Organisation for Economic Co-operation and Development

ENV/CBC/MONO(2021)11

Unclassified

English - Or. English

6 November 2023

ENVIRONMENT DIRECTORATE
CHEMICALS AND BIOTECHNOLOGY COMMITTEE

Cancels & replaces the same document of 23 August 2021

SUPPORTING DOCUMENT TO THE OECD GUIDELINE 497 ON DEFINED APPROACHES FOR SKIN SENSITISATION

**Series on Testing and Assessment,
No. 336**

JT03530938

SERIES ON TESTING AND ASSESSMENT
NO. 336

SUPPORTING DOCUMENT TO THE OECD GUIDELINE 497 ON DEFINED APPROACHES FOR SKIN SENSITISATION

IOMC
INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS
A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate
ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT
Paris 2021

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 36 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in eleven different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides**; **Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents;** and **Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

Also published in the Testing and Assessment link

For this and many other Environment, Health and Safety publications, consult the OECD's World Wide Web site (www.oecd.org/chemicalsafety/)

or contact:

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

This document is the Supporting Document for the OECD Guideline on Defined Approaches on Skin Sensitisation, published as Guideline No. 497. It provides detailed information on the identification and curation of robust reference chemicals, predictive performance and uncertainty in the DAs and their individual data information sources.

This Supporting Document and its annexes were used as a basis to develop the Guideline on Defined Approaches for Skin Sensitisation. The Supporting Document and its annexes were prepared by the countries and institutes leading the project: United States, the European Commission Joint Research Centre, and Health Canada, with input and review from various members of the dedicated Expert Group on Defined Approaches on Skin Sensitisation. The Supporting Document was endorsed in April 2021 by the Working Party of the National Coordinators of the Test Guidelines Programme, and is published under the responsibility of the Chemicals and Biotechnology Committee.

Table of contents

Tables

Figures

1. Introduction

1. Project 4.116 to develop a Guideline (GL) on Defined Approaches for Skin Sensitisation was added to the OECD Test Guidelines work plan. Following a special session of the WNT in December 2017, an Expert Group on Defined Approaches for Skin Sensitisation (EG DASS), including experts in Integrated Approaches to Testing and Assessment (IATAs), Defined Approaches (DAs), quantitative analyses, Quantitative Structure Activity Relationships (QSARs), and skin sensitisation, was convened in early 2018. A joint session of the OECD IATA Case Studies Project and the OECD QSAR Toolbox Management Group was held in November 2018 to discuss issues related to QSAR data in Defined Approach Guidelines that would be covered under the Mutual Acceptance of Data (MAD). An evaluation framework was developed at the request of the WNT during the special session in December 2017, and subsequently agreed upon and considered by the expert group for the evaluation of the DAs included in the GL (**Annex 1**, Evaluation Framework).

2. A first set of three relatively simple, rule-based DAs using validated OECD *in chemico* and *in vitro* test methods and *in silico* predictions is included in the GL and supporting information relative to these DAs is presented in this document. Some of the information included in this document has been published previously in OECD Guidance Documents or in scientific journals.

3. The three DAs described in the GL and in this supporting document are:

- The 2 out of 3 (2o3) defined approach to skin sensitization hazard identification based on *in chemico* (KE 1) and *in vitro* (KE 2/3) data (1, 2).
- The integrated testing strategy (ITSv1) for UN GHS potency categorisation based on *in chemico* (KE 1), *in vitro* (KE 3), and *in silico* (Derek Nexus) data (3, 4), with an updated data interpretation procedure (DIP) based on expert group recommendation.
- A modification of the integrated testing strategy (ITSv2) for UN GHS skin potency categorisation based on *in chemico* (KE 1), *in vitro* (KE 3), and *in silico* (OECD QSAR Toolbox) data, with an updated DIP based on the expert group recommendation.

4. These three DAs have been shown to provide similar or superior predictive capacity to the murine Local Lymph Node Assay (LLNA) (5) for hazard identification (*i.e.* sensitiser versus non-sensitiser) when compared to human reference data. In addition, the ITSv1 and ITSv2 DAs can be used to discriminate chemicals into potency sub-categories according to the UN Globally Harmonized System of Classification and Labelling of Chemicals (UN GHS; Category 1A = strong sensitisers; Category 1B = other sensitisers, and No Categorization (NC = not classified) (6), where the potency sub-categorisation performance was also comparable to or exceeded that of the LLNA when compared to human reference data.

5. The selection criteria for evaluation and inclusion in the GL of these initial DAs were 1) *in vitro* methods used in the DA were validated OECD Test Guideline methods, and 2) the DA relied on relatively simple, rule-based DIP.

6. It has to be noted that in some cases, the OECD Key Event Based Test Guidelines (TG 442C, 442D, and 442E) (6, 7, 8) include other methodologically and functionally similar assays (*i.e.* “me too” methods) or assays that address the same key event endpoint. However, only the test methods used in the DAs mentioned above, explicitly stated in this supporting document and associated GL, are considered. Though the other validated *in vitro* methods may perform equally, it will be the responsibility of the test method developers to demonstrate the suitability and validity of the replacement *in vitro* method in the DAs.

7. The DAs described in this supporting document were originally described in Annex I: Case studies to the 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 Series on Testing & Assessment Guidance Document 256 (10). In GD 256, the performance of the DAs as evaluated by the DAs developers is reported. Prior to initiating expert review of the DAs, the DA developers were given the opportunity to update the DAs in advance of this evaluation.

8. It should be noted that other DAs are available for predicting skin sensitisation hazard and potency, or for determining a point of departure for quantitative risk assessment, and though they are not presently included in the GL they may be added at a later stage following review and approval.

9. Based on the works in institutes and agencies from the leads and member countries, and on recommendations from the OECD EG DASS from 2017 to 2020, information on the reference data, applicability, limitations and uncertainties, and predictive performance of the individual DAs has been updated in the GL and this supporting document to include comparisons with Local Lymph Node Assay (LLNA) and HPPT reference data (see **Section 5** for performance summaries and **Annex** 2, Reference data matrix and comparisons, for detailed comparisons). The LLNA and human reference data have been extensively curated, and the reference chemical lists were refined based on exclusion/inclusion criteria agreed upon by the EG (**Annexes 3** and **4**).

10. Based on suggestions from the EG DASS, the ITS DA was updated to predict potency categories for skin sensitisation used by UN GHS, rather than the ECETOC skin sensitisation categories (11) used in the original DA (3, 4).

11. In addition, upon request by the EG DASS, a second version of the ITS was revised to include OECD QSAR Toolbox (referred to as OECD QSAR TB hereafter) predictions of skin sensitisation potential based on structural analogues in place of structural alerts resulting from the proprietary software Derek Nexus (referred to as Derek hereafter); this version is referred to as ITSv2. In addition, the OECD EG recommended, for both the ITSv1 (using Derek) and the ITSv2 (using OECD QSAR TB), using a score of 6 instead of the original score of 7 to better identify UN GHS 1A (strong) sensitizers, and to extend the applicability of the ITSv1 and ITSv2 to chemicals for which *in silico* predictions cannot be generated. Background information is consolidated herein to support the scientific validity of the DAs included in the OECD GL.

12. Sub-Groups were formed under the EG DASS to progress more efficiently on specific issues identified:

- Sub-Group on LLNA data curation
- Sub-Group on human data curation
- Sub-Group on Applicability domain definition
- Sub-Group on Uncertainty analysis

2. *In vivo* murine and human reference classifications

13. The performance assessment of Defined Approaches (DAs) was evaluated against highly curated sets of *in vivo* reference data. To this end Sub-Groups were formed under the EG DASS charged with the curation of 1) LLNA data (EG DASS LLNA sub-group) and 2) human data (human data sub-group) to agree upon the final *in vivo* reference classifications for the sets of reference chemicals used in the evaluation of the DAs.

2.1. LLNA reference data

14. To review LLNA studies, the EG DASS LLNA sub-group proposed criteria for including/excluding LLNA results in the reference database (See **Annex 3**, Report of the LLNA sub-group on the curation and evaluation of the LLNA reference data for further details). These criteria are based on the essential test method components from the LLNA Performance Standards for the validation of modifications to the traditional LLNA as described in OECD TG 429 (5). According to these criteria: 1) the test chemical must be applied topically to both ears of the mice, 2) lymphocyte proliferation must be measured during the induction phase of skin sensitisation and in the lymph nodes draining the site of test chemical application, 3) a vehicle control must be included in each study, 4) either individual or pooled animal data may be collected.

15. Additional inclusion criteria are based 1) on the availability of concentrations tested and corresponding stimulation index (SI)1 values, 2) *in vivo* administration of 3H-methyl thymidine or other radiolabelled markers, 3) sodium dodecyl sulphate (SDS) was not applied as pre-treatment, and 4) exclusion of studies performed with natural extracts because of variable and/or ill-defined composition.

16. Furthermore, the EG DASS LLNA sub-group provided the EG DASS with a proposal for how to interpret individual LLNA results with an Estimated Concentration three (EC3)2 outside the measured dose range (*i.e.* extrapolated EC3 values). The criteria used for this purpose were consistently applied to the relevant LLNA studies to determine the reliability of the extrapolated EC3 and its suitability for UN GHS sub-categorisation (UN GHS sub-category 1A vs. sub-category 1B).

17. Considerations were also applied on how to handle negative LLNA results for chemicals not tested up to a concentration of 100% and without a documented scientific rationale for the maximum test concentration selected. The EG DASS LLNA sub-group accepted study results as negative if for all test concentrations Stimulation Indexes (SI) values were < 3 and if the test chemical was evaluated up to a highest concentration tested of at least 50%. Negative study results were also accepted when a valid scientific reason (in accordance with TG 429) (5) was provided for why the highest tested concentration was lower than 50%.

18. When multiple LLNA test results were available for the same chemical, the EG DASS LLNA sub-group decided to use the Median Like Location Parameter (MLLP) approach to

1 Stimulation Index (SI): A value calculated to assess the skin sensitization potential of a test chemical that is the ratio of the lymphocyte proliferation in treated groups to that in the concurrent vehicle control group.

2 Estimated concentration three (EC3): Estimated concentration of a test chemical needed to produce a stimulation index of three.

derive an overall reference classification (see **Annex 3**, Report of the LLNA sub-group on the curation and evaluation of the LLNA reference data for further details).

19. By applying these criteria an unambiguous classification could be obtained for 168 (85.7%) of the chemicals, of which 135 (80.4%) were classified as sensitisers (UN GHS Skin Sens. 1) and 33 (19.6%) were no classification (NC). See **Annex 3, Table 2.1**., GHSBIN

20. For 123 chemicals with an unambiguous classification as skin sensitisers, also a UN GHS sub-categorisation could be obtained. Of these, 38 (30.9%) were classified as Skin Sens. 1A and 85 (69.1%) as Skin Sens. 1B. No classification was obtained for 33 chemicals. See **Annex 3, Table 2.1**, GHSSUB

21. DA performance was assessed based on the 168 chemicals with unambiguous binary reference classifications and the 156 chemicals with unambiguous potency sub-categorization. Additional logic was applied to identify chemicals with borderline classifications. Further details on the borderline considerations (**Table 1.1**, GHSBORDER) may be found in **Annex 3**, Report of the LLNA sub-group on the curation and evaluation of the LLNA reference data.

Table 2.1. Distribution of the LLNA reference classifications over the UN GHS hazard category/sub-categories (N = 194)

Mode	GHS class/sub-category				
	1			NC/1B	NC
	1A	1	1B		
GHS_{BIN}	135			na	33
GHS_{SUB}	38	na	85		
GHS_{BORDER}	34	20	78	31	31

22. In addition, another assessment of the LLNA data was performed. In this analysis different methods of combining multiple test results into an overall classification, the Median Sensitisation Potency Estimate (MSPE), and the Weight of Evidence (WoE) score methods were used and were compared to the standard approach applied (MLLP) (See **Annex 3,** Report of the LLNA sub-group on the curation and evaluation of the LLNA reference data for further details). Where, for a given chemical, these results differed from each other, it was concluded that the respective reference classification was uncertain or borderline, requiring further discussion. In each of these cases, an overall classification based on rule-based expert judgement was proposed. The analysis also for the first time introduced a proposal for how to address borderline classifications in general.

23. Previous analyses have suggested a low reproducibility of LLNA test results (*i.e.* EC3 values). Nevertheless, an analysis performed by the EG DASS LLNA sub-group showed that LLNA-derived classifications are highly reproducible when quality criteria as defined by the EG DASS were applied to the individual test results. In all cases where the reproducibility of classifications was clearly < 100%, this could be traced back to the fact that either the respective chemicals were borderline cases (*i.e.* their potency was close to the 1A/1B border or they were presumed to be sensitising at high concentrations only) or nominal classification based on the rules applied for assessment was overruled by the EG DASS based on expert knowledge. These conclusions, however, are based on a comparatively small number of chemicals with multiple test results and should therefore be considered preliminary, see **Annex 3**, Report of the LLNA sub-group on the curation and evaluation of the LLNA reference data for further details.

2.2. Human reference data

24. Previous work provided a preliminary demonstration that many of the DAs under OECD consideration had superior performance to the LLNA when compared to expert derived

human classifications (12). To judge the performance of these DAs, the EG DASS recognised the need for an in-depth analysis of the reliability and robustness of human data consistent with what was done for the LLNA data for assessing the performance of the DAs.

25. The human data sub-group, (referred to hereafter as HDSG) was composed of scientists from various risk assessment institutions from the United States and Europe and was assisted by scientists from the Research Institute for Fragrance Materials (RIFM). For the construction and review of the available Human Predictive Patch Test (HPPT) database, the following specific tasks were undertaken:

- curation of the HPPT skin sensitisation database,
- analysis of the variability and uncertainty in the HPPT database,
- development of a framework for using HPPT data to classify chemicals with respect to their skin sensitisation potential based on the UN GHS, and
- classification, when possible, of the reference chemicals used by the OECD EG DASS for DA performance assessment with respect to their skin sensitisation potential.

26. To that end, a database of 2277 HPPT test results referenced in more than 1700 publications from the late 1950s until and including December 2019 was compiled and investigated. This database was built starting from a database collated previously at the United States National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM). Most of these references used were “monographs” on fragrance ingredients, published by the Research Institute for Fragrance Materials (RIFM) since the early 1970s in the journal Food and Chemical Toxicology. Additional references from the published literature were also included.

27. Human diagnostic patch test data, which, in principle, can also be used for classification under the UN GHS, were not included due to considerations reported in Section 2 of the HDSG report (see **Annex 4**, Report of the Human Data Sub-Group on the Curation and Evaluation of Human Reference Data for further details).

28. The database contains information on chemical identity, test design, and test results and provides bibliographic information on the (mostly unpublished) original test reports as well as on later publications citing these primary sources. A system for categorising the HPPT results with respect to their relative reliability has been developed; 2255 test results with acceptable relative reliability scores (RRS < 5) were taken forward for further analyses. These activities are reported in detail in Section 3 of the HDSG report (**Annex 4**), while Section 4 of the HDSG report provides an overview of the database including basic descriptive statistics.

29. In Section 5 of the HDSG report (**Annex 4**), the multiple factors potentially affecting HPPT data variability and uncertainty are qualitatively discussed in some detail. For a more quantitative assessment, either the necessary data do not exist or more work is needed, which is beyond the scope of the current phase of the OECD project on the DA guideline. When compared with the respective UN GHS classification criteria, the vast majority (2188, or 97%) of the available HPPT results with sufficient reliability did not translate into an unambiguous classification, including sub-categorisation, *i.e.* UN GHS Skin Sens. 1A, 1B, or No Classification (NC). About 73% (1642) of the test results did not allow for an unambiguous binary classification decision (Skin Sens. 1 vs. NC). This is mainly due to the fact that HPPT are normally only performed at one dose level which is often either too high or too low to reliably allocate sensitisers into sub-categories based on potency. As a consequence, the HDSG developed concepts and scoring methods to decide on sub-categorisation of the available data, which are reported in detail in Section 6 of the HDSG report (**Annex 4**).

30. Section 7 of the report describes how the individual test results were translated into "extrapolated" classifications by applying the concepts developed in the previous sections to the 453 test results available for 104 chemicals in the EG DASS reference chemical list.

Furthermore, it explains how, in the case of discordant individual test results, the overall classification was obtained by means of a weight-of-evidence (WoE) assessment.

31. In Section 8.2, the reproducibility of HPPT-based reference classifications was analysed and characterised quantitatively. Although the results are based on only 14 - 34 chemicals with multiple test results (and therefore conclusions should be drawn with care), HPPT-based UN GHS reference classifications were found to be highly reproducible on average.

32. Finally, in the HDSG report, the resulting reference classifications are provided in an overview table. Human data suitable for classification (albeit not always unambiguous) could be obtained for 104 of the 196 EG DASS reference chemicals. Of these, the HDSG could conclude on the hazard classification of 66 (63.5%) chemicals.

33. Of the 66 chemicals with a hazard classification, 55 were classified as sensitisers and 11 were not classified (considered to be non-sensitisers) (See Table 2.2, GHSBIN). Potency sub-categorisation could be obtained for 63 chemicals, 21 were classified as UN GHS 1A, 31 as UN GHS 1B, and 11 as not classified (NC) (see Table 2.2, GHSSUB).

34. DA performance was assessed based on the 66 chemicals with unambiguous binary reference classifications and the 63 chemicals with unambiguous potency sub-categorization. For some chemicals, binary classification or sub-categorisation were associated with significant uncertainty. This was expressed by introducing a third classification "GHSBORDER" to indicate sensitisers with significant uncertainty about their sub-categorisation (GHSBORDER = 1) (N=55) or uncertainty about their classification as sensitisers (GHSBORDER = NC/1B) (N=38). Further details on borderline considerations (Table 2.2, GHSBORDER) may be found in **Annex 4,** Report of the Human Data Sub-Group on the Curation and Evaluation of Human Reference Data.

Table 2.2. Distribution of HPPT reference classifications over the UN GHS hazard category/sub-

Mode	GHS class/sub-category				
	1			NC/1B	NC
	1A	1	1B		
GHS_{BIN}	55			na	11
GHS_{SUB}	21	na	31		
GHS_{BORDER}	14	12	29	38	

35. An analysis of the reproducibility of human data was performed on the basis of multiple studies available for individual chemicals and the outcome showed that human data are highly reproducible (90-100%) when quality criteria as defined by the EG DASS were applied. In all cases where the reproducibility of classifications was clearly < 100%, this could be traced back to the fact that either the respective chemicals were borderline cases (*i.e.* their potency was close to the 1A/1B border or they were presumed to be sensitising at high concentrations only) or nominal classification based on the rules applied for assessment was overruled by the EG DASS based on expert knowledge. These conclusions, however, are based on a comparatively small number of chemicals with multiple test results and should therefore be considered preliminary, see **Annex 4**, Report of the Human Data Sub-Group on the Curation and Evaluation of Human Reference Data for further details.

3. Sources of uncertainty: *in chemico*, *in vitro*, *in silico*, DA predictions

3.1. Impact of Log P on the performance of *in chemicolin vitro* assays and ITSv1, ITSv2 and 2o3 Defined Approaches for Skin Sensitisation

36. Previous published analyses (13) indicated a marked reduction in sensitivity of the h-CLAT vs. LLNA data for 31 (27 sensitisers, 4 non-sensitisers) chemicals with Log P > 3.5 compared to 143 chemicals with Log P $\leq$ 3.5. This study provided the basis for inclusion of a provision in OECD TG 442E on the h-CLAT as stand-alone method that "negative results with test chemicals with Log P greater than 3.5 should not be considered".

37. Based on this observation, an analysis was conducted as a joint effort between the United Kingdom and Denmark to investigate the impact of Log P on the results from *in chemico*, *in vitro* and *in silico* methods and results generated with the DAs in predicting LLNA and human reference classifications. A summary is presented here, and the complete report can be found in **Annex 5,** Impact of Log P on the performance of *in chemico*/*in vitro* assays and ITSv1, ITSv2 and 2o3 Defined Approaches for Skin Sensitisation.

38. Specifically, the performance of the DPRA, KeratinoSens™, h-CLAT, Derek Nexus, OECD QSAR TB, and ITSv1, ITSv2, 2o3 DASS were assessed against the LLNA data in the OECD DASS reference dataset (N = 168). Depending on the information source or DA used, the number of chemicals assessed ranged from 167-168. Note that the analysis was performed with all the chemicals in the dataset with agreed *in vivo* classifications, including inconclusive results of the DAs, borderline results in the *in chemico/in vitro* methods and out of domain *in silico* predictions.

39. The focus of the analysis was on the sensitivity of the various information sources and for the DAs for chemicals with Log P > 3.5 compared to the rest of the chemicals because the specificity measures for such chemicals are too uncertain to be conclusive due to the low number of negative reference chemicals with Log P > 3.5 (N=6). Furthermore, for human health protection it is desirable to minimise the rate of false negative results.

40. When evaluated against LLNA positive and negative reference classifications, the drop in sensitivity for chemicals with Log P > 3.5 (N=39) was in the range of 17 - 29% for the *in vitro* assays and in the range of 1 - 29% for the DAs. The lowest reduction in sensitivity observed was for the *in silico* tools (1 - 5%) and ITSv1 and ITSv2 (11 -12%), which include such tools in the DIP. This can be explained by the fact that *in silico* tools were trained with information from the LLNA and the GPMT, which to some extent mitigate the impact of the reduced sensitivity of the *in chemico*/*in vitro* assays. Regarding specificity calculations for chemicals with Log P > 3.5 it was observed that these have a significant level of uncertainty due to the low number of negative chemicals in this physicochemical range (N=6), see **Annex 5**, Impact of Log P on the performance of *in chemico*/*in vitro* assays and ITSv1, ITSv2 and 2o3 Defined Approaches for Skin Sensitisation for further details.

41. When the DAs were evaluated against the human reference classifications for chemicals with Log P > 3.5, it was concluded that no firm conclusions can be drawn on the impact of sensitivity due to the limited number of chemicals available (N=12, 6 sensitisers and 6 non-sensitisers).

3.2. Meta-analysis of LLNA and human reference data for lipophilic chemicals

42. To shed light on the analysis performed by the United Kingdom and Denmark on the impact of Log P on sensitivity, a further evaluation was performed taking into account additional different lines of evidence. A summary is presented here, and the complete report can be found in (See **Annex 6**, Analysis of LLNA reference data to conclude on predictivity of alternative methods for skin sensitization for lipophilic chemicals for further details).

43. The reduced sensitivity for the *in vitro* assays compared to the *in silico* tools and DAs encompassing such tools (*i.e.* ITSv1 and ITSv2) had been hypothesized to be caused by a potentially limited exposure by poorly soluble chemicals when tested in the aqueous (KeratinoSens™, h-CLAT) or partly aqueous (DPRA) incubation media (hypothesis presented by Denmark to DASS expert group; 5.7.2020).

44. An in-depth evaluation of the data generated with the cell-based assays shows that most of the test chemicals with a Log P > 3.5 producing false negative results compared to LLNA classifications, while not inducing the markers for positivity (luciferase induction or CD86 / CD54 expression), still lead to cytotoxicity in KeratinoSens™ and/or h-CLAT tests, indicating that cells have been sufficiently exposed to the chemicals.

45. Furthermore, for the DPRA, a recent study (14) investigated which of the 82 chemicals initially tested by Gerberick *et al.* (15) to assess DPRA predictivity lead to visible precipitation in the DPRA. Precipitation would indicate that the solution is oversaturated and that the dissolved exposure concentration is below the nominal concentration. Analysing the data in that study, the predictivity of the DPRA as stand-alone method is still at 80% (20/25) for those partly dissolved chemicals. This indicates that the reaction can proceed with the amount of the test chemical being in solution, and additional test chemical can be dissolved as the dissolved test chemical has reacted. These observations indicate that a limited exposure for more hydrophobic chemicals due to insolubility is not a general reason for not detecting the biological activity (in cell assays) or reactivity (in *in chemico*) assays.

46. An initial analysis of the human reference data for chemicals with Log P >3.5 indicated instead, that most of the false-positive LLNA results cluster in this particular physiochemical range (See **Annex 6,** Analysis of LLNA reference data to conclude on predictivity of alternative methods for skin sensitization for lipophilic chemicals for further details). This observation was investigated in a meta-analysis using different reference data: (i) The human data curated by the DASS expert group, (ii) the expert judgment on human sensitization data published by Basketter et al. (16) and (iii) a weight-of-evidence analysis based on DASS data and clinical data. All three analyses indicate that in this physicochemical range the specificity of the LLNA is reduced, which may explain the apparent reduced sensitivity of *in vitro* data when evaluated against the LLNA only. Further support comes from the analysis of studies identifying false-positives in the LLNA vs. guinea pig data and *in silico* structural alerts (see **Annex 6**, Analysis of LLNA reference data to conclude on predictivity of alternative methods for skin sensitization for lipophilic chemicals for further details). This analysis indicates that that there is an uncertainty for the LLNA positive *in vivo* reference data at high Log P and this should be taken into consideration when assessing negative calls from *in vitro* assays or DAs in this physicochemical range.

3.3. Impact of borderline results on the performance of the 2o3 Defined Approach for Skin Sensitisation

47. In most toxicological test methods used for regulatory purposes, continuous data is translated into a binary classification (“positive” or “negative”) using cut-off values irrespective of the data’s variability. Any test result is however subject to variation and these variations increase the uncertainty of a test result especially when close to a (classification) cut-off (in the borderline range). Thus, the *in chemico*/*in vitro* assays used in the DAs may generate different outcomes when results are close to the threshold for discriminating a positive from a negative outcome.

48. The concept of borderline ranges, *i.e.* ranges where the results have lower confidence, is already embedded in some of the test methods providing dichotomous classifications. For example, paragraph 24 of OECD TG 442C on the DPRA test method describes ranges close to the threshold used to discriminate between positive and negative results where additional testing should be performed. However, it was unclear how these ranges were derived. Furthermore, no such a range was defined for either the KeratinoSens™ (TG 442D) nor for the h-CLAT (TG 442E) test.

49. To investigate the impact of *in chemico* and *in vitro* borderline results on predictions made by the 2o3 DA, work was conducted as a joint effort between Germany, Switzerland and the Business and Industry Advisory Committee (BIAC) at the OECD.

50. Borderline ranges have been statistically calculated for each of the test methods composing the 2o3 DA (*i.e.*, DPRA, KeratinoSens™ and h-CLAT) based on the dataset from the formal validation studies of these assays. The ranges extracted from the report (see **Annex 7,** Impact of borderline results on the performances of the 2o3 Defined Approach for Skin Sensitisation for further details) are summarised below:

Table 3.1. Summary of the experimentally derived borderline ranges for the 2o3 DA

	Endpoint	Cut-off	TG borderline range	Validation study mean
DPRA (OECD TG 442C)	Mean peptide depletion [%]	6.38	3-10[a]	4.95 - 8.32
	Cysteine-only depletion [%]	13.89	9-17[a]	10.56 - 18.47
KeratinoSens (OECD TG 442D)	Luciferase induction (fold-change)	1.5	n/a	1.35 - 1.67
h-CLAT (OECD TG 442E)	Relative fluorescence intensity CD54	200	n/a	157 - 255
	Relative fluorescence intensity CD86	150	n/a	122 - 184

Note: See **Annex 7** for additional details.

51. These ranges were used to identify borderline results within the OECD DASS reference dataset and assess their impact on the performances of the 2o3 DA. The final goal was to propose a decision logic to guide the users on the conclusion for borderline cases.

52. When the result of a method was found to be within the above borderline ranges, two options were considered as follows:

- Option 1.1: All borderline predictions (positive if above the decision threshold, and negative if below the decision threshold) were considered uncertain. In this case:
  - If at least two test methods were positive AND non-borderline: the 2o3 DASS UN GHS Cat. 1 (sensitiser) prediction was made.
  - If at least two test methods were negative AND non-borderline: the 2o3 DASS UN GHS NC prediction was made.
  - In all other cases, the 2o3 DASS prediction was considered inconclusive.
- Option 1.2: Only negative borderline predictions were considered uncertain, whereas positive borderline predictions were considered as a positive outcome. In this case:

- The 2o3 DASS prediction UN GHS Cat. 1 was made if at least two test methods were positive AND
    - non-borderline, OR
    - borderline positive (one or more test methods).
- The DASS prediction UN GHS NC was made if at least two test methods were negative AND non-borderline.
- In all other cases, the DASS prediction was considered inconclusive.

53. Considering borderline outcomes within options 1.1 and 1.2 results in an overall decreased in false negatives of the 2o3 DA, with only a slight increase in the false positives. Furthermore, it implies that additional data and/or information or considerations is needed for up to 17% inconclusive chemicals.

54. The borderline approach proposed received general support by the Expert Group and there was also agreement to make use of option 1.1 as the default option for the 2o3 DA. However, option 1.2 may still be considered as an alternative possibility, depending upon the intended use and regulatory context. The decision logic in the GL reflects these decisions (see **Section 2.1.4** of the GL and **Annex 7**, Impact of borderline results on the performances of the 2o3 Defined Approach for Skin Sensitisation for further details).

4. Reference database chemical space characterisation

55. The concerns related to the lower sensitivity of the 2o3 DA for high Log P hydrophobic chemicals were mitigated by considerations of the borderline results, since *in vitro*/*in chemico* test data for a relevant proportion of chemicals with false negative results against *in vivo* reference data have been found to fall in the borderline range. See **Section 3.3** below and **Annex 7** for details.

4.1. Characterisation of the chemical space of the DAs

56. The information reported in this document is meant to support a consistent approach to describe the chemical space covered by defined approaches by determining:

- Chemical reactivity domain of tested chemicals
- Physicochemical properties of tested chemicals

4.1.1. Chemical reactivity domain of the tested chemicals

57. The chemical reactivity domains covered by the chemicals in the reference database were determined with the OECD QSAR Toolbox 4.4 using the profiler "Protein binding alerts for skin sensitisation by OASIS" v2.7. The 168 tested chemicals with reference classifications cover the following reaction domains:

Table 4.1. List of chemical reactivity domains found for the chemicals tested using the 2o3

Chemical reactivity domain	Number of Chemicals
Acylation	14*
Michael addition	29*
No alert found	80
Nucleophilic addition	1
SN2	15*
SNAr	3
Schiff base formation	28*
SNVinyl	1*
Quinoide oxime structure	2
Nitroquinones, naphthoquinone(s)/imines	2
Total	177
Radical reactions (only if autoxidation and/or skin metabolism simulator)	23

*Note:** Iodopropynyl butylcarbamate and benzyl salicylate were categorised as Acylation and SN2; p-Mentha-1,8-dien-7-al was categorised as Michael addition + Schiff base; p-benzoquinone and Brandowski's base were categorised as Michael Addition, Quinoide oxime structure, Nitroquinones, naphthoquinone(s)/imines; Kathon CG was categorised as SN2, SNVinyl. Therefore, 168 chemicals were classified in 177 categories as there are 6 chemicals classified in more than one category which were counted in each category. Chemicals were only categorised as radical reactions if the autoxidation or

skin metabolism simulators were used.

58. The largest reactivity domain group correspond to the “No alert found”, which are the chemicals for which the OECD QSAR Toolbox profiler could not find any alert in the parent structure. Of these 80 chemicals, 15 show alerts if the autooxidation or skin metabolism simulators are used. The second largest groups of domains are the Michael addition (N=29) and Schiff base formation (N=28). And the third largest groups of domains are SN2 (N=15) and Acylation (N=14). The rest of the domains contain 3 or less chemicals each. Alerts for Radical reactions were obtained for 23 chemicals but only when using the autoxidation or skin metabolism simulator.

4.1.2. Physicochemical properties of the tested chemicals

59. The properties that were considered relevant for skin sensitisation and agreed upon by the expert group to be used to characterize the chemical space were: molecular weight (MW), boiling point (BP), melting point (MP), vapour pressure (LogVP), octanol-water partition coefficient (LogP), and water solubility (LogWS).

60. The physicochemical property values used to describe the chemical space covered by the tested chemicals were obtained from the OPERA model, which has been extensively validated against experimental data (Mansouri *et al.* 2018) and that is accessible via the EPA’s chemistry dashboard (https://comptox.epa.gov/dashboard). Experimental values for the physicochemical properties were preferred to the predicted ones, which were only considered when no experimental values available (c.a. 55% of the values were predicted).

61. The properties of mixtures, chemicals of unknown or variable composition, complex reaction products or of biological materials (UVCBs), and natural products cannot be calculated by the means mentioned above and need to be considered as special cases.

62. The summary of the physicochemical property ranges that describe the chemical space of the chemicals tested in the three DAs are shown below. Note that 167 chemicals out of the 168 reference chemicals were used to determine the ranges, since the mixture Kathon CG was excluded as the physicochemical properties could not be calculated.

Table 4.2. Summary of the physicochemical property ranges that describe the chemical space of the chemicals tested in the three DAs

Property	Min-Max
MW(g/mol)	[30.0 - 512.6]
LogP	[-3.9 - 9.4]
LogWS(mol/L)	[-7.6 - 1.2]
MP(ºC)	[-122.5 - 252.7]
BP(ºC)	[-19.1 - 445.3]
LogVP(Pa)	[-18.7 - 11.6]

63. The distributions of the physicochemical properties of the chemicals tested in the DAs are shown next. The values for each property for each chemical are shown as a dot plots and the densities of the distribution of the values are shown as violin plots (light blue area). The chemicals are colour-coded in order to indicate if their corresponding predictions are considered of low (red crosses) or high confidence (black dots) for each DA. See **Annexes 5-7** for further information on the factors determining the confidence of the DAs predictions.

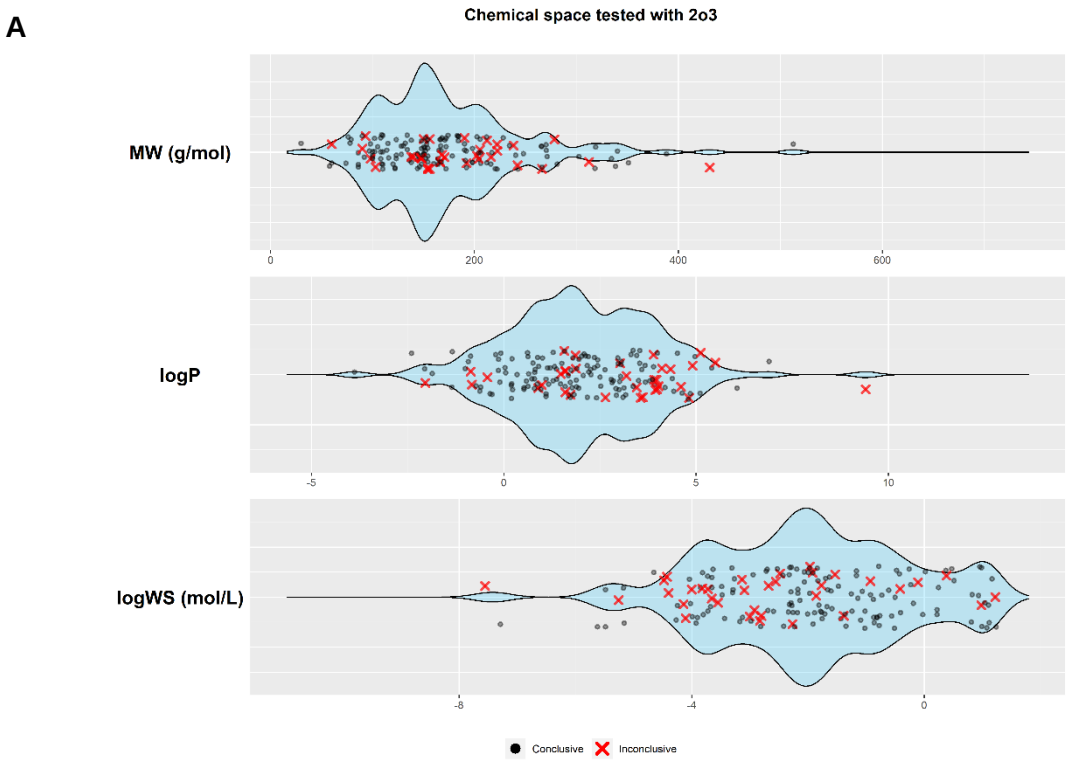
Chemical space covered by the 2o3

64. The 2o3 with consideration of the borderline ranges (see **Annex 7**) provides a total of 34 inconclusive predictions, out of 168 chemicals with reference classifications agreed upon by the EG DASS (20%).

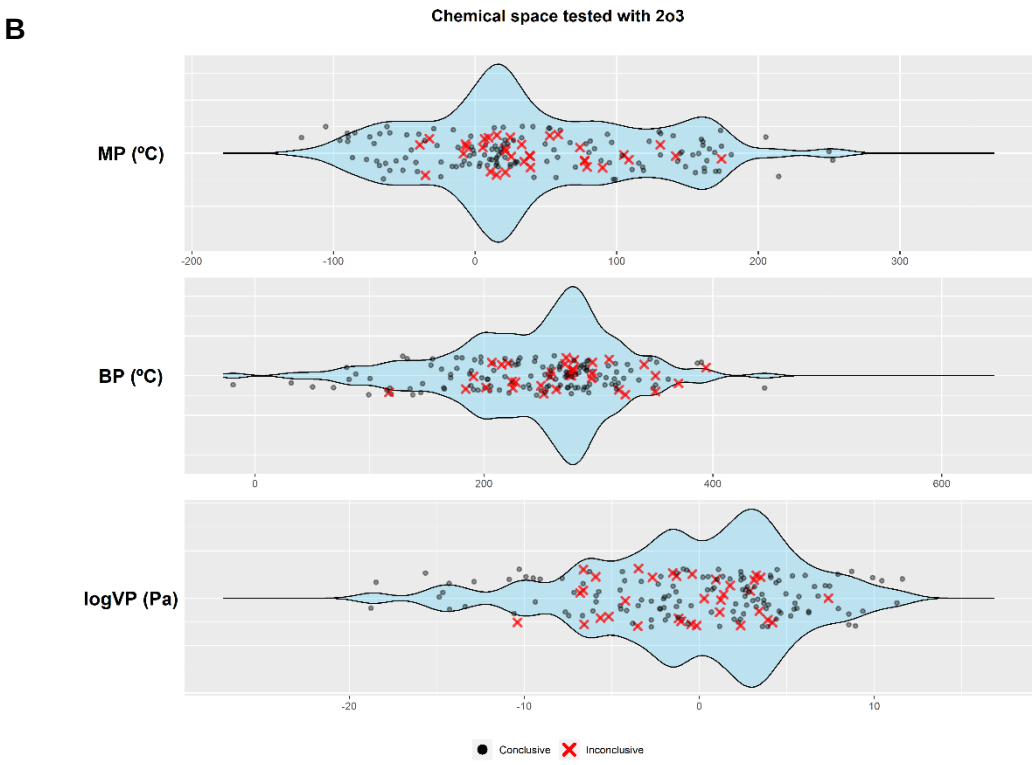
65. The distribution of these inconclusive predictions seems to be in general quite homogenous. Only the MP, BP, and LogVP show ranges in which there are less inconclusive predictions: MP<-50ºC, BP<200ºC, and LogVP<-7.5 and LogVP>5.

Figure 4.1. Physicochemical properties of the 168 reference chemicals that have been tested using the 2o3 DA (composed of DPRA, KeratinoSens™, and h-CLAT)

A) From top to bottom: molecular weight (MW (g/mol)), octanol-water partition coefficient (LogP), water solubility (LogWS(mol/L))



B) From top to bottom: melting point (MP (ºC)), boiling point (BP (ºC)), vapour pressure (LogVP(Pa)).



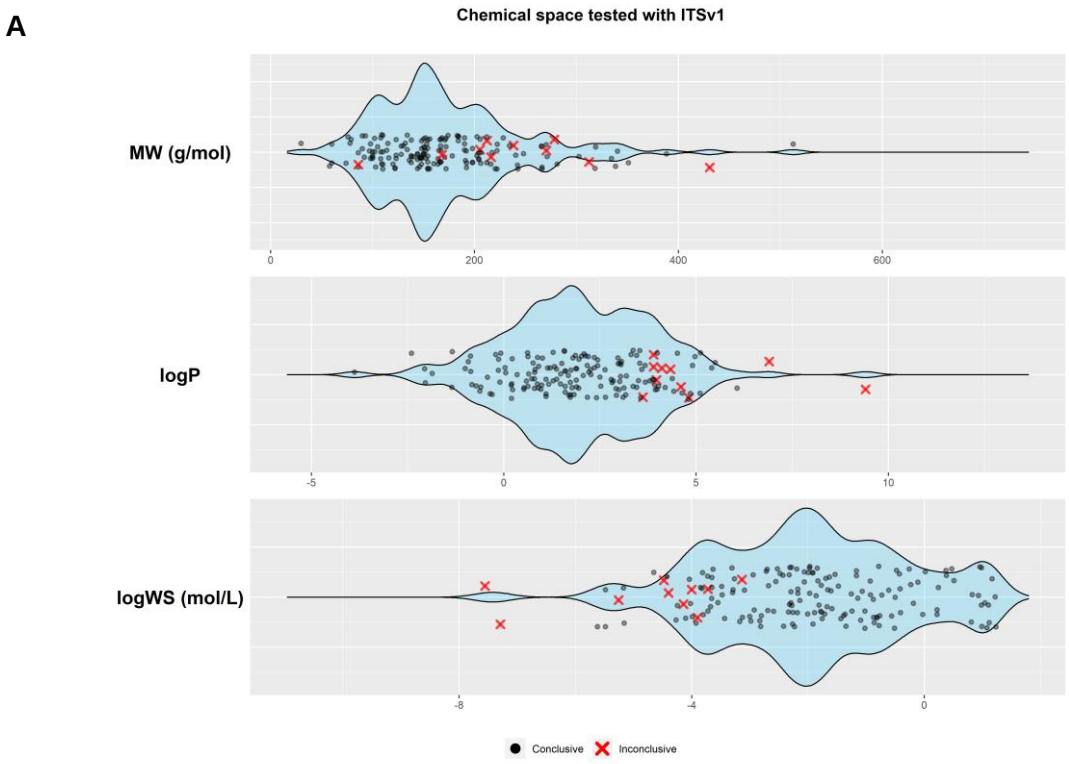
Note: 167 chemicals are included in the plots as the physicochemical properties could not be calculated for the mixture Kathon CG..

Chemical space covered by the ITSv1

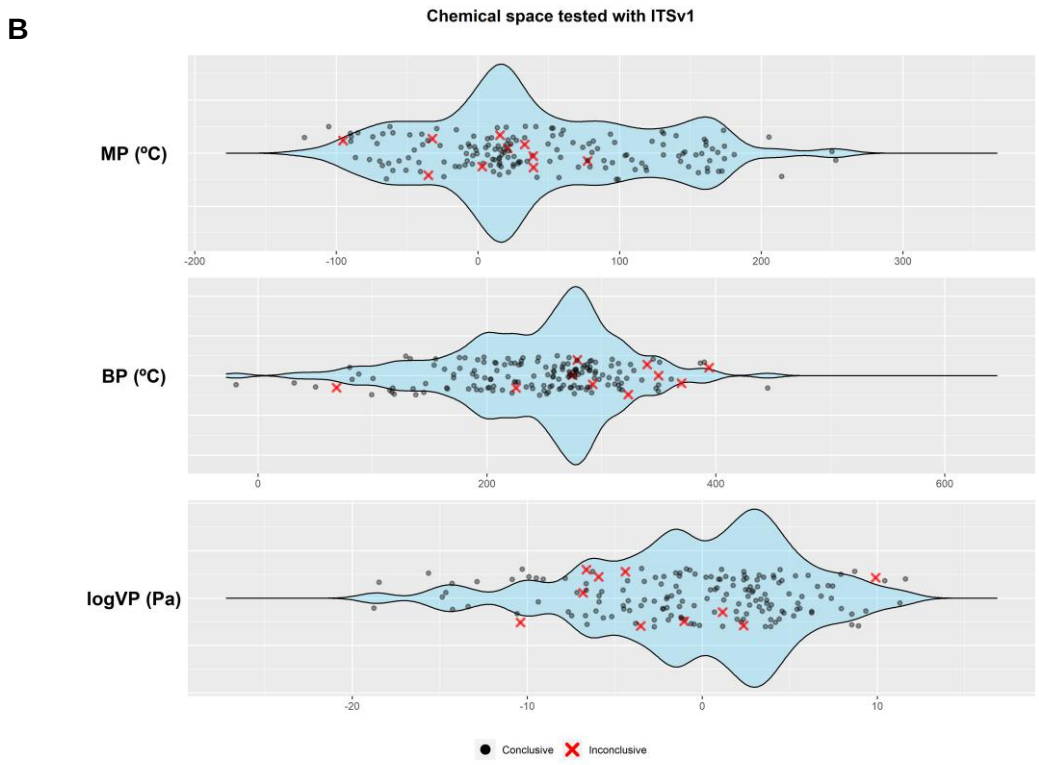
66. The ITSv1 provides 10 inconclusive predictions out of 168 predicted chemicals (6%). These inconclusive predictions correspond to negative predictions for chemicals with Log P > 3.5. ITSv1 can also provide low confidence predictions when Derek contribution is out of domain, but there is no such case within the 168 chemicals tested with ITSv1.

Figure 4.2. Physicochemical properties of the 168 reference chemicals that have been tested using the ITSv1 (composed of DPRA, h-CLAT, and Derek).

A) From top to bottom: molecular weight (MW (g/mol)), octanol-water partition coefficient (LogP), water solubility (LogWS(mol/L))



B) From top to bottom: melting point (MP (ºC)), boiling point (BP (ºC)), vapour pressure (LogVP(Pa)).



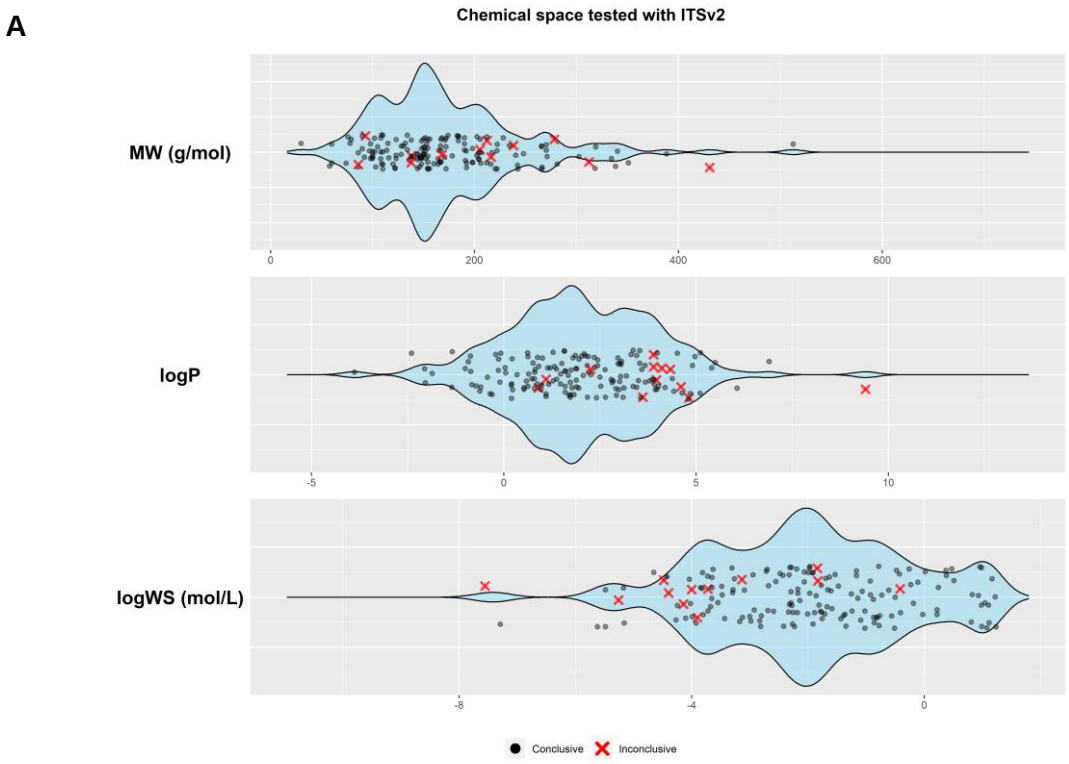
Note: 167 chemicals are included in the plots as the physicochemical properties could not be calculated for the mixture Kathon CG.

Chemical space covered by the ITSv2

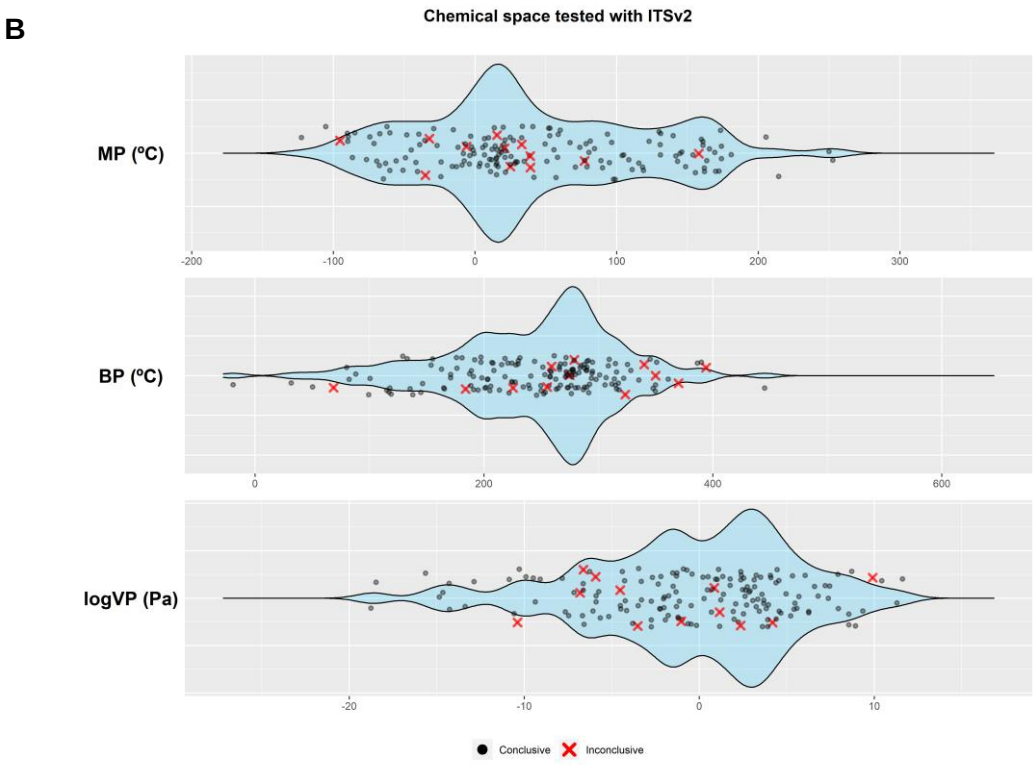
67. The ITSv2 provides 12 inconclusive predictions out of 168 predicted chemicals (7%). These low confidence predictions are in 9/12 cases due to the limitations of the negative predictions of chemicals with Log P > 3.5. The other three low confidence predictions, aniline, anisyl alcohol, and salicylic acid are low confidence predictions because the OECD QSAR Toolbox prediction is out of domain and the scores of the other two components together are, 1,1, and 2, respectively.

Figure 4.3. Physicochemical properties of the 168 reference chemicals that have been tested using the ITSv2 (composed of DPRA, h-CLAT, and OECD QSAR TB prediction).

A) From top to bottom: molecular weight (MW (g/mol)), octanol-water partition coefficient (LogP), water solubility (LogWS(mol/L))



B) From top to bottom: melting point (MP (ºC)), boiling point (BP (ºC)), vapour pressure (LogVP(Pa)).



Note: 167 chemicals are included in the plots as the physicochemical properties could not be calculated for the mixture Kathon CG.

5. Performance of individual methods and DAs

68. The performances of the individual methods and DAs that are included in the GL have been summarised here. The performances are shown for LLNA and human hazard, and potency sub-categorization. The performances are based on reference classifications agreed upon by the EG DASS. In addition, the performance of the individual methods and DAs for pre/pro-haptens is shown in **Section 5.2**.

5.1. Summary of performance of the individual methods and DAs vs LLNA and human reference data

Table 5.1. Predictivity of the individual methods and DAs for LLNA and human skin sensitisation hazard

Method/DA	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity
Reference Data	**vs LLNA**			**vs Human**		
LLNA Call (N=56)	-	-	-	0.58	0.94	0.22
DPRA Call (no borderlines, N=156, 60)	0.81	0.74	0.87	0.86	0.82	0.89
DPRA Call (all, N= 168, 65)	0.78	0.70	0.85	0.82	0.82	0.82
KS Call (no borderlines, N=157, 59)	0.71	0.74	0.68	0.77	0.76	0.78
KS Call (all, N=168, 66)	0.70	0.71	0.70	0.76	0.71	0.82
h-CLAT Call (high confidence only) (N=113, 51)	0.78	0.92	0.64	0.69	0.88	0.5
h-CLAT Call (all, N=167, 65)	0.74	0.81	0.67	0.71	0.87	0.55
Derek Nexus (in domain only, N=165, 64)	0.82	0.91	0.73	0.71	0.96	0.46
Derek Nexus (all, N=168, 66)	0.81	0.89	0.73	0.69	0.93	0.45
OECD TB (AW SS) (in total domain only, N=148, 58)	0.83	0.94	0.73	0.74	0.92	0.56
OECD TB (AW SS) (all, N=167, 65)	0.81	0.90	0.73	0.74	0.83	0.64
2o3 Call (conclusive only, N=134, 55)	0.84	0.82	0.85	0.88	0.89	0.88
2o3 Call (all, N=168, 65)	0.79	0.74	0.85	0.83	0.83	0.82
ITSv1 Call (conclusive only, N=159, 55)	0.81	0.92	0.70	0.69	0.93	0.44
ITSv1 Call (all, N=168, 65)	0.80	0.87	0.73	0.74	0.93	0.55
ITSv2 Call (conclusive only, N=156, 64)	0.80	0.93	0.67	0.69	0.94	0.44
ITSv2 Call (all, N=168, 66)	0.79	0.87	0.70	0.73	0.91	0.55

*The chemicals used to obtain the statistics may vary for the different methods providing more or less robust measures. The total number of chemicals used for each method is indicated in brackets vs LLNA and HDSG, respectively. The results for "h-CLAT Call (high confidence only)" do not include borderlines neither negative predictions for Log P > 3.5 chemicals. Due to the biased nature of the dataset towards sensitisers, sensitivity measures will typically be more robust than specificity. Please refer to **Annex 2** or specific sections of the GL for details on the exact number of chemicals used to obtain the statistics shown above.

Table 5.2. Predictivity of the DAs for LLNA and Human skin sensitisation potency (GHS)

Method/DA	Overall Accuracy	Average Balanced Accuracy	Overall Accuracy	Average Balanced Accuracy
Reference Data	**vs LLNA GHS**		**vs Human GHS**	
LLNA Call (N=47)	-	-	0.60	0.64
ITSV1 Call (conclusive only, N=146, 60)	0.71	0.78	0.68	0.72
ITSv1 Call (all, N=156, 63)	0.69	0.77	0.69	0.74
ITSv2 Call (conclusive only, N=141, 57)	0.71	0.77	0.70	0.73
ITSv2 Call (all, N=156, 63)	0.67	0.75	0.68	0.73

*The chemicals used to obtain the statistics may vary for the different methods providing more or less robust measures. The total number of chemicals used for each method/DA is indicated in brackets vs LLNA and HDSG, respectively. Due to the biased nature of the dataset towards sensitisers, sensitivity measures will typically be more robust than specificity. Please refer to **Annex 2** or specific sections of the GL for details on the exact number of chemicals used to obtain the statistics shown above.

Table 5.3. Potency classification performance of the ITSv1 DA in comparison to LLNA reference data, based on the GHS 1A/1B sub-categorisation

	LLNA		
ITSv1 DA	NC	1B	1A
NC	21	11	0
1B	9	55	10
1A	0	12	28
Inconclusive	3	7	0

78% average balanced accuracy overall

ITSv1 Statistics by Class:

Class-wise performance (N=146)	NC (N=30)	1B (N=78)	1A (N=38)
Sensitivity (%)	70	71	74
Specificity (%)	91	72	89
Balanced Accuracy (%)	80	71	81

Note: Sensitivity is ability to detect membership within class, specificity is ability to detect membership outside of class, and balanced accuracy is average of sensitivity and specificity. Statistics reflect high confidence predictions only; inconclusive predictions are shown in grey.

Table 5.4. Potency classification performance of the ITSv2 DA in comparison to LLNA reference data (N = 153 chemicals), based on the GHS 1A/1B sub-categorisation

	LLNA		
ITSv2 DA	NC	1B	1A
NC	20	9	0
1B	10	54	10
1A	0	12	26
Inconclusive	3	10	2

77% average balanced accuracy overall

ITSv2 Statistics by Class:

Class-wise performance (N=145)	**NC (N=30)**	**1B (N=75)**	**1A (N=36)**
Sensitivity (%)	67	72	72
Specificity (%)	92	70	89
Balanced Accuracy (%)	79	71	80

Note: Sensitivity is ability to detect membership within class, specificity is ability to detect membership outside of class, and balanced accuracy is average of sensitivity and specificity. Statistics reflect high confidence predictions only; inconclusive predictions are shown in grey.

Table 5.5. Potency classification performance of the ITSv1 DA in comparison to Human reference data, based on the GHS 1A/1B sub-categorisation

	Human		
ITSv1 DA	NC	1B	1A
NC	4	4	0
1B	5	24	7
1A	0	3	13
Inconclusive	2	0	1

72% average balanced accuracy overall

ITSv1 vs Human Statistics by Class:

Class-wise performance (N=60)	**NC (N=9)**	**1B (N=31)**	**1A (N=20)**
Sensitivity (%)	44	77	65
Specificity (%)	92	59	93
Balanced Accuracy (%)	68	68	79

Note: Sensitivity is ability to detect membership within class, specificity is ability to detect membership outside of class, and balanced accuracy is average of sensitivity and specificity. Statistics reflect high confidence predictions only; inconclusive predictions are shown in grey.

Table 5.6. Potency classification performance of the ITSv2 DA in comparison to Human reference data (N = 60 chemicals), based on the GHS 1A/1B sub-categorisation

	Human		
ITSv2 DA	NC	1B	1A
NC	4	3	0
1B	5	24	6
1A	0	3	12
Inconclusive	2	1	3

73% average balanced accuracy overall

ITSv2 vs Human Statistics by Class:

Class-wise performance (N=57)	**NC (N=9)**	**1B (N=30)**	**1A (N=18)**
Sensitivity (%)	44	80	67
Specificity (%)	94	59	92
Balanced Accuracy (%)	69	70	80

Note: Sensitivity is ability to detect membership within class, specificity is ability to detect membership outside of class, and balanced accuracy is average of sensitivity and specificity. Statistics reflect high confidence predictions only; inconclusive predictions are shown in grey.

Table 5.7. Potency classification performance of the LLNA in comparison to Human reference data, based on the GHS 1A/1B sub-categorisation

	Human		
LLNA	NC	1B	1A
NC	2	3	0
1B	6	17	7
1A	0	3	9

64% average balanced accuracy overall

LLNA vs Human Statistics by Class:

Class-wise performance (N=47)	**NC (N=8)**	**1B (N=23)**	**1A (N=16)**
Sensitivity (%)	25	74	56
Specificity (%)	92	46	90
Balanced Accuracy (%)	59	60	73

Note: Sensitivity is ability to detect membership within class, specificity is ability to detect membership outside of class, and balanced accuracy is average of sensitivity and specificity.

69. As shown by the potency sub-categorisation statistics above, overall balanced accuracy and class-wise performance of the ITSv1 and ITSv2 vs. human HPPT reference data was comparable to and/or exceeded that of the LLNA vs. human HPPT reference data. It is noted that due to the imbalanced nature of the reference data and the small numbers of chemicals, the measures of accuracy are more uncertain for smaller classes, *e.g.* for NC chemicals.

5.2. Supplementary analyses of specificity by inclusion of additional "potential LLNA negatives

70. A supplementary small analysis was conducted because of the concern that the application of the strict EG agreed criteria for identifying LLNA negative reference chemicals resulted in the identification of only 33 LLNA negative chemicals. The relatively small number of LLNA negatives in the reference data set would namely imply that the validation measures of the specificity of the proposed DASS approaches would be more uncertain that the equivalent measures of the sensitivity of the proposed DASS approaches, because the measure of sensitivity would be based on more than 4 times as many LLNA positive chemicals as the measure of specificity. Hence a supplementary analysis employing less strict criteria for identifying reference chemicals, which may be regarded as LLNA negatives, was conducted and is summarized in **Annex 8**. The results of this supplementary specificity analysis showed that such relaxed criteria for LLNA negatives would result in marked lower specificity values for the three DASS approaches. The supplementary analysis was performed after the EG agreement was obtained regarding the criteria for identification of the LLNA negative reference chemicals. However, the results of the supplementary specificity analysis support that EG agreed and applied strict criteria for identifying LLNA negative reference chemicals is reasonable, although the number of such chemicals was relatively low, *i.e.* approximately 4 times lower than the number of LLNA positive reference chemicals resulting in a higher uncertainty for the measured specificity values than that for the sensitivity values.

5.3. Analysis of the performance of DAs at predicting pre/pro haptens

71. Pre- and pro-haptens are considered difficult chemicals to predict as they require metabolic or autoxidation transformation to become sensitisers. Most non animal methods have no or limited metabolic capacity, and their ability to predict skin sensitisation for pre/pro haptens is often questioned. The chemicals of the dataset suspected to be pre-, pro-, or both (17, 18), were categorised as such and the performance of the DAs predicting them is shown in Table 5.8. Metol was considered a pre/pro-hapten based on Urbisch et al. data (2).

Table 5.8. Comparison of individual methods and DAs vs LLNA for pre/pro haptens (only high confidence and conclusive predictions are included)

	DPRA vs LLNA	KeratinoSens™ vs LLNA	h-CLAT vs LLNA	2o3 vs LLNA	Derek Nexus vs LLNA	OECD QSAR Toolbox vs LLNA	ITSv1 vs LLNA	ITSv2 vs LLNA
False negative	7	9	1	4	0	0	0	0
False positive	0	0	0	0	0	0	0	0
True negative	0	0	0	0	0	0	0	0
True positive	19	18	22	19	27	27	29	28
Inconclusive	3	2	6	6	2	2	0	1
n	29	29	29	29	29	29	29	29
Accuracy	-	-	-	-	-	-	-	-
Sensitivity	0.73	0.67	0.96	0.83	1.00	1.00	1.00	1.00
Specificity	-	-	-	-	-		-	-

72. There were 29 chemicals categorised as pre/pro haptens in the dataset of 168 chemicals (17%) and all of them were positive in the LLNA. Of these, only 9 chemicals had human data available and all were found to be sensitisers in humans.

73. Since all chemicals in the subset were positive in the LLNA, there are no known FP predictions, and consequently no specificity or accuracy. Regarding false negatives, DPRA and KeratinoSens™ have 7 and 9, respectively. These represent 26% of the false negatives of DPRA and 42% of KeratinoSens™. The 2o3 has a low number of false negatives, 4 which represents 20% of all its false negatives. h-CLAT has one false negative, which represents 13% of its false negatives, and ITSv1, ITSv2, Derek Nexus, and the OECD QSAR Toolbox have 0 false negatives. In terms of conclusive and inconclusive results, the 2o3 and h-CLAT are the methods with more inconclusive predictions, 6/29=21%, followed by DPRA with 3/29=11%, and KeratinoSens™, Derek, and the OECD QSAR Toolbox with 2/29=7%, and ITSv2 with only 1/29=3%. ITSv1 is the only method that has no inconclusive predictions in this subset of chemicals. DPRA and KeratinoSens™ show the lower sensitivities of all the methods, which are both below 0.75. The 2o3 has a sensitivity of 0.83, h-CLAT of 0.96, Derek Nexus, the OECD QSAR Toolbox, ITSv1, and ITSv2 have sensitivities of 1.0 as they correctly predict all the chemicals in the subset for which they report conclusive predictions.

5.3.1. Observations for pre/pro-haptens:

74. In terms of DAs and individual methods, ITSv1 and ITSv2 show excellent performance for pre/pro haptens as they predict all chemicals correctly, with only 1 inconclusive result for ITSv2, which would be false negative if it was considered. The 2o3, shows a sensitivity of 0.83 but with more inconclusive results than the other two methods (N=6). Regarding the individual methods, DPRA and KeratinoSens™ have sensitivities below 0.75 and are the lowest of all the methods compared. h-CLAT is the individual method with the highest sensitivity, 0.92, but also the one with the highest number of inconclusive predictions (N=6).

6. References

1. Bauch C, Kolle SN, Ramirez T, Eltze T, Fabian E, Mehling A, Teubner W, van Ravenzwaay B, Landsiedel R. (2012) Putting the parts together: combining *in vitro* methods to test for skin sensitizing potentials. Regul Toxicol Pharmacol, 63:489-504.
2. Urbisch D, Mehling A, Guth K, Ramirez T, Honarvar N, Kolle S, Landsiedel R, Jaworska J, Kern PS, Gerberick F, Natsch A, Emter R, Ashikaga T, Miyazawa M, Sakaguchi H. (2015). Assessing skin sensitization hazard in mice and men using non-animal test methods, Regul Toxicol Pharmacol, 71:337-51.
3. Nukada Y, Miyazawa M, Kazutoshi S, Sakaguchi H, Nishiyama N. (2013).Data integration of non-animal tests for the development of a test battery to predict the skinsensitizing potential and potency of chemicals. Toxicol *In vitro*, 27:609-18.
4. Takenouchi O, Fukui S, Okamoto K, Kurotani S, Imai N, Fujishiro M, Kyotani D, Kato Y, Kasahara T, Fujita M, Toyoda A, Sekiya D, Watanabe S, Seto H, Hirota M, Ashikaga T, Miyazawa M. (2015). Test battery with the human cell line activation test, direct peptide reactivity assay and DEREK based on a 139 chemical data set for predicting skin sensitizing potential and potency of chemicals. J Appl Toxicol, 35:1318-32.
5. OECD (2010), OECD Guidelines for Chemical Testing No. 429. Skin sensitisation: Local Lymph Node assay. Organisation for Economic Cooperation and Development, Paris. Available at: [http://www.oecd.org/env/testguidelines]
6. United Nations (UN) (2019), Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Eighth revised edition, New York and Geneva, United Nations Publications. Available at: [https://unece.org/ghs-rev8-2019]
7. OECD (2015), OECD Guideline for the Testing of Chemicals No. 442C: *In chemico* Skin Sensitisation: Direct Peptide Reactivity Assay (DPRA). Paris, France: Organisation for Economic Cooperation and Development. Available at: http://www.oecd.org/env/testguidelines
8. OECD (2018), OECD Key Event based test Guideline 442D: *In vitro* Skin Sensitisation Assays Addressing AOP Key Event on Keratinocyte Activation. Organisation for Economic Cooperation and Development, Paris. Available at: [http://www.oecd.org/env/testguidelines].
9. OECD (2018), OECD Key event-based test Guideline 442E: *In vitro* Skin Sensitisation Assays Addressing the Key Event on Activation of Dendritic Cells on the Adverse Outcome Pathway for Skin Sensitisation. Organisation for Economic Cooperation and Development, Paris. Available at: [http://www.oecd.org/env/testguidelines].
10. OECD (2016), Series on Testing & Assessment No. 256: 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, Annex 1 and Annex 2. ENV/JM/HA(2016)29. Organisation for Economic Cooperation and Development, Paris. Available at: [https://community.oecd.org/community/iatass].

11. ECETOC Technical Report 087 (2003), Contact Sensitisation: Classification According to Potency. Available at: [https://www.ecetoc.org/publication/tr-087-contact-sensitisation-classification-according-to-potency/]

12. Kleinstreuer N, Hoffmann S, Alepee N, *et al.* (2018) Non-Animal Methods to Predict Skin Sensitization (II): an assessment of defined approaches. Crit Rev Toxicol Feb 23:1-16. doi: 10.1080/10408444.2018.1429386

13. Takenouchi, O., Miyazawa, M., Saito, K., Ashikaga, T. & Sakaguchi, H. Predictive performance of the human Cell Line Activation Test (h-CLAT) for lipophilic chemicals with high octanol-water partition coefficients. J. Toxicol. Sci. 38, 599–609 (2013).

14. Yamamoto, Y., Wanibuchi, S., Sato, A., Kasahara, T., Fujita, M., 2019. Precipitation of test chemicals in reaction solutions used in the amino acid derivative reactivity assay and the direct peptide reactivity assay. Journal of Pharmacological and Toxicological Methods. 100.10.1016/j.vascn.2019.106624

15. Gerberick, G. F., Vassallo, J. D., Foertsch, L. M., Price, B. B., Chaney, J. G., Lepoittevin, J. P., 2007. Quantification of chemical peptide reactivity for screening contact allergens: A classification tree model approach. Toxicological Sciences. 97, 417-427

16. Basketter DA, Alépée N, Ashikaga T, Barroso J, Gilmour N, Goebel C, Hibatallah J, Hoffmann S, Kern P, Martinozzi-Teissier S, Maxwell G, Reisinger K, Sakaguchi H, Schepky A, Tailhardat M, Templier M. (2014). Categorization of chemicals according to their relative human skin sensitizing potency, Dermatitis, 25(1):11-21.

17. Patlewicz, G., Casati, S., Basketter, D. A., Asturiol, D., Roberts, D. W., Lepoittevin, J.-P., Worth, A. P., & Aschberger, K. (2016). Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? Regulatory Toxicology and Pharmacology, 82. https://doi.org/10.1016/j.yrtph.2016.08.007

18. Casati S, Aschberger K, Asturiol D, Basketter D, Dimitrov S, Dumont C, Karlberg AT, Lepoittevin JP, Patlewicz G, Roberts DW and Worth A (2016). Ability of non-animal methods for skin sensitisation to detect pre- and pro-haptens: Report and recommendations of an EURL ECVAM expert meeting; EUR 27752 EN; doi:10.2788/01803

7. List of Annexes to this Document[3]

Annex 1: Evaluation framework

Annex 2: Reference Data Matrix and Comparison

Annex 3: Report on the curation and evaluation of the LLNA reference data used for assessing performance of Defined Approaches for Skin Sensitisation

Annex 4: Report of the Human Data Sub-Group on the Curation and Evaluation of Human Reference Data

Annex 5: Impact of LogP on the performance of *in chemicolin vitro* assays and ITSv1, ITSv2 and 2o3 Defined Approaches for Skin Sensitisation

Annex 6: Analysis of LLNA reference data to conclude on predictivity of alternative methods for skin sensitization for lipophilic chemicals

Annex 7: Impact of borderline results on the performances of the 2o3 Defined Approach for Skin Sensitisation

Annex 8: Supplementary analyses of specificity by inclusion of additional "potential LLNA negatives"

[3] All files are available on the WNT community site