob Safford, Cindy A. Ryan, Frank Gerberick, Rebecca J. Dearman, and Ian Kimber. 2005. "Predictive Identification of Human Skin Sensitization Thresholds." *Contact Dermatitis* 53(5):260–67.

Basketter, David A., John F. McFadden, Frank Gerberick, Amanda Cockshott, and Ian Kimber. 2009. "Nothing Is Perfect, Not Even the Local Lymph Node Assay: A Commentary and the Implications for REACH." *Contact Dermatitis* 60(2):65–69.

Dumont, Coralie, João Barroso, Izabela Matys, Andrew Worth, and Silvia Casati. 2016. "Analysis of the Local Lymph Node Assay (LLNA) Variability for Assessing the Prediction of Skin Sensitisation Potential and Potency of Chemicals with Non-Animal Approaches." *Toxicology in Vitro* 34:220–28.

Dumont, Coralie, Pilar Prieto, David Asturiol, and Andrew Worth. 2015. "Review of the Availability of In Vitro and In Silico Methods for Assessing Dermal Bioavailability." *Applied In Vitro Toxicology* 1(2):147–64.

Enoch, S. J., J. C. Madden, and M. T. D. Cronin. 2008. "Identification of Mechanisms of Toxic Action for Skin Sensitisation Using a SMARTS Pattern Based Approach." *SAR and QSAR in Environmental Research* 19(5–6):555–78.

EPA. 2018. *Draft Interim Science Policy: Use of Alternative Approaches for Skin Sensitization as a Replacement for Laboratory Animal Testing*.

Fitzpatrick, Jeremy M., David W. Roberts, and Grace Patlewicz. 2017. "What Determines Skin Sensitization Potency: Myths, Maybes and Realities. The 500 Molecular Weight Cut-off: An Updated Analysis." *Journal of Applied Toxicology* 37(1):105–16.

Fujita, Masaharu, Yusuke Yamamoto, Sayaka Wanibuchi, Yasuhiro Katsuoka, and Toshihiko Kasahara. 2019. "The Underlying Factors That Explain Why Nucleophilic Reagents Rarely Co-Elute with Test

1109 Chemicals in the ADRA." *Journal of Pharmacological and Toxicological Methods* 96:95–105.

1110 Hoffmann, Sebastian. 2015. "LLNA Variability: An Essential Ingredient for a Comprehensive Assessment
1111 of Non-Animal Skin Sensitization Test Methods and Strategies." *ALTEX - Alternatives to Animal*
1112 *Experimentation* 32(4 SE-Short communications).

1113 Hoffmann, Sebastian, Nicole Kleinstreuer, Nathalie Alépée, David Allen, Anne Marie Api, Takao Ashikaga,
1114 Elodie Clouet, Magalie Cluzel, Bertrand Desprez, Nichola Gellatly, Carsten Goebel, Petra S. Kern,
1115 Martina Klaric, Jochen Kühnl, Jon F. Lalko, Silvia Martinozzi-Teissier, Karsten Mewes, Masaaki
1116 Miyazawa, Rahul Parakhia, Erwin van Vliet, Qingda Zang, and Dirk Petersohn. 2018. "Non-Animal
1117 Methods to Predict Skin Sensitization (I): The Cosmetics Europe Database." *Critical Reviews in*
1118 *Toxicology* 48(5):344–58.

1119 Kimber, I., D. A. Basketter, K. Berthold, M. Butler, J. L. Garrigue, L. Lea, C. Newsome, R. Roggeband, W.
1120 Steiling, G. Stropp, S. Waterman, and C. Wiemann. 2001. "Skin Sensitization Testing in Potency and
1121 Risk Assessment." *Toxicological Sciences* 59(2):198–208.

1122 Kleinstreuer, Nicole C., Sebastian Hoffmann, Nathalie Alépée, David Allen, Takao Ashikaga, Warren
1123 Casey, Elodie Clouet, Magalie Cluzel, Bertrand Desprez, Nichola Gellatly, Carsten Göbel, Petra S.
1124 Kern, Martina Klaric, Jochen Kühnl, Silvia Martinozzi-Teissier, Karsten Mewes, Masaaki Miyazawa,
1125 Judy Strickland, Erwin van Vliet, Qingda Zang, and Dirk Petersohn. 2018. "Non-Animal Methods to
1126 Predict Skin Sensitization (II): An Assessment of Defined Approaches." *Critical Reviews in Toxicology*
1127 48(5):359–74.

1128 Kligman, A. M. and D. A. Basketter. 1995. "A Critical Commentary and Updating of the Guinea Pig
1129 Maximization Test." *Contact Dermatitis* 32(3):129–34.

1130 Madden, J. C., S. Webb, S. J. Enoch, H. E. Colley, C. Murdoch, R. Shipley, P. Sharma, C. Yang, and M. T. D.
1131 Cronin. 2017. "In Silico Prediction of Skin Metabolism and Its Implication in Toxicity Assessment."
1132 *Computational Toxicology* 3:44–57.

1133 Myatt, G. J., E. Ahlberg, Y. Akahori, D. Allen, A. Amberg, L. T. Anger, A. Aptula, S. Auerbach, L. Beilke, P.
1134 Bellion, R. Benigni, J. Bercu, E. D. Booth, D. Bower, A. Brigo, N. Burden, Z. Cammerer, M. T. D.
1135 Cronin, K. P. Cross, L. Custer, M. Dettwiler, K. Dobo, K. A. Ford, M. C. Fortin, S. E. Gad-McDonald, N.
1136 Gellatly, V. Gervais, K. P. Glover, S. Glowienke, J. Van Gompel, S. Gutsell, B. Hardy, J. S. Harvey, J.
1137 Hillegass, M. Honma, J. H. Hsieh, C. W. Hsu, K. Hughes, C. Johnson, R. Jolly, D. Jones, R. Kemper, M.
1138 O. Kenyon, M. T. Kim, N. L. Kruhlak, S. A. Kulkarni, K. Kümmerer, P. Leavitt, B. Majer, S. Masten, S.
1139 Miller, J. Moser, M. Mumtaz, W. Muster, L. Neilson, T. I. Oprea, G. Patlewicz, A. Paulino, E. Lo
1140 Piparo, M. Powley, D. P. Quigley, M. V. Reddy, A. N. Richarz, P. Ruiz, B. Schilter, R. Serafimova, W.
1141 Simpson, L. Stavitskaya, R. Stidl, D. Suarez-Rodriguez, D. T. Szabo, A. Teasdale, A. Trejo-Martin, J. P.
1142 Valentin, A. Vuorinen, B. A. Wall, P. Watts, A. T. White, J. Wichard, K. L. Witt, A. Woolley, D.
1143 Woolley, C. Zwickl, and C. Hasselgren. 2018. "In Silico Toxicology Protocols." *Regulatory Toxicology*
1144 *and Pharmacology* 96.

1145 Natsch, Andreas. 2010. "The Nrf2-Keap1-ARE Toxicity Pathway as a Cellular Sensor for Skin Sensitizers—
1146 Functional Relevance and a Hypothesis on Innate Reactions to Skin Sensitizers." *Toxicological*
1147 *Sciences* 113(2):284–92.

1148 Natsch, Andreas, Roger Emter, Hans Gfeller, Tina Haupt, and Graham Ellis. 2015. "Predicting Skin
1149 Sensitizer Potency Based on In Vitro Data from KeratinoSens and Kinetic Peptide Binding: Global

Versus Domain-Based Assessment." *Toxicological Sciences* 143(2):319–32.

Natsch, Andreas, Hans Gfeller, Tina Haupt, and Gerhard Brunner. 2012. "Chemical Reactivity and Skin Sensitization Potential for Benzaldehydes: Can Schiff Base Formation Explain Everything?" *Chemical Research in Toxicology* 25(10):2203–15.

Natsch, Andreas and Tina Haupt. 2013. "Utility of Rat Liver S9 Fractions to Study Skin-Sensitizing Prohaptens in a Modified KeratinoSens Assay." *Toxicological Sciences* 135(2):356–68.

Nukada, Yuko, Takao Ashikaga, Hitoshi Sakaguchi, Sakiko Sono, Nanae Mugita, Morihiko Hirota, Masaaki Miyazawa, Yuichi Ito, Hitoshi Sasa, and Naohiro Nishiyama. 2011. "Predictive Performance for Human Skin Sensitizing Potential of the Human Cell Line Activation Test (h-CLAT)." *Contact Dermatitis* 65(6):343–53.

OECD. 1992. "Test No. 406: Skin Sensitisation:OECD Guidelines for the Testing of Chemicals, Section 4." *OECD Publishing, Paris*.

OECD. 2010a. "Test No. 429: Skin Sensitisation: Local Lymph Node Assay, OECD Guidelines for the Testing of Chemicals, Section 4." *OECD Publishing, Paris*.

OECD. 2010b. "Test No. 442A: Skin Sensitization: Skin Sensitization: Local Lymph Node Assay: DA, OECD Guidelines for the Testing of Chemicals, Section 4." *OECD Publishing, Paris*.

OECD. 2011. "Report of the Expert Consultation on Scientific and Regulatory Evaluation of Organic Chemistry Mechanism-Based Structural Alerts for the Identification of Protein-Binding Chemicals." *Series on Testing and Assessment No. 139*.

OECD. 2014. *The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins*. OECD.

OECD. 2017. *Guidance Document on the Reporting of Defined Approaches and Individual Information Sources to Be Used within Integrated Approaches to Testing and Assessment (IATA) for Skin Sensitisation*. OECD.

OECD. 2018a. "Test No. 442B: Skin Sensitization: Local Lymph Node Assay: BrdU-ELISA or –FCM, OECD Guidelines for the Testing of Chemicals, Section 4." *OECD Publishing, Paris*.

OECD. 2018b. "Test No. 442D: In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method, OECD Guidelines for the Testing of Chemicals, Section 4." *OECD Publishing, Paris*.

Roberts, D. W. and A. O. Aptula. 2014. "Electrophilic Reactivity and Skin Sensitization Potency of S N Ar Electrophiles." *Chemical Research in Toxicology* 27(2):240–46.

Roberts, D. W., A. O. Aptula, M. T. D. Cronin, E. Hulzebos, and G. Patlewicz. 2007. "Global (Q)SARs for Skin Sensitisation–Assessment against OECD Principles‖." *SAR and QSAR in Environmental Research* 18(3–4):343–65.

Roberts, D. W., A. O. Aptula, and G. Y. Patlewicz. 2011. "Chemistry-Based Risk Assessment for Skin Sensitization: Quantitative Mechanistic Modeling for the S N Ar Domain." *Chemical Research in Toxicology* 24(7):1003–11.

Roberts, D. W., R. Fraginals, J. P. Lepoittevin, and C. Benezra. 1991. "Refinement of the Relative

Alkylation Index (RAI) Model for Skin Sensitization and Application to Mouse and Guinea-Pig Test Data for Alkyl Alkanesulphonates." *Archives of Dermatological Research* 283(6):387–94.

Roberts, D. W. and D. L. Williams. 1982. "The Derivation of Quantitative Correlations between Skin Sensitisation and Physio-Chemical Parameters for Alkylating Agents, and Their Application to Experimental Data for Sultones." *Journal of Theoretical Biology* 99(4):807–25.

Roberts, David W. 2018. "Is a Combination of Assays Really Needed for Non-Animal Prediction of Skin Sensitization Potential? Performance of the GARD™ (Genomic Allergen Rapid Detection) Assay in Comparison with OECD Guideline Assays Alone and in Combination." *Regulatory Toxicology and Pharmacology* 98:155–60.

Roberts, David W. and Anne Marie Api. 2018. "Chemical Applicability Domain of the Local Lymph Node Assay (LLNA) for Skin Sensitisation Potency. Part 4. Quantitative Correlation of LLNA Potency with Human Potency." *Regulatory Toxicology and Pharmacology* 96:76–84.

Roberts, David W., Anne Marie Api, Robert J. Safford, and Jon F. Lalko. 2015. "Principles for Identification of High Potency Category Chemicals for Which the Dermal Sensitisation Threshold (DST) Approach Should Not Be Applied." *Regulatory Toxicology and Pharmacology* 72(3):683–93.

Roberts, David W. and Aynur O. Aptula. 2008. "Determinants of Skin Sensitisation Potential." *Journal of Applied Toxicology* 28(3):377–87.

Roberts, David W., Aynur O. Aptula, and Grace Patlewicz. 2006. "Mechanistic Applicability Domains for Non-Animal Based Prediction of Toxicological Endpoints. QSAR Analysis of the Schiff Base Applicability Domain for Skin Sensitization." *Chemical Research in Toxicology* 19(9):1228–33.

Roberts, David W. and Andreas Natsch. 2009. "High Throughput Kinetic Profiling Approach for Covalent Binding to Peptides: Application to Skin Sensitization Potency of Michael Acceptor Electrophiles." *Chemical Research in Toxicology* 22(3):592–603.

Roberts, David W. and Grace Patlewicz. 2018. "Non-Animal Assessment of Skin Sensitization Hazard: Is an Integrated Testing Strategy Needed, and If so What Should Be Integrated?" *Journal of Applied Toxicology* 38(1):41–50.

RUSSELL, W. M. S. and R. L. BURCH. 1959. *The Principles of Humane Experimental Technique.* London: Methuen & Co. Ltd.

Schulz, K. F., D. G. Altman, and D. Moher. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMJ* 340(mar23 1):c332–c332.

Sibbald, B. and M. Roland. 1998. "Understanding Controlled Trials. Why Are Randomised Controlled Trials Important?" *BMJ (Clinical Research Ed.)* 316(7126):201.

Sumpter, Tina L., Stephen C. Balmert, and Daniel H. Kaplan. 2019. "Cutaneous Immune Responses Mediated by Dendritic Cells and Mast Cells." *JCI Insight* 4(1).

Urbisch, Daniel, Annette Mehling, Katharina Guth, Tzutzuy Ramirez, Naveed Honarvar, Susanne Kolle, Robert Landsiedel, Joanna Jaworska, Petra S. Kern, Frank Gerberick, Andreas Natsch, Roger Emter, Takao Ashikaga, Masaaki Miyazawa, and Hitoshi Sakaguchi. 2015. "Assessing Skin Sensitization Hazard in Mice and Men Using Non-Animal Test Methods." *Regulatory Toxicology and*

Pharmacology 71(2):337–51.

Vocanson, M., A. Hennino, A. Rozières, G. Poyet, and J. F. Nicolas. 2009. "Effector and Regulatory Mechanisms in Allergic Contact Dermatitis." *Allergy* 64(12):1699–1714.

Wang, Chia Chi, Ying Chi Lin, Shan Shan Wang, Chieh Shih, Yi Hui Lin, and Chun Wei Tung. 2017. "SkinSensDB: A Curated Database for Skin Sensitization Assays." *Journal of Cheminformatics*.

Tables

Table 1. Sources of data for the development of *in silico* methods

Database	Description
NTP-ICE	Integrated Chemical Environment (ICE), an open access database with results from NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM)
SkinSensDB	SkinSensDB is a collection of data from published literature to facilitate the development of AOP-based computational prediction methods(Wang et al. 2017)
ECHA-CHEM	European Chemicals Agency (ECHA) database is an open access database containing data for chemicals manufactured and imported in Europe. Although the summaries are publicly available, extracting data in large amounts requires special consideration as the studies are proprietary
TOXNET-HSDB	Hazardous Substances Data Bank (HSDB) is an open source database that provides information on human exposure to potentially hazardous chemicals
EURL-ECVAM-DB-ALM	The European Union Reference Laboratory for alternatives to animal testing database service on alternative methods to animal experimentation is an open access database, containing information on percutaneous absorption
CosIng	European Commission database of current and historical data for cosmetic substances and ingredients
RIFM	The Research Institute For Fragrance Materials (RIFM) monographs contain human health and toxicological data for fragrance and flavor raw materials.
Proprietary	Databases generated within a specific institution. Structure activity relationship (SAR) fingerprints
Literature	Manual curation of peer-reviewed articles and published training sets such as (MTD Cronin, 1994)

Table 2. Examples of deriving an overall assessment and confidence for the "skin sensitization *in vitro*" endpoint using the "AOP 2 out of 3" approach

Assessment and confidence scores			**Overall Assessment**	**Explanation**
Covalent Interaction with Skin Proteins	**Events in Keratinocytes**	**Events in Dendritic cells**	**Skin sensitization** ***in vitro***	
Positive, high confidence	Positive, high confidence	Negative, low confidence	Positive, high confidence	Positive, high confidence is the majority assessment
Negative, high confidence	Negative, high confidence	Positive, high confidence	Negative, medium confidence	Negative is the majority assessment, the confidence score is lowered based on a consideration of a third high confidence result
Positive, high confidence	Positive, medium confidence	Positive, low confidence	Positive, medium confidence	Positive is the majority assessment, the confidence score is medium based on three aligned calls with different confidence scores
Negative, high confidence	Negative, medium confidence	Positive, medium confidence	Negative, medium confidence	Negative is the majority assessment, the lower confidence score of the two aligned calls propagates to the overall assessment
Negative, Low confidence		Positive, Low confidence	Inconclusive	The assessments are non-concordant and the confidence scores are aligned

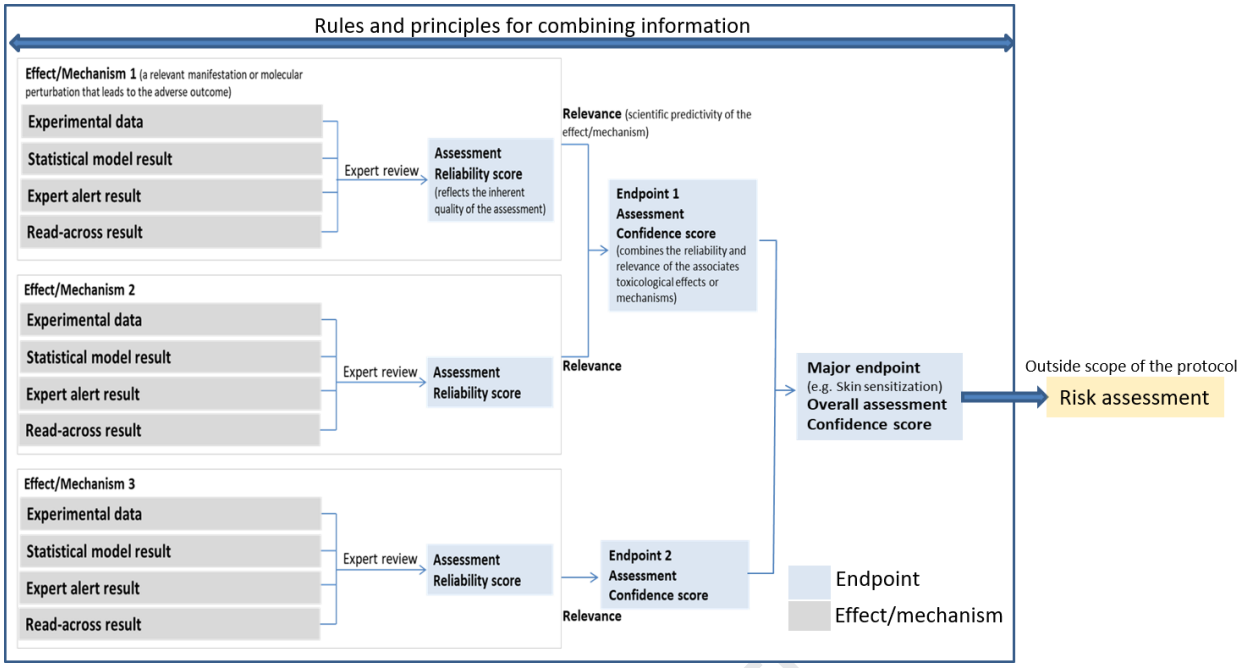
Table 3: Factors increasing and decreasing confidence in a human study (Schulz, Altman, and Moher 2010; Sibbald and Roland 1998)

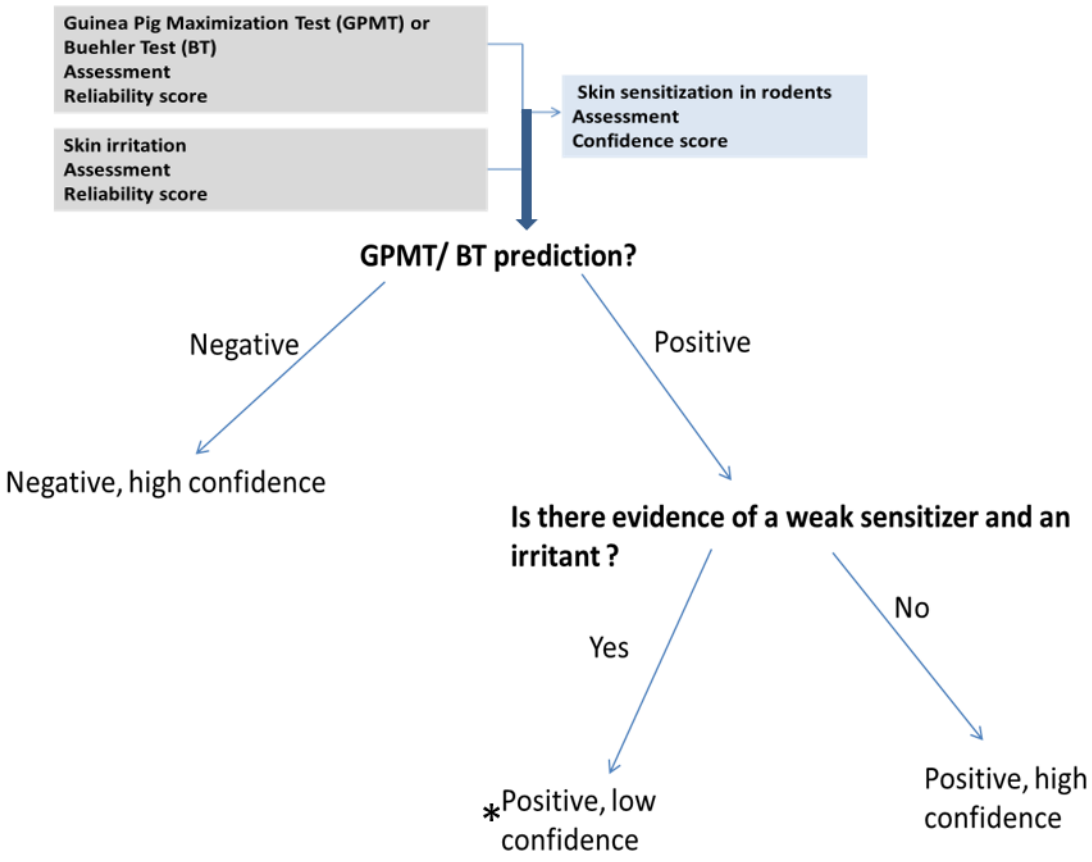
Factors increasing confidence	**Factors decreasing confidence**
Objective clearly stated and linked to measured outcome	Ambiguous objective, poorly linked to measured outcome
Randomized controlled study Randomized double-blind study	Uncontrolled and not randomized (or case report) No blinded control in study
Study conducted long enough to observe the effect	Study duration too short to observe the effect
Control substance application matches test substance application and represents the real-world exposure	Control substance application does not match test substance application or does not represents the real-world exposure scenario
Outcome clearly defined and measured through a quantitative endpoint	Subjective outcome based on perception
Statistical rationale behind determination of sample size	No rationale behind sample size selection
Description of study population available for review	No description of study population available

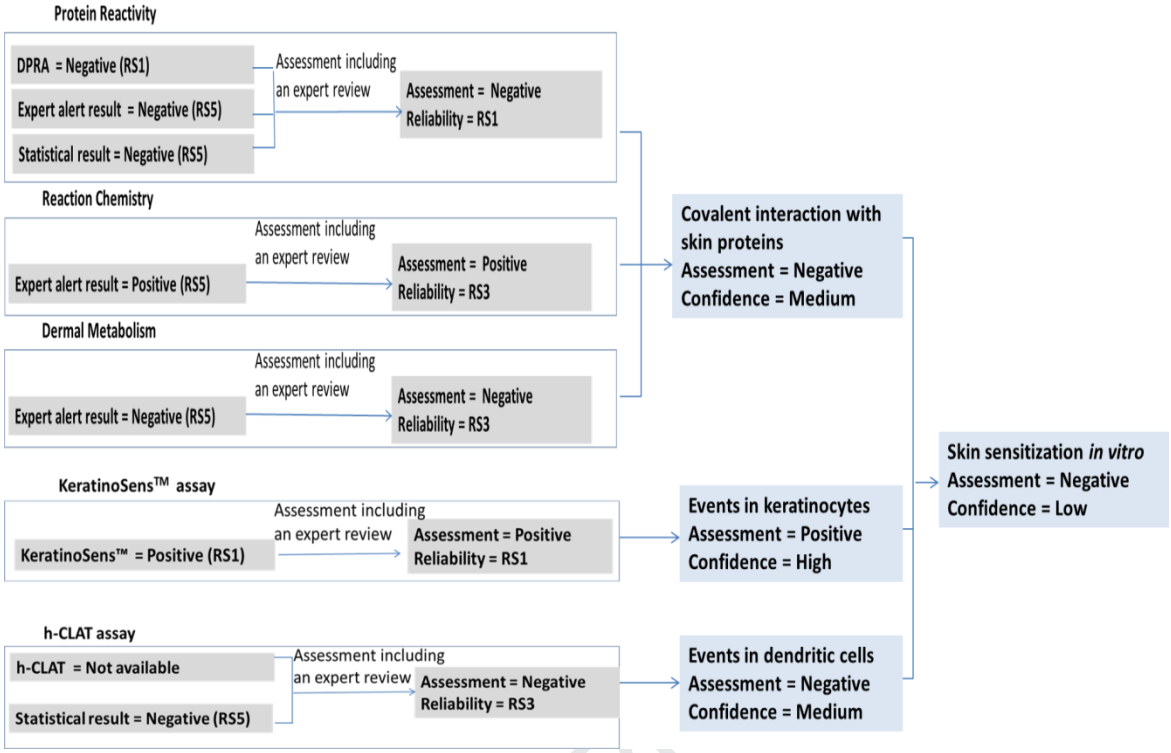
Table 4: Elements of an *in silico* toxicology report

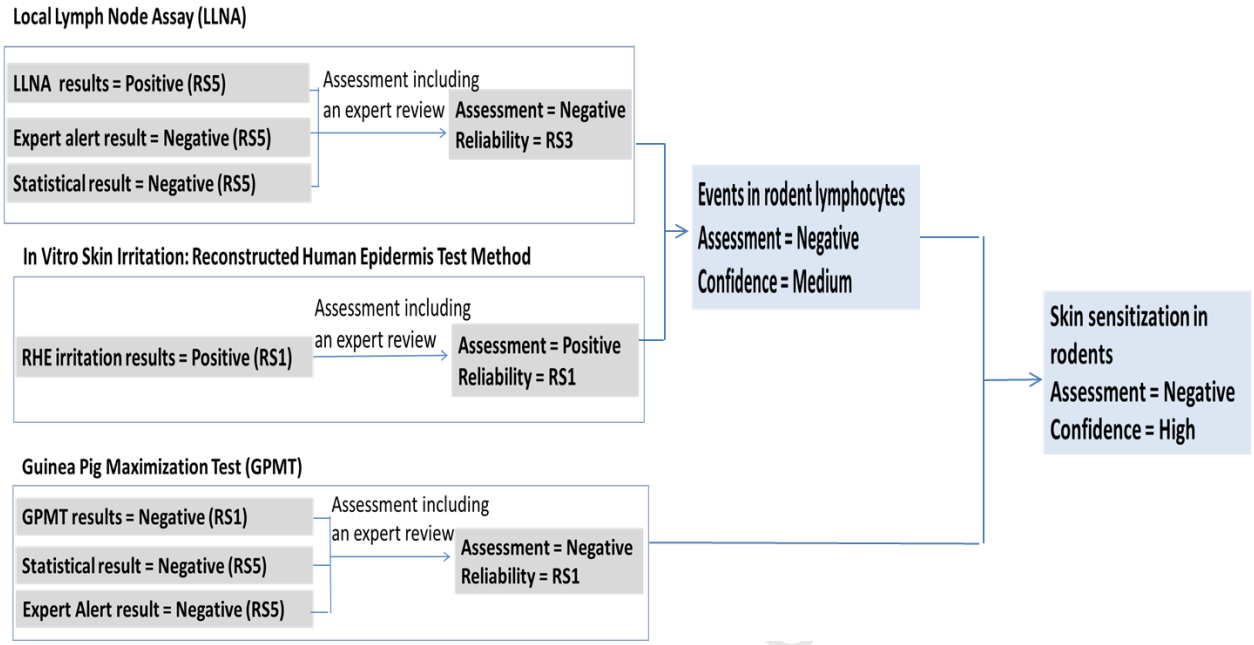
Section	Content
Title page	- Title (including information on the decision context) - Who generated the report and from which organization - Who performed the *in silico* analysis and/or expert review, including their organization - Date when this analysis was performed - Who the analysis was conducted for
Executive summary	- Provide a summary of the study - Describe the toxicity or properties being predicted - Include a table or summary showing the following: o The chemical(s) analyzed o Summary of *in silico* results, reviewed experimental data and overall assessment for each toxicological effect or mechanism o Summary of toxicological endpoint assessment and confidence o Summary of supporting information
Purpose	- Specification of the problem formulation
Materials and methods	- QSAR model(s), expert alerts, and other models used with version number(s) and any parameters set as part of the prediction (e.g., QMRF[1] format) - Databases searched with version number(s) - Description of any IATAs, DIPs, DAs used - Tools used as part of any read-across with version number(s)
Results of Analysis	- Details of the results and expert review of the *in silico* models and any experimental data, including results of the applicability domain analysis - Report of any read-across analysis, including source analogs and read-across justifications
Conclusion	- Summarize the overall analysis including experimental data, *in silico* methods and expert review - Final prediction that is based on expert judgment
References	- Complete bibliographic information or links to this information, including test guidelines referred to in the experimental data, etc.
Appendices (optional)	- Full (or summary) study reports used or links to the report, detailed (or summary) *in silico* reports, reports on the models used (e.g., QMRF reports)

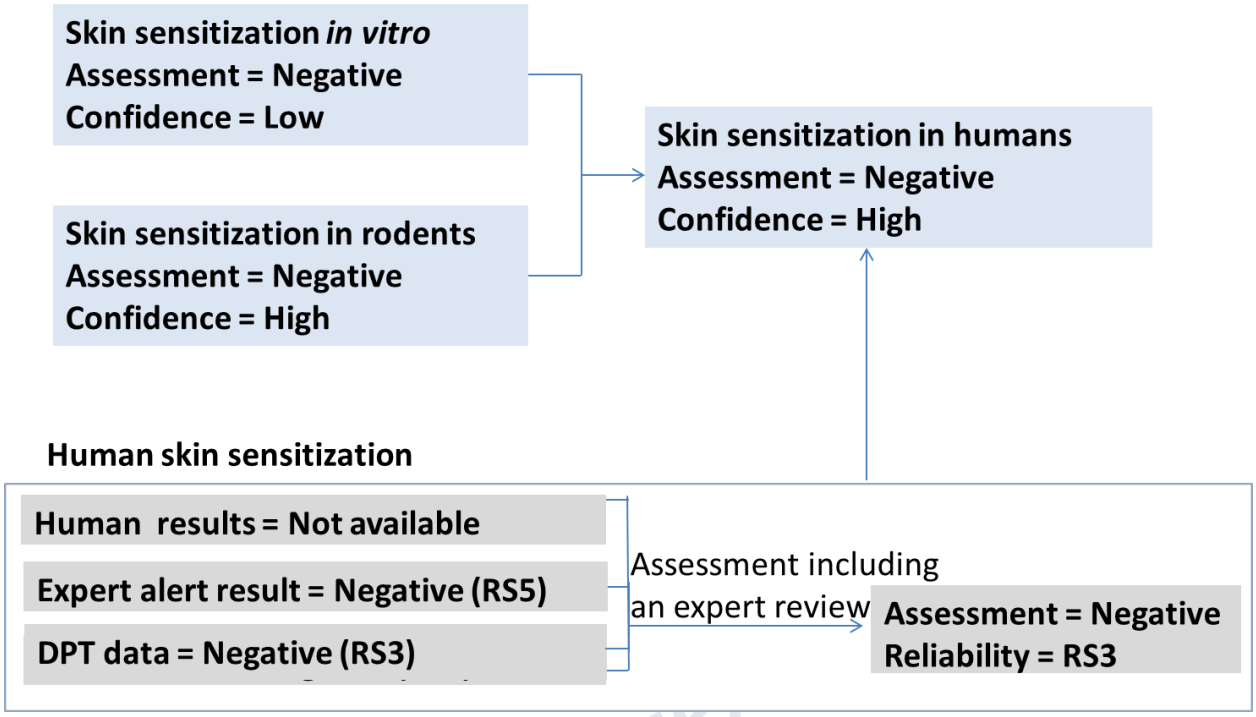
[1]QMRF – QSAR Model Reporting Format

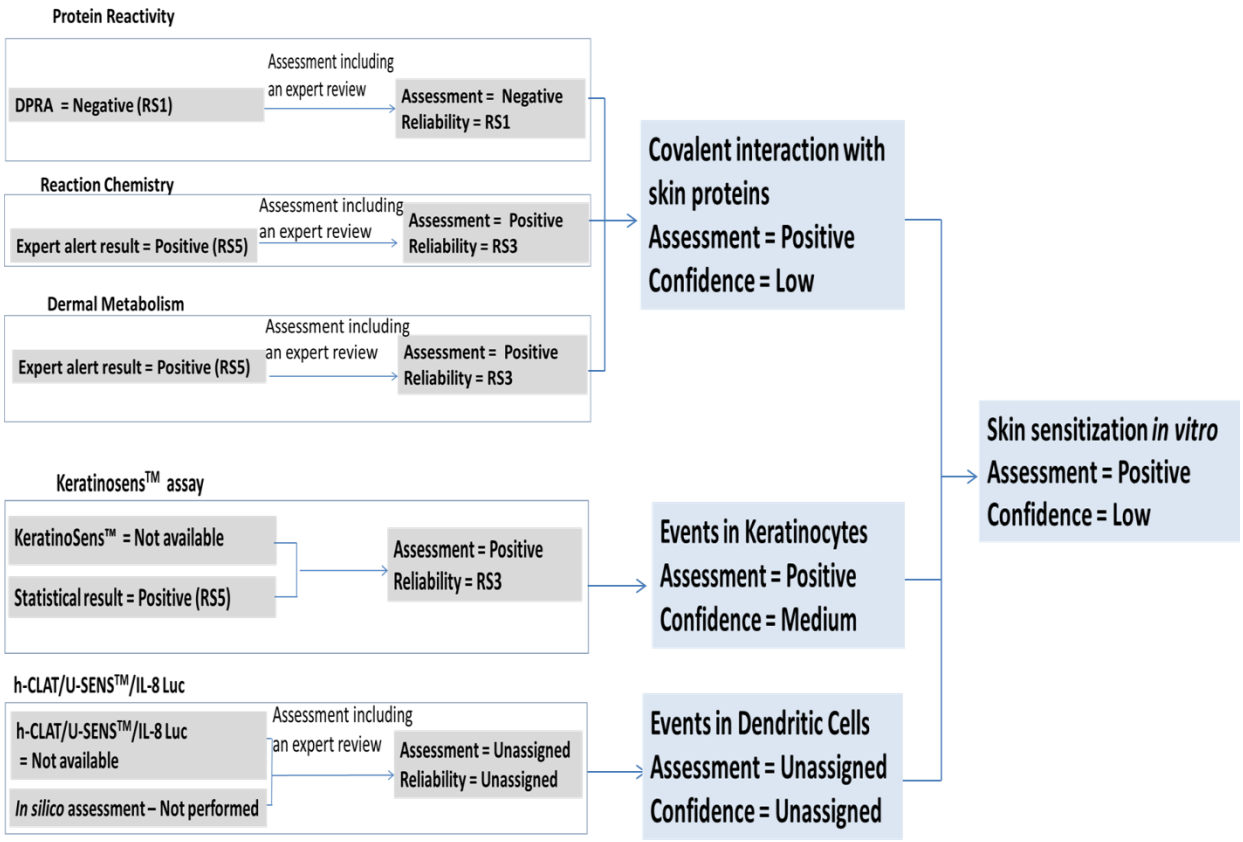


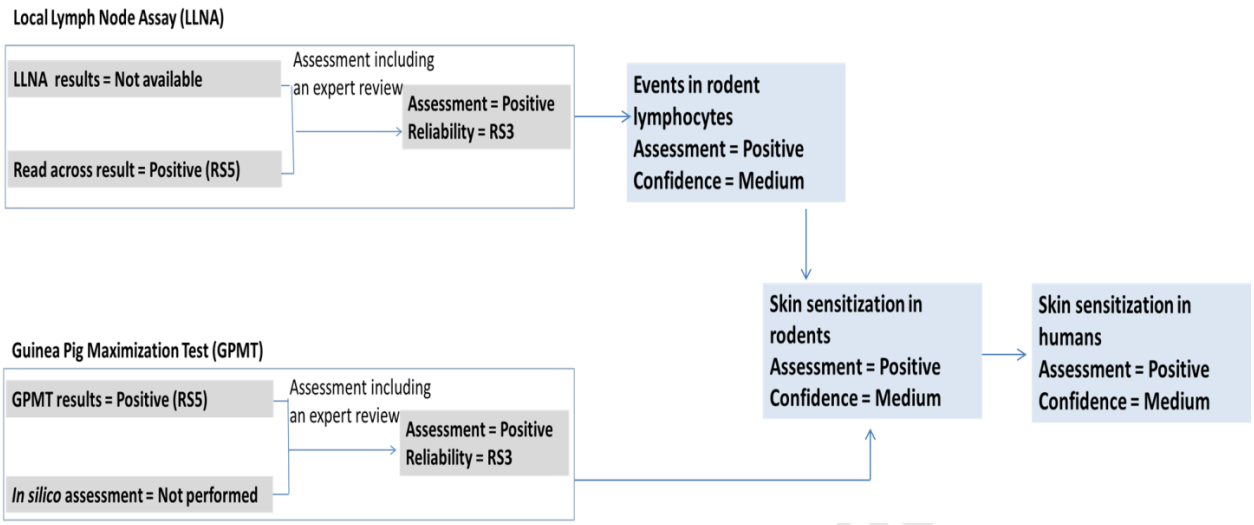

Guinea Pig Maximization Test (GPMT) or
Buehler Test (BT)
Assessment
Reliability score
Skin irritation
Assessment
Reliability score
Skin sensitization in rodents
Assessment
Confidence score
GPMT/ BT prediction?
Negative
Positive
Negative, high confidence
Is there evidence of a weak sensitizer and an irritant ?
Yes
No
*Positive, low confidence
Positive, high confidence

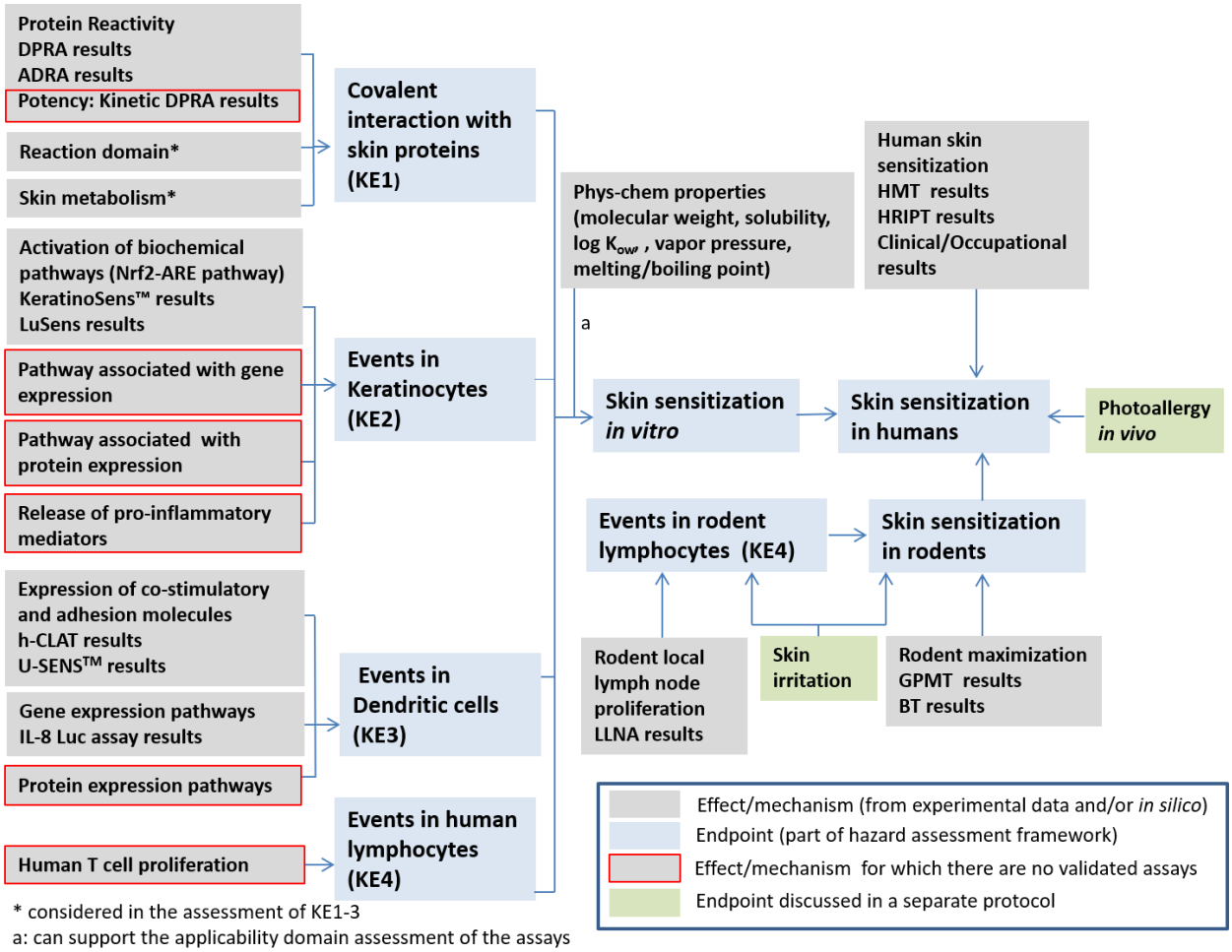




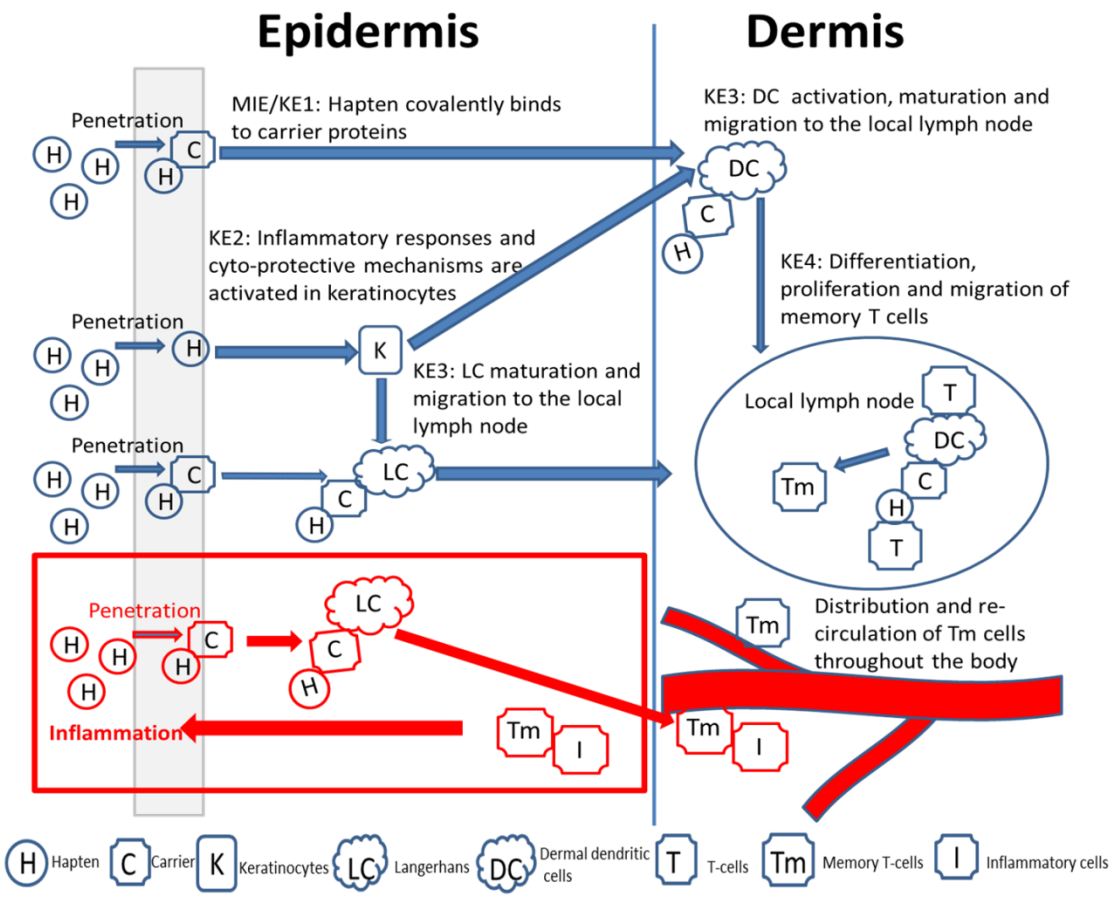


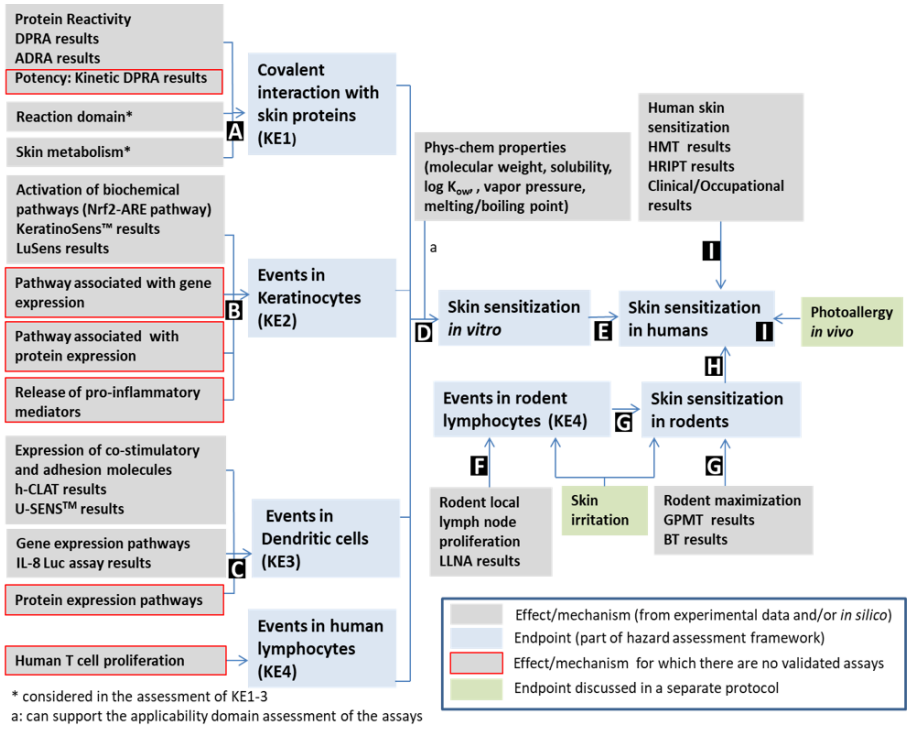

Protein Reactivity
DPRA = Negative (RS1)
Assessment including an expert review
Assessment = Negative
Reliability = RS1
Reaction Chemistry
Expert alert result = Positive (RS5)
Assessment including an expert review
Assessment = Positive
Reliability = RS3
Dermal Metabolism
Expert alert result = Positive (RS5)
Assessment including an expert review
Assessment = Positive
Reliability = RS3
Covalent interaction with skin proteins
Assessment = Positive
Confidence = Low
KeratinosensTM assay
KeratinoSens™ = Not available
Statistical result = Positive (RS5)
Assessment = Positive
Reliability = RS3
Events in Keratinocytes
Assessment = Positive
Confidence = Medium
h-CLAT/U-SENSTM/IL-8 Luc
h-CLAT/U-SENSTM/IL-8 Luc = Not available
In silico assessment – Not performed
Assessment including an expert review
Assessment = Unassigned
Reliability = Unassigned
Events in Dendritic Cells
Assessment = Unassigned
Confidence = Unassigned
Skin sensitization in vitro
Assessment = Positive
Confidence = Low





* considered in the assessment of KE1-3
a: can support the applicability domain assessment of the assays


Epidermis
Dermis
Penetration
MIE/KE1: Hapten covalently binds to carrier proteins
KE3: DC activation, maturation and migration to the local lymph node
DC
KE2: Inflammatory responses and cyto-protective mechanisms are activated in keratinocytes
KE4: Differentiation, proliferation and migration of memory T cells
K
KE3: LC maturation and migration to the local lymph node
Local lymph node
T
LC
Tm
Distribution and re-circulation of Tm cells throughout the body
Inflammation
I
Hapten
Carrier
Keratinocytes
Langerhans
Dermal dendritic cells
T-cells
Memory T-cells
Inflammatory cells



A Discussed in section 3.1

B Discussed in section 3.2

C Discussed in section 3.3

D Discussed in section 3.4

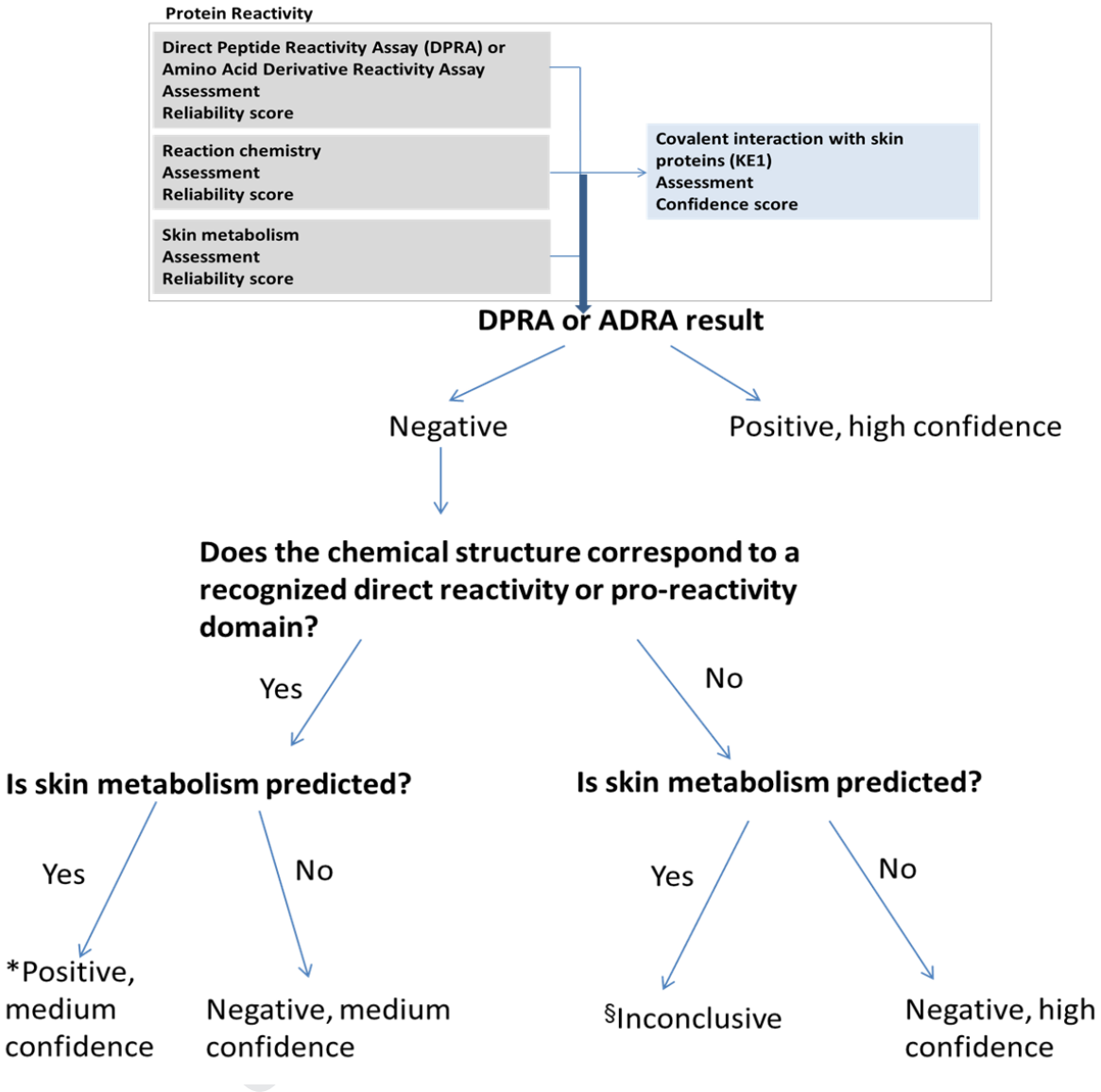
E Discussed in section 3.5

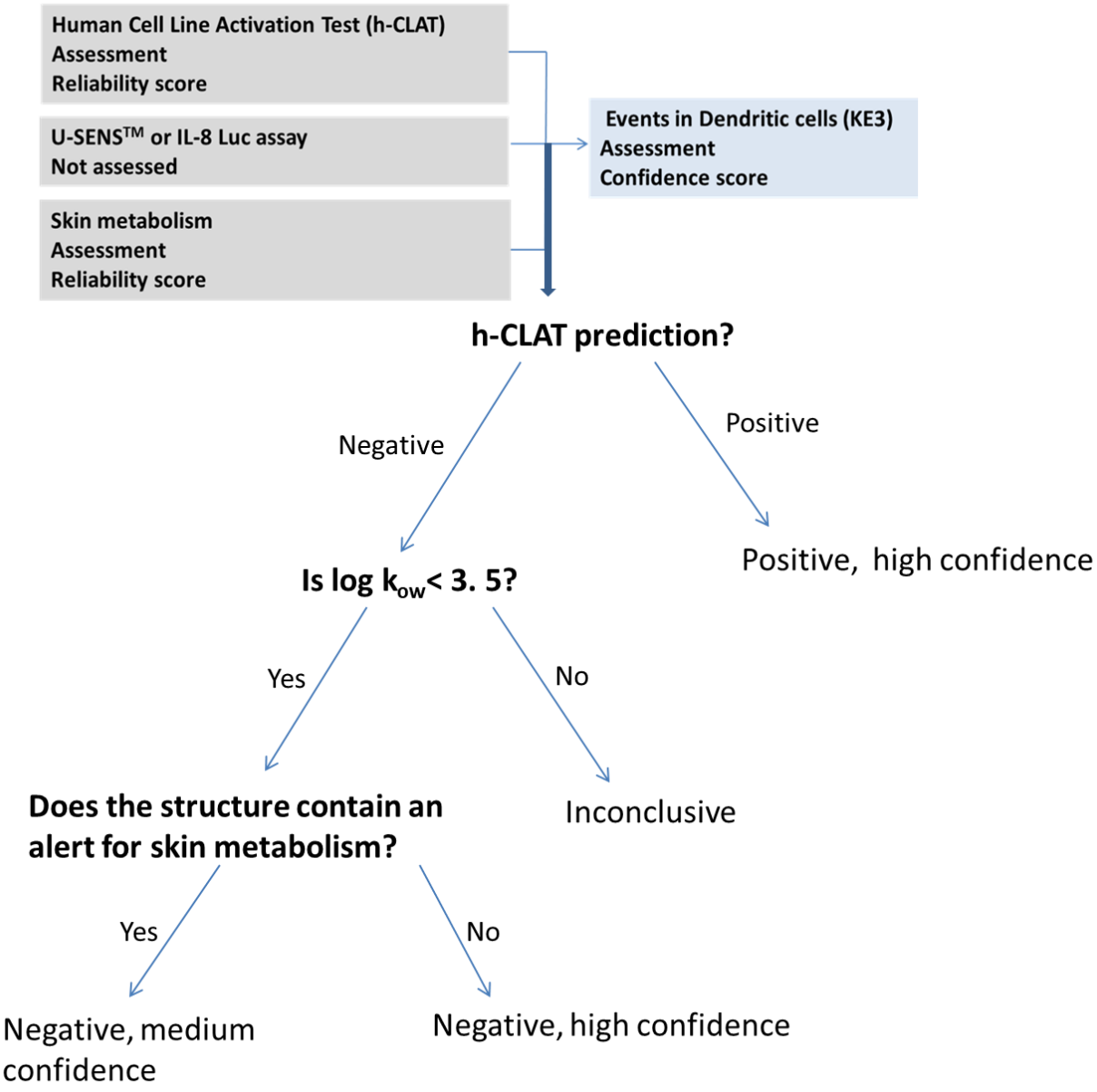
F Discussed in section 3.6

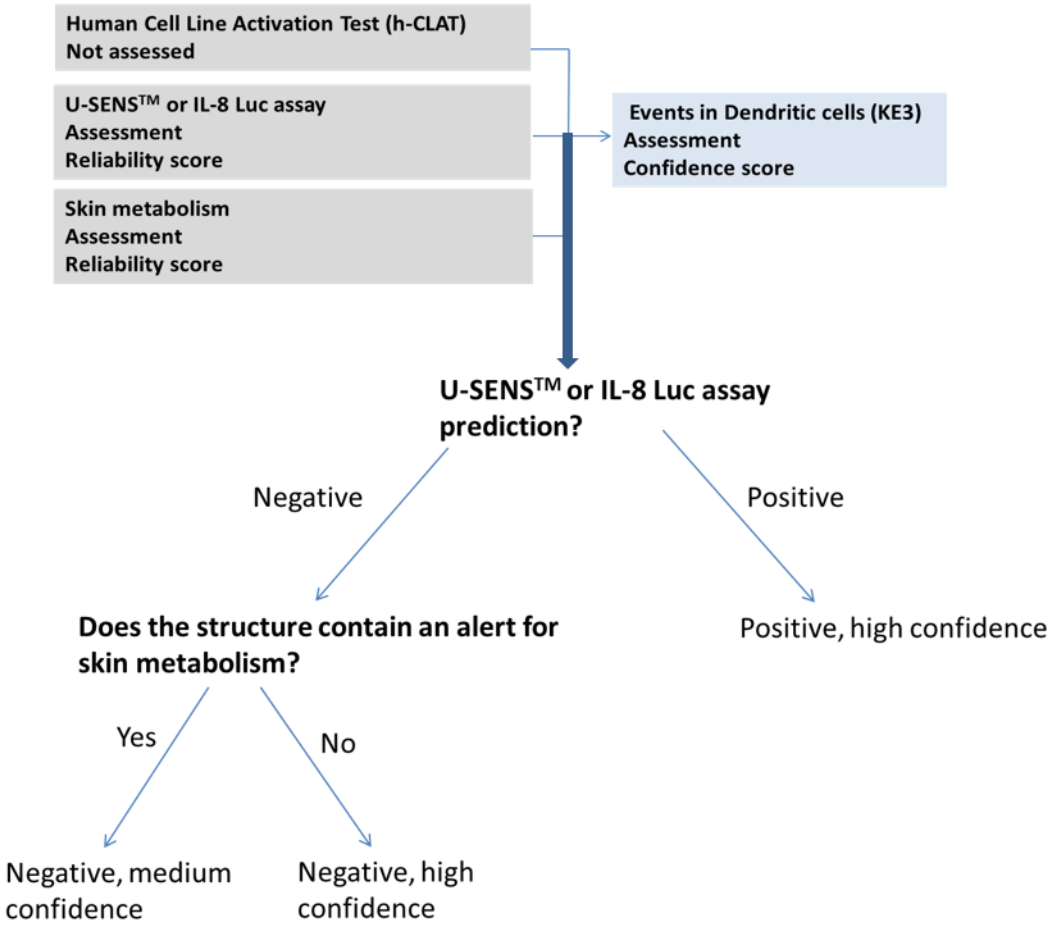
G Discussed in section 3.7

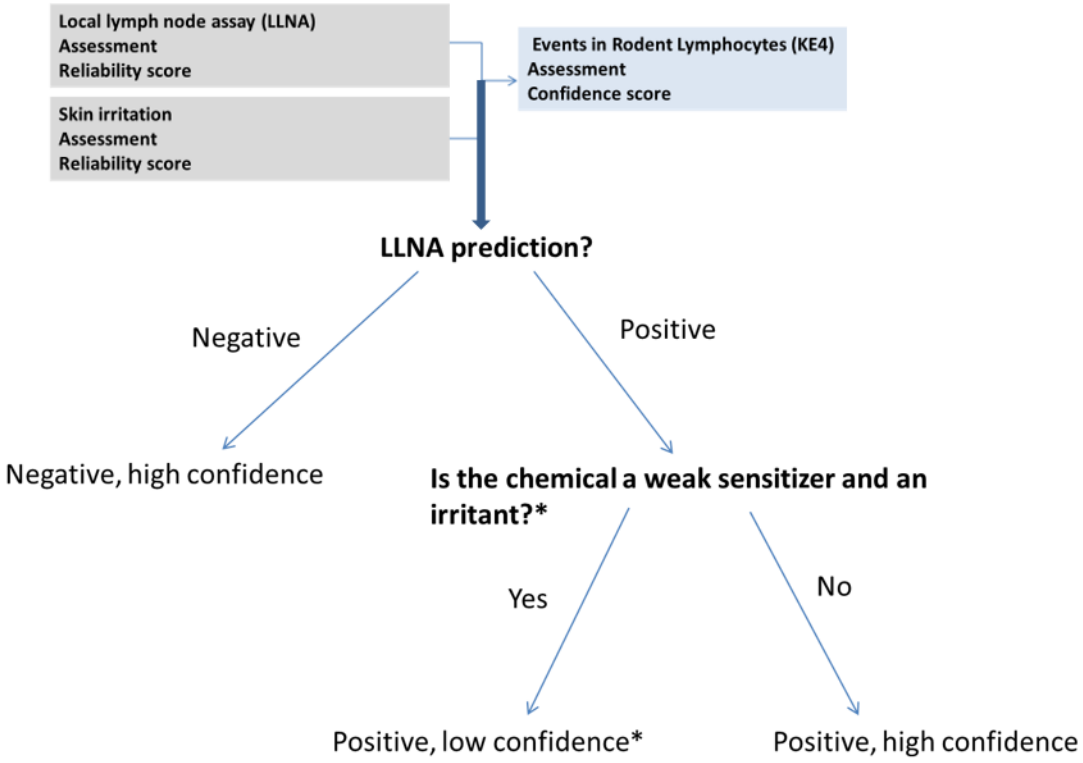
H Discussed in section 3.8

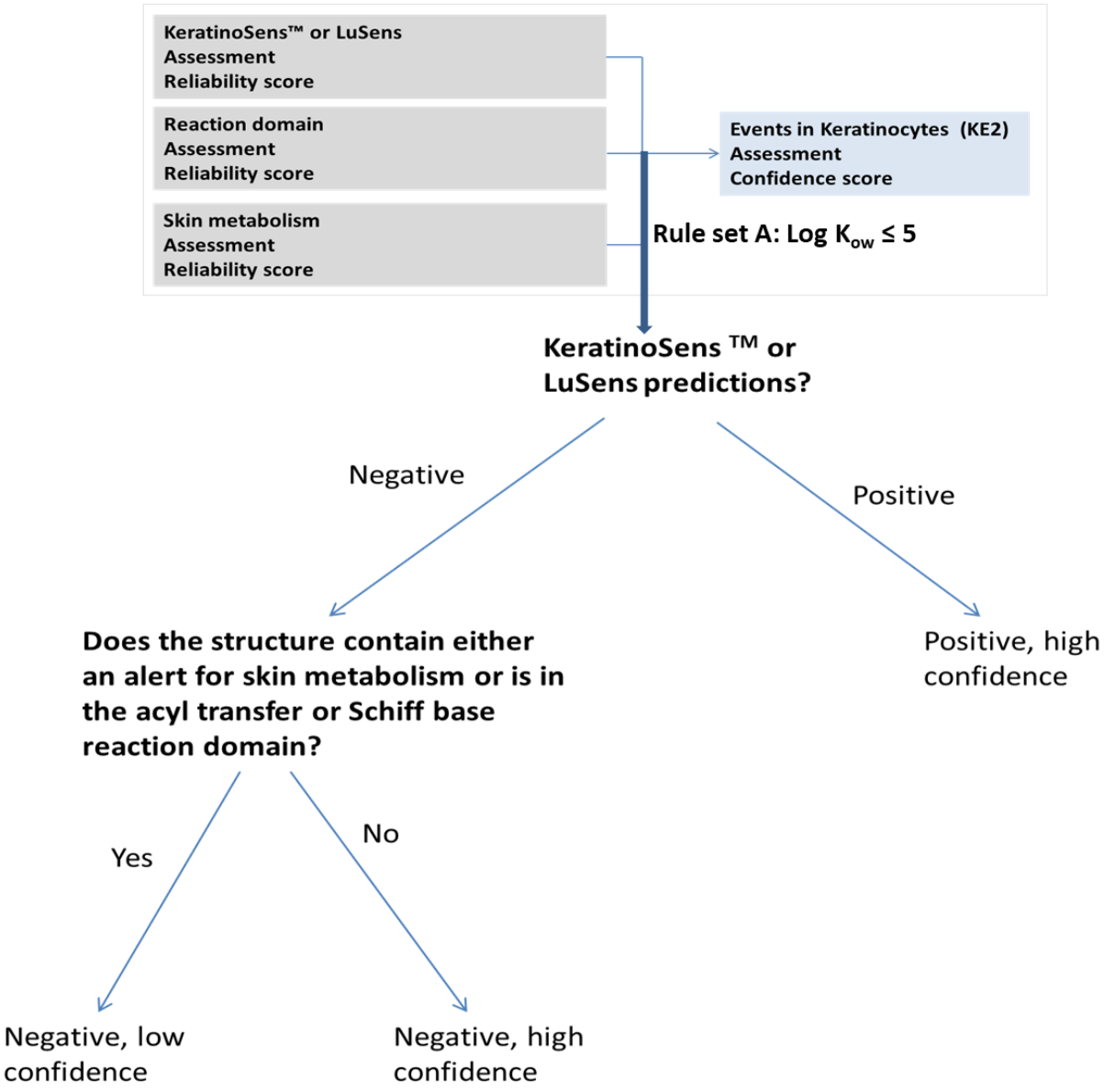
I Discussed in section 3.9

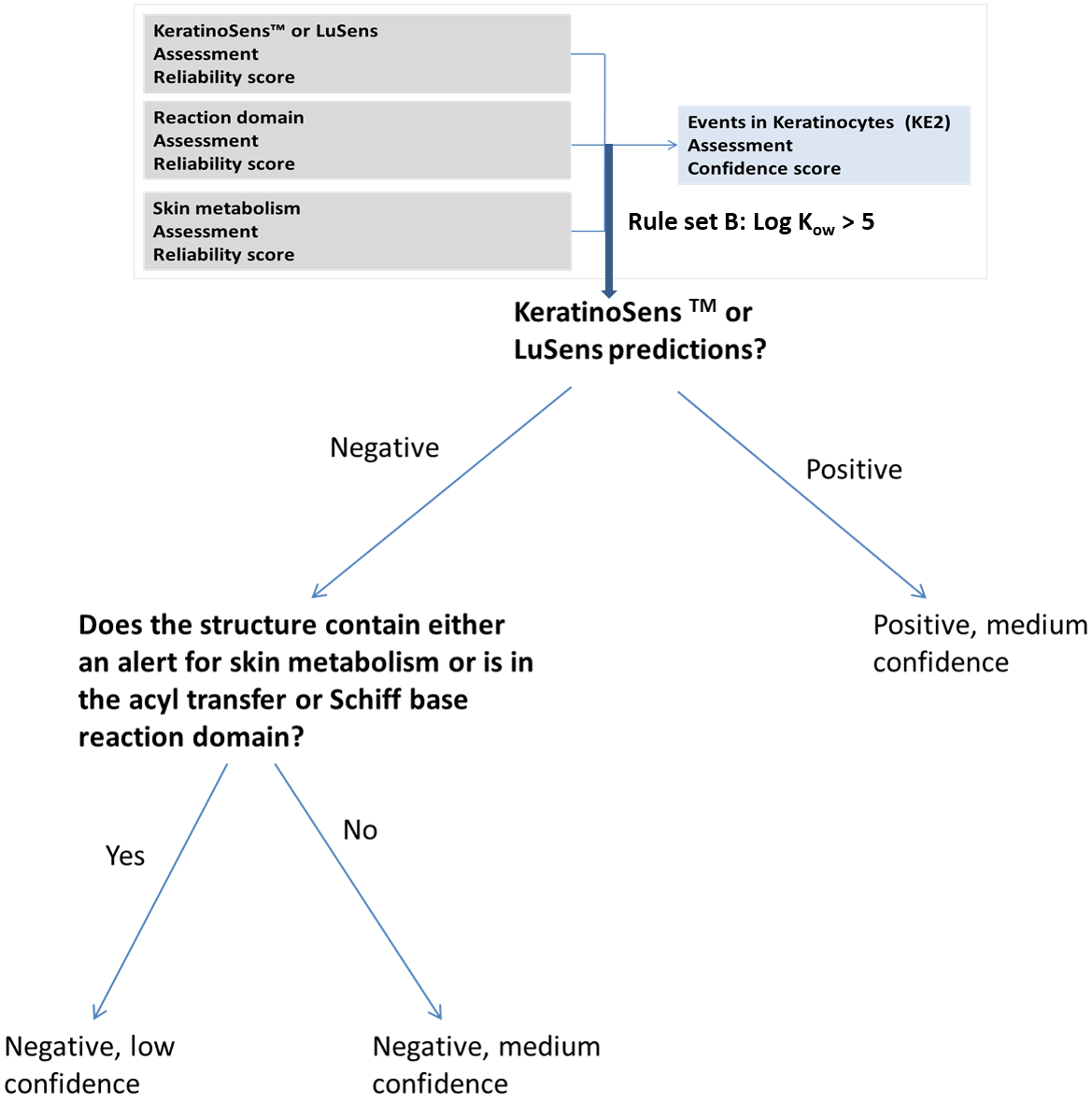

Protein Reactivity
Direct Peptide Reactivity Assay (DPRA) or Amino Acid Derivative Reactivity Assay
Assessment
Reliability score
Reaction chemistry
Assessment
Reliability score
Skin metabolism
Assessment
Reliability score
Covalent interaction with skin proteins (KE1)
Assessment
Confidence score
DPRA or ADRA result
Negative
Positive, high confidence
Does the chemical structure correspond to a recognized direct reactivity or pro-reactivity domain?
Yes
No
Is skin metabolism predicted?
Is skin metabolism predicted?
Yes
No
Yes
No
*Positive, medium confidence
Negative, medium confidence
§Inconclusive
Negative, high confidence


Human Cell Line Activation Test (h-CLAT)
Assessment
Reliability score
U-SENS™ or IL-8 Luc assay
Not assessed
Skin metabolism
Assessment
Reliability score
Events in Dendritic cells (KE3)
Assessment
Confidence score
h-CLAT prediction?
Negative
Positive
Positive, high confidence
Is log kow< 3. 5?
Yes
No
Inconclusive
Does the structure contain an alert for skin metabolism?
Yes
No
Negative, medium confidence
Negative, high confidence


Human Cell Line Activation Test (h-CLAT)
Not assessed
U-SENS™ or IL-8 Luc assay
Assessment
Reliability score
Skin metabolism
Assessment
Reliability score
Events in Dendritic cells (KE3)
Assessment
Confidence score
U-SENS™ or IL-8 Luc assay prediction?
Negative
Positive
Does the structure contain an alert for skin metabolism?
Positive, high confidence
Yes
No
Negative, medium confidence
Negative, high confidence




KeratinoSens™ or LuSens
Assessment
Reliability score
Reaction domain
Assessment
Reliability score
Skin metabolism
Assessment
Reliability score
Events in Keratinocytes (KE2)
Assessment
Confidence score
Rule set A: Log K_{ow} ≤ 5
KeratinoSens ™ or LuSens predictions?
Negative
Positive
Positive, high confidence
Does the structure contain either an alert for skin metabolism or is in the acyl transfer or Schiff base reaction domain?
Yes
No
Negative, low confidence
Negative, high confidence


KeratinoSens™ or LuSens
Assessment
Reliability score
Reaction domain
Assessment
Reliability score
Skin metabolism
Assessment
Reliability score
Events in Keratinocytes (KE2)
Assessment
Confidence score
Rule set B: Log K_{ow} > 5
KeratinoSens ™ or LuSens predictions?
Negative
Positive
Does the structure contain either an alert for skin metabolism or is in the acyl transfer or Schiff base reaction domain?
Positive, medium confidence
Yes
No
Negative, low confidence
Negative, medium confidence

Highlights

1. Details a hazard assessment framework for skin sensitization that includes experimental data and *in silico* results
2. Defines rules and principles for deriving an assessment from the available information
3. Outlines criteria to be considered as part of an expert review of an assessment
4. A method for assigning confidence to skin sensitization assessments is proposed

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

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: