Research Article

Integrated Skin Sensitization Assessment Based on OECD Methods (II): Hazard and Potency by Combining Kinetic Peptide Reactivity and the “2 out of 3” Defined Approach

Andreas Natsch1 and George Frank Gerberick2

1Fragrances S&T, Ingredients Research, Givaudan Schweiz AG, Kemptthal, Switzerland; 2GF3 Consultancy, LLC, Cincinnati, OH, USA

Abstract

Depending on regulatory requirements, the skin sensitization risk for new chemicals with potential consumer skin contact must be assessed by experimental testing by (i) binary hazard assessment to identify sensitizers, (ii) subclassification of sensitizers according to the Global Harmonized System (GHS), and (iii) derivation of a point of departure (PoD) for risk assessment. The Organisation for Economic Co-operation and Development (OECD) recently published a test guideline incorporating the “2 out of 3” defined approach (2o3 DA) for skin sensitization hazard assessment and added the kinetic direct peptide reactivity assay (kDPRA) as a stand-alone test guideline method for GHS subclassification. The 2o3 DA requires that at least two in vitro tests are conducted. The cell-based tests and the kDPRA generate, next to a binary outcome with a fixed threshold, continuous concentration-response data, which can be used in quantitative regression models to derive a PoD. The sequence of testing for the 2o3 DA is flexible. Here we compare different testing sequences and how they can be combined with kDPRA data to provide a PoD in parallel to hazard identification (hazard ID) and GHS subclassification. A set of 188 chemicals with available in vitro data was evaluated for the final PoD using these different testing sequences. The results indicate that testing can start with DPRA/kDPRA and either of the cell-based assays, and that testing can stop after two congruent tests without major impact on the final PoD for chemicals within the applicability domain of the kDPRA.

1 Introduction

A significant effort is underway to develop next-generation risk assessment (NGRA) approaches for skin sensitization that do not rely on new animal test data. New approach methodologies (NAMs), i.e., non-animal test methods, have been developed to identify skin sensitization hazards, and these have a new focus on determining potency information for risk assessment purposes (Bernauer et al., 2021; Dent et al., 2018; Ezendam et al., 2016; Gilmour et al., 2020; Kleinreuer et al., 2018). The ban on animal testing for new cosmetic ingredients, which was implemented in Europe within the cosmetics legislation (Regulation (EC) No 1223/2009), led to the rapid development of NAMs by both academic and industrial laboratories (Ezendam et al., 2016). Three OECD guidelines have been published that cover mechanistic key events (covalent binding to protein, keratinocyte activation and dendritic cell activation). These three mechanistic events map to key events 1-3 of the skin sensitization AOP (OECD, 2014). Eight non-animal test methods have been approved and are included in OECD TGs (direct peptide reactivity assay, DPRA; amino acid derivative reactivity assay, ADRA; kinetic DPRA, kDPRA; ARE-Nrf2 luciferase assay KeratinoSens™, KS; ARE-Nrf2 luciferase assay, LuSens; human cell line activation test, h-CLAT; U937 cell line activation test, U-SENS™; and interleukin-8 reporter gene assay, IL-8 Luc Assay) (OECD, 2018a,b, 2021b).

Recent work has focused on finding ways to combine NAM data to generate integrated approaches to testing and assessment (IATA) or defined approaches (DA). DA for skin sensitization contain fixed data interpretation procedures (DIP) on how to combine data obtained from different in chemico, in vitro and in silico methods to conclude whether a substance is a skin sensitiz-
er and, if so, to define its potency as a skin sensitizer (Gilmour et al., 2020; Hoffmann et al., 2018; Kleinsteuber et al., 2018). Two simple DAs for assessing skin sensitization have been published in a new guideline (OECD, 2021a). OECD TG 497 includes the 2o3 DA and integrated testing strategy (ITSV1 and ITS v2) DA. In the 2o3 DA, a hazard assessment is provided by two concordant, non-borderline (non-BL) results from DPRA, KS and h-CLAT (Bauch et al., 2012; Natsch et al., 2021; Urbisch et al., 2015). The 2o3 DA does not provide information on the skin sensitization potency. While ITS v1 and v2 integrate an in silico prediction, the 2o3 is based only on experimental data from OECD validated tests.

Assessing skin sensitization potency is needed for the binary subclassification of sensitizers into 1A (strong sensitizers) and 1B (other sensitizers) in the UN Global Harmonized System (GHS). The kDPRA assay, which has been recently added to OECD TG 442C, is a standalone assay for the application of subcategory 1A (Natsch et al., 2020; OECD, 2021b; Wareing et al., 2020). An assessment of potency on a more granular scale is needed for NGRA of new chemical entities. Thus, it is advantageous for risk assessors to have available approaches that can provide continuous PoD values so that more quantitative assessments can be made to help protect workers and consumers.

Linear regression models using KS and kinetic peptide reactivity data have been proposed to provide a PoD value in the form of a predicted EC3 value in the local lymph node assay (LLNA) (Natsch et al., 2015, 2018). Building on this previous approach using regression models, updated quantitative models using input data from the kDPRA, the KS and the h-CLAT were generated to calculate a PoD (Natsch and Gerberick, 2022). The predictive models were produced using a comprehensive database that included test data from the accepted OECD methods. All models were examined using a set of case studies selected based on multiple LLNA reference data in the OECD database. The robustness of the models was characterized by comparing a comprehensive historical database versus the curated dataset provided by the OECD working group on DA. The predicted PoD were within or close to the variation of the historical LLNA data for most of the case studies. Overall, the models predict the in vivo value with a median fold-misprediction factor of around 2.5.

The various models offer risk assessors flexibility in the choice of tests, and a PoD value can still be determined when there are compatibility issues or when chemicals are outside the chemical domain of an individual assay.

In this paper, it is demonstrated how the kDPRA and these quantitative models can be combined in different testing sequences in the 2o3 DA to provide at the same time (i) hazard ID, (ii) GHS subclassification, and (iii) PoD determination based on the validated in vitro tests. The integrated assessment presented is solely based on in vitro data from three OECD test guidelines. Thus, this work further advances the 3Rs for skin sensitization testing as it gives practical guidance on how to finally combine the methods and evaluates these proposed strategies on a large number of chemicals.

2 Materials and methods

Database used

The analysis in this paper is based on a comprehensive database on 188 chemicals with data in the kDPRA, KS, h-CLAT and the LLNA, and no new data were generated for this study (Tab. ESM1-1; the data presented are a subset of the larger database presented in a parallel paper (Natsch and Gerberick, 2022)). For 154 of these chemicals, data are also available in the OECD reference database (OECD DB) compiled by the OECD DA working group (OECD, 2021c). LLNA data from published historical compilations are available for all the 188 chemicals. In parallel, for the subset in the OECD database, a curated LLNA value is available based on evaluating the original data with a set of fixed rules (OECD, 2021d). The analysis for accuracy of the PoD determination was made with the historical LLNA data and in parallel with the curated LLNA data, using the historical data only for the chemicals that were not in the OECD DB.

Regression models and statistics

The data normalizations and calculations are described in a parallel paper (Natsch and Gerberick, 2022). Based on the test data from the OECD tests, a prediction spreadsheet can be used to calculate a predicted EC3 as PoD based on regression equations. The following four regression models are used here and are implemented in this prediction spreadsheet:

For a PoD assessment based on KS and kDPRA data:

$$EQ1 \quad pEC3 = 0.42 + 0.40 \times \log k_{\text{max norm}} + 0.15 \times \log EC1.5_{\text{norm}} + 0.36 \times \log IC50_{\text{norm}} - 0.21 \times \log VP_{\text{norm}}$$

For a PoD assessment based on h-CLAT and kDPRA data:

$$EQ4 \quad pEC3 = 0.18 + 0.36 \times \log k_{\text{max norm}} + 0.21 \times \log MIT_{\text{norm}} + 0.35 \times \log CV75_{\text{norm}} - 0.19 \times \log VP_{\text{norm}}$$

For a PoD assessment based on KS, h-CLAT and kDPRA data:

$$EQ5 \quad pEC3 = 0.20 + 0.34 \times \log k_{\text{max norm}} + 0.20 \times \log MIT_{\text{norm}} + 0.09 \times \log EC1.5_{\text{norm}} + 0.21 \times \log CV75_{\text{norm}} + 0.11 \times \log IC50_{\text{norm}} - 0.19 \times \log VP_{\text{norm}}$$

Abbreviations

2o3 DA, “2 out of 3” defined approach; AD, applicability domain; AOP, adverse outcome pathway; BL, borderline; DA, defined approach; DPRA, direct peptide reactivity assay; GHS, Global Harmonized System; hazard ID, hazard identification; h-CLAT, human cell line activation test; IATA, integrated approach to testing and assessment; ITS, integrated testing strategy; kDPRA, kinetic direct peptide reactivity assay; KS, KeratinoSens™; LLNA, local lymph node assay; NAM, new approach methodology; NGRA, next-generation risk assessment; PoD, point of departure; OECD, Organisation for Economic Co-operation and Development; OECD DB, OECD reference database on defined approaches; QRA, quantitative risk assessment; TG, test guideline; VP, vapor pressure

1 doi:10.14573/altex.2201142s1
For a PoD assessment based on KS and h-CLAT data:

\[ \text{EQ6} \quad \text{pEC3} = 0.09 + 0.276 \times \log \text{MIT}_\text{norm} + 0.22 \times \log \text{EC1.5}_\text{norm} + 0.34 \times \log \text{CV75}_\text{norm} + 0.06 \times \log \text{IC50}_\text{norm} - 0.12 \times \log \text{VP}_\text{norm} \]

The parameters used in these equations are (i) from the kDPRA the Log \( k_{\text{max}}_\text{norm} \), the normalized, logarithmic rate constant, (ii) from the KS the Log IC50\text{norm}, the normalized IC50 value (concentration for 50% reduction in cellular viability) and the Log EC1.5\text{norm}, the normalized EC1.5 value indicating the concentration for 1.5-fold induction of luciferase activity, and (iii) from the h-CLAT the normalized Log MIT\text{norm}, indicating the lowest concentration for either 1.5-fold CD86 or 2-fold CD54 induction, and the Log CV75\text{norm} indicating concentration for 25% reduction in viability. In addition, the Log VP\text{norm} describes the volatility for chemicals evaporating significantly from the LLNA vehicle within 60 min.

To assess the prediction accuracy of quantitative models, the ratio between the larger and the smaller values of the measured and predicted EC3 value was calculated in each case to give the fold-misprediction. Median and geometric means were calculated for this measure of the data fit, and the number of chemicals mispredicted by > 5-fold or by > 10-fold in either direction are listed.

For assessment of subclassification, sensitizers were discriminated from non-sensitizers with the 2o3 DA, taking borderline (BL) outcomes in the individual tests into account as described in OECD TG 497 (OECD, 2021a). Data are presented as a three-way classification table. For analysis of this prediction of three classes, only the OECD data were used as BL analysis could not be done on the additional published h-CLAT data.

### 3 Results

#### 3.1 An economical testing sequence to include GHS subclassification and PoD determination into the 2o3 DA

In the 2o3 DA, a hazard assessment is provided by two concordant, non-BL results from DRPA, KS and h-CLAT (OECD, 2021a). The testing sequence does not affect the outcome of this hazard assessment. Here, we provide the most economical testing sequence and indicate two alternative approaches (either starting with h-CLAT or conducting all assays by default). The goal of all these testing sequences, as described here, is to provide hazard ID, GHS subclassification, and PoD determination based on results from DRPA, kDPRA, KS, and h-CLAT.

An efficient testing sequence is shown in Figure 1. Testing starts with DRPA and KS since these tests are more economical in most test laboratories and lead to fewer inconclusive/BL outcomes as compared to the h-CLAT (OECD, 2021c). Thus, fewer instances will require the third test to be conducted. Also, during the validation of the 2o3 DA at the OECD, it was clearly shown that the sequence of testing does not affect the outcome of the 2o3 DA. Two non-BL negative results lead to a negative call (Scenario 1), while two non-BL positive results are sufficient for classification as a sensitizer (Scenario 2). If the chemical is within the applicability domain (AD) of the kDPRA, conducting the kDPRA provides information on whether the chemical must be subclassified as 1A. The combined concentration-response information from a positive kDPRA and a positive KS is then applied in the regression model in the standardized prediction spreadsheet using EQ1 to derive the PoD. However, if the chemical is not within the AD of the kDPRA (Scenario 3a), it is recommended to perform the h-CLAT to gather more evidence on potency by applying EQ6. According to the 2o3 scheme, if either the DRPA or KS was negative or BL, the h-CLAT must be conducted. Two negative, non-BL outcomes again indicate a non-sensitizer (Scenario 5), and a BL outcome leads to an inconclusive assessment (Scenario 6). A positive h-CLAT with a positive result from either KS or DRPA leads to classification. If the DRPA and the h-CLAT are positive (Scenario 4), chemicals within the AD of the kDPRA can then be subclassified based on the kDPRA and assessed for PoD with regression model EQ4.

If the DRPA is negative, two positives in KS and h-CLAT can lead to classification (Scenario 3b), and a PoD can be derived based on EQ6. In this case, a subclassification of 1B can be made directly: a negative call in the DRPA, and hence, a negative call in the kDPRA is sufficient for chemicals to be classified as 1B. (Note: This is also consistent with the outcome from the alternative validated DA, ITS, whereby a chemical negative in the DRPA is not classified as a 1A sensitizer, as it cannot reach a score of 6 or 7 (OECD, 2021a)). However, in Scenario 3a (i.e., a chemical not in the AD of the kDPRA that is positive in the DRPA), the GHS subclassification is inconclusive. In this case, the outcome of the PoD with EQ6 may still be used for a WoE assessment to indicate whether the LLNA potency is predicted to be at an EC3 < 2%. However, according to the OECD guideline, this would not be sufficient for a conclusive 1B classification.

For Scenario 6, i.e., an inconclusive outcome of the 2o3 due to BL results, the result can be due to BL negative results. In this case, no relevant PoD can be calculated as no EC1.5 or MIT or reaction rate is derived from the BL tests. On the other hand, if the result is BL positive, EC1.5 or MIT values are available, and a PoD can be calculated but has a lower certainty. These values were still given in ESM1 (13 cases), as the OECD guideline states that borderline outcomes could still be used in a weight-of-evidence.

This proposed testing sequence might be further simplified if chemical reactivity is expected, e.g., based on structural alerts. Then the kDPRA could be directly done instead of the DRPA. A positive result in the kDPRA (> 13.89% Cys peptide depletion) may then be used as a positive rating along with a positive result from KS and/or h-CLAT. If the kDPRA result was negative, the DRPA would still need to be conducted to confirm the negative result. This approach may save tests if a chemical has a high likelihood of a positive outcome in the (k)DPRA.

#### 3.2 Alternative testing sequences

**All data is generated**

The tiered economic testing strategy in Figure 1 with conditional testing in h-CLAT based on the outcome of the first two tests may be considered time-consuming by some users. An alternative option is to test a new chemical directly in KS, h-CLAT and DRPA by default. If two tests are positive, and one is the DRPA, the kDPRA is conducted. In this case, the hazard ID and the GHS sub-
In this case, the PoD is more frequently derived with EQ4 instead of EQ1, as all chemicals positive in the first two assays (i.e., h-CLAT and DPRA) will be assessed based on EQ4.

### 3.3 PoD outcome for chemicals with available KS, h-CLAT and kDPRA data

Table 1 summarizes the prediction accuracy for the three different testing sequences, namely (i) prediction of the PoD according to Figure 1, (ii) prediction based on EQ5 / EQ6 in cases when all data are generated, and (iii) with h-CLAT done first (Fig. ESM2-12). The individual predictions and the scenario/equation used for each chemical with these three approaches are given in ESM11, along with the correlation between the different assessments for each chemical (Fig. ESM11, 1-3). In all three cases, the prediction accuracy is quite similar and leads to a comparable number of > 5-fold or > 10-fold (i.e., a full potency class)
mispredictions vs. the LLNA result. As is obvious from Table 1, the scatter plots (Fig. ESM1 1-6), and the data on the individual chemicals in ESM1, for most chemicals, the predicted PoD are similar when using the different testing sequences, and there is no tendency that one testing sequence is, in general, less conservative. The number of overpredicted chemicals, however, is lower when using all evidence, as the negative evidence for chemicals positive in only two assays is taken into account, and this approach (EQ5) therefore also leads to a slightly better correlation with in vitro data (see Fig. ESM11, 4-6).

### 3.4 Analysis of significant over- and under-predictions

To analyze individual mispredictions, we focused on the outcome of the testing sequence in Figure 1. Table ESM3-1 lists all the chemicals that are > 5-fold underpredicted, i.e., their potency as assessed by the LLNA is significantly higher than the predicted PoD. The chemicals in this Table are grouped, and an individual discussion is given for each chemical. In summary, a set of 6 chemicals is underpredicted as weak sensitizers with predicted EC3 of 9.2%-55%, while they are moderate sensitizers in the LLNA. These include *inter alia* primary amines/pro-haptens and amine-reactive chemicals, which are outside of the AD of the kDPRA (OECD, 2021b). For a larger group (n = 12), the predicted PoD indeed indicates a significant sensitization potency (predicted EC3 0.05%-5%), but the individual values are clearly below the strong to extreme potency observed in the LLNA. This indicates that the dynamic range for the exact potency assessment of some extreme sensitizers using the regression models is limited. However, a high sensitization potential is predicted for most chemicals in this group based on the in vitro data.

Table ESM3-2 provides data and discussion on the chemicals with a predicted PoD below the LLNA EC3, i.e., a higher potency is predicted in vitro. This group contains six false positives in the 2o3 vs. LLNA outcome. For four of those, positive human sensitization evidence or a strong alkylating potency indicate that the LLNA actually underpredicts the sensitization potential. In contrast, for two others, the reported human sensitization potential is rather weak (propyl paraben and benzocaine) and clearly overrated by the in vitro approach. A further group (n = 5) contains very reactive and volatile chemicals. Although EQ1 corrects for high volatility, it does not fully predict the weak sensitization in LLNA observed for these highly reactive chemicals that evaporate rapidly under LLNA conditions (see supplementary data file 1 in Natsch et al., 2015). However, these chemicals may be significantly more potent under (partial) occlusion or when present in a product limiting evaporation. Hence, this conservative assessment by the in vitro derived PoD may be appropriate. Another set of chemicals (n = 5) is clearly overpredicted when assessed vs. LLNA data, but either clinical data or human repeat insult patch tests indicate that these are very relevant human sensitizers, and the in vitro prediction could better reflect the human sensitization potency. No human data are available for the remaining seven chemicals, but they include several highly reactive chemicals.

The analysis in Table 1 and in ESM3 is based on a comparison with the comprehensive historical LLNA database. An additional analysis was conducted based on the OECD curated EC3 values, taking the historical database values only where no curated EC3 was available. This analysis is shown and compared to the above analysis in ESM2. The outcome of both analyses is almost congruent.

### 3.5 Hazard ID and GHS subclassification outcome for chemicals in the OECD database

If the kDPRA is combined with the 2o3 DA in a testing strategy, chemicals can be rated both for hazard and for GHS potency class. As indicated above, this is independent of the testing sequence with all three testing proposals leading to the same outcome. In Table 2, we show the outcome of the classification rating on the
4 Discussion

The 2o3 DA has been accepted as an OECD standard for hazard ID. At the same time, the kDPRA can be used as a stand-alone test for GHS subclassification once a chemical is identified as a skin sensitizer. Thus, combining these two approaches for classification and subclassification, as illustrated here, is a straightforward strategy. This combination will not require further validation for both the hazard and the subclassification decision as both prediction models were validated and implemented in the OECD TG 497 and 442C for chemicals considered within the AD (OECD, 2021a,b).

For this classification approach, only the positive/negative answers from the validated prediction models in KS/h-CLAT/DPRA or the validated binary classification according to a quantitative threshold (Log $k_{\text{max}} = -2$) in the kDPRA are used. However, the data generated are more granular (quantitative kinetic rate constant over several orders of magnitude in kDPRA and concentration-response data over three orders of magnitude in the cell-based assays). As shown in the parallel analysis (Natsch and Gerberick, 2022), this concentration-response data can be used to estimate a PoD. Thus, the same test results generated for the GHS (subclassification can be used for the potency assessment and to derive a PoD in the integrated testing and assessment sequences provided here.

The different sequences can start with either of the two cell-based assays or generate data with all three tests as a default. Different predictive equations can be applied for PoD determination depending on the generated data. The analysis of the outcome for the individual chemicals indicates that the different testing sequences using other predictive equations overall lead to surprisingly similar predictions (Fig. S2-S4). This confirms previous observations on data redundancy especially between quantitative data from h-CLAT and KS (Natsch et al., 2015; Natsch and Gerberick, 2022). Still, it also indicates that the different testing sequences are all valid approaches and neither of them leads to an overall less conservative risk assessment. Since the various in vitro assessments correlate better with each other than with the in vivo data (see Fig. S2-S7 in ESM2), the key open question is whether other in vitro assays will provide further, more orthogonal information for a further improved PoD determination, or whether this asymptotic fit to in vivo data when adding more in vitro information also partly reflects the limitations of the in vivo data source.

In any case, it is of the utmost importance to understand the sources of uncertainty in the in vitro and in vivo datasets. Part of the uncertainty comes from the biological variability of both the LLNA and the in vitro data. For the LLNA, analysis of repeated studies (Dumont et al., 2016; Hoffmann, 2015) indicates that the typical standard deviation of EC3 values is 1.8-fold in either direction, but larger discrepancies were noted in some cases. This will lead to uncertainty of the in vivo comparator, especially in instances where only one LLNA study is available. Biological variability in the in vitro data (Gabbert et al., 2020; Leontaridou et al., 2017) will further increase uncertainty, and therefore, variability in both datasets will always limit the fit between them.

However, this data variability can only explain part of the prediction inaccuracy. A further part of the uncertainty is that the in vitro tests are not yet a perfect reflection of the sensitization process, as they all only measure surrogates of some key events (e.g., no T cell activation). On the other hand, as illustrated by the detailed analysis of the individual chemicals with > 5-fold misprediction, part of the inaccuracy may also be because the LLNA is not a perfect model of potency for all chemicals, reminding us that the LLNA itself measures only part of the sensitization process (antigen-presentation triggered cell proliferation in the lymph node). Thus, for some chemicals that are negative in the LLNA but positive in the in vitro assessment, data from human studies and/or the alkylation potential observed in peptide reactivity studies indicate that the LLNA may be false-negative, and

---

**Tab. 2: GHS sub-classification of the chemicals in the OECD database by the 2o3 DA combined with kDPRA**

<table>
<thead>
<tr>
<th>Prediction 2o3 DA with kDPRA</th>
<th>LLNA result</th>
<th>NC (n = 26)</th>
<th>1B (n = 85)</th>
<th>1A (n = 38)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>21</td>
<td>16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1B</td>
<td>3</td>
<td>34</td>
<td>7 (4)</td>
<td></td>
</tr>
<tr>
<td>1A</td>
<td>1</td>
<td>14</td>
<td>26 (29)</td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>84%</td>
<td>53%</td>
<td>79% (88%)</td>
<td></td>
</tr>
<tr>
<td>Underpredicted</td>
<td>NA</td>
<td>25%</td>
<td>21% (12%)</td>
<td></td>
</tr>
<tr>
<td>Overpredicted</td>
<td>16%</td>
<td>22%</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Inconclusive</td>
<td>n = 8</td>
<td>n = 21</td>
<td>n = 5</td>
<td></td>
</tr>
</tbody>
</table>

*The values are based on applying only the prediction model of the kDPRA and 2o3 DA. The values calculated when taking the applicability domain (AD) of the kDPRA into account and applying Scenario 3a in Figure 1 (using EQ6 for chemicals outside of AD of kDPRA) are given in brackets.*
the in vitro result may give a more accurate estimation of the sensitization risk. Similarly, several of those chemicals for which the potency is overestimated by the in vitro PoD are critical skin sensitizers from human clinical studies, especially some preservatives and glove allergens. When analyzing the underpredictions, on the other hand, the in vitro PoD appears not to perfectly cover the dynamic range for very potent sensitizers. Thus, it is noteworthy that some of the extreme sensitizers are predicted as strong sensitizers based on the PoD, but the predictive models do not yet reflect their full potency in the LLNA.

Turning to the GHS classification and subclassification outcome, the predictivity is better for predicting non-sensitizers and strong (1A) sensitizers in the LLNA, and the predictivity for the LLNA 1B sensitizers is less accurate with around 22-25% mispredictions in either direction. While a more limited predictivity for the intermediate class (where misprediction to either side is possible) is an intrinsic property for any three-way classification scheme, the absolute number of correct classifications may be considered relatively low. Here we thus provide a detailed analysis for the individual mispredicted chemicals regarding the GHS classification (ESM4). Next to general limitations of prediction accuracy based on data variability discussed above, some of the predictive limitations for correct classifications can be attributed to (i) limitations of the applicability domain (AD) of the in vitro assays and partial coverage of key events, (ii) only partial coverage of the human sensitization potential and potency by the LLNA model, (iii) the fact that some in vivo and in vitro results are very close to the decision threshold (LLNA EC3 of 2% / kDPRA threshold of Log $k_{\text{max}} = -2$).

The kDPRA has an important weight in the potency determination (Natsch and Gerberick, 2022). Thus, it is critical to assess whether a chemical is in the AD of the kDPRA. The OECD TG indicates that test chemicals with exclusive lysine-reactivity as observed in Dpra or ADRA are outside of the AD of the kDPRA as the kinetic reactivity with lysine residues is covered neither by the kDPRA nor the testing schemes shown here. Such chemicals, if positive in both KS and h-CLAT, may still be assessed with the regression models. Thus, the PoD for glutaraldehyde – a chemical not in the AD of the kDPRA and mispredicted for potency using kDPRA only – is predicted based on EQ6 with a PoD of 0.6%, which is still higher than the LLNA EC3 of 0.1% but in the correct GHS class. For another amine-reactive chemical, 3,4-dihydrocumarin potency is underrated. While it is possible to measure amine reactivity of these chemicals, it may be a significant challenge to derive quantitative potency models based on the limited number of typical amine reactive chemicals as a training set (with the exception of aldehydes, for which we have provided a model (Natsch et al., 2018)). A second limitation indicated for the kDPRA is “aromatic amines, catechols or hydroquinones”, which may require further data to confirm their weak reactivity if their Log $k_{\text{max}} < -2$. Thus, there are two cases among the seven mispredicted chemicals rated as 1B instead of 1A (1,4-phenylenediamine and 2-amino-phenol) that are rated as 1A if EQ6 is applied.

Next to considering the applicability of the in vitro tests, it is also key to look at a WoE when assessing the wrong in vitro classifications vs. the LLNA outcome. Thus, the analysis of the LLNA data as performed by the OECD data review indicated a limitation of the LLNA for specificity vs. human data (Natsch et al., 2021; OECD, 2021a). This is partly because the review criteria required a higher maximal test concentration to conclude on a negative call in the LLNA as compared to the validation of the LLNA (Kolle et al., 2020). Also, the estimate of specificity vs. human data is based on a relatively low number of chemicals, but it indicates that the database does contain some false-positive chemicals in the LLNA. Indeed, among the 16 FN in 2o3 vs. LLNA data, there are seven chemicals for which the WoE indicates that they are not, or extremely weak, human sensitizers (ESM4). On the other hand, among the over-predicted chemicals, as discussed above for the PoD, the sensitization potency and correct GHS class could be underestimated by the LLNA for several cases and could be more correctly reflected by the in vitro PoD (ESM4).

The integrated assessment discussed here is solely based on in vitro data from the three OECD TG, and no in silico assessment is integrated into this approach, differently from almost all published approaches for an integrated evaluation of the sensitization potential (Del Bufalo et al., 2018; Hirota et al., 2018; Jaworska et al., 2015; Macmillan and Chilton, 2019; Strickland et al., 2017; Takenouchi et al., 2015). There are some benefits to the present approach of conducting an assessment based solely on validated OECD test methods and not, from the start, integrating an in silico prediction: (i) Most in silico tools were developed and trained partly on the database with available in vitro and in vivo data, and rule-based approaches based on structural alerts in principle have an unlimited number of degrees-of-freedom. Using in silico tools on the same database without separating test and training set may thus lead to an overfitted model. This problem is minimal for the PoD models used here as they are based on only 3-6 input variables and trained on >180 chemicals. (ii) When conducting an assessment solely based on in vitro data, an independent, parallel assessment can then be made applying the in silico tools to increase certainty and obtain a more holistic picture. If the in silico tool is already integrated into the initial prediction, this is not possible without double-accounting. (iii) In silico tools as implemented, e.g., in OECD TG 497 (OECD, 2021a) have a relatively strict definition in their AD for known chemical features, especially to make conclusive negative predictions. Thus, using an in silico tool by default has implications on the overall AD for new chemicals to be assessed. Approaches to perform a WoE assessment on existing chemicals have been described using only human data (Basketter et al., 2014) or combining human, animal, in vitro and in silico data. For new chemicals, the human and animal part would be lacking, but the here proposed integrated in vitro approach can then be combined with parallel in silico predictions for a WoE.

Here we focused the analysis with regard to the LLNA outcome. When assessing hazard ID, looking at the human data is important (Natsch et al., 2021; OECD, 2021a) as the LLNA may also have limitations in specificity if used at too high concentrations or if not taking irritation into account as indicated above. However, quantitative human data on potency is available only for a minority of chemicals, and since we are discussing how to combine potency assessment into the 2o3 DA, the key analysis...
presented here was performed vs. the LLNA potency data. Nevertheless, we indicate in the discussion on individual chemicals (ESM3 and ESM4) the semi-quantitative potency information from human data when available (Api et al., 2017; Basketter et al., 2014; OECD, 2021e) as these data further help to assess in which cases the in vitro data truly underestimate potency but also highlight cases where the NAM assessment may lead to a more correct and more conservative human risk assessment.

The proposed testing sequences for (sub)classifications and PoD determination are a proposal to make the best use of the data generated by testing according to TG 442C, 442D and 442E. The PoD could be used directly in risk assessment, and in the absence of other evidence, a default assessment factor may be introduced to account for uncertainty (Natsch et al., 2015). As risk assessors transition to using NAM data for potency assessment, a PoD derived from these regression models could be integrated into existing risk assessment schemes such as quantitative risk assessment (QRA) (Api et al., 2020). However, the assessment certainly does not stop there: Analysis of the prediction accuracy of close analogues with both in vitro and in vivo data will help refinement of the uncertainty for specific chemicals (Natsch et al., 2018). The large database provided in this and the parallel analysis (Natsch and Gerberick, 2022) and the increasing database from other initiatives will further help to investigate in which chemical domains certainty is higher or lower and will provide read-across analogues to conduct such an uncertainty analysis in the specific chemical domain of the molecule to be assessed. Furthermore, depending on the chemical domain, further non-guide-line methods can be applied to test specific parameters, such as metabolic activation by metabolic systems, reactivity with amine groups, or epidermal disposition. Such further evidence can then refine the PoD derived from the presented standard testing sequences.

Electronic supplementary material

ESM1 provides the in vitro and in vivo data in Sheet 1; Sheet 2 provides all the individual predictions and fold-mispredictions for 188 chemicals with the three different testing sequences; Sheet 3 shows the graphical correlations between the different predictions and between predictions and in vivo results.

ESM2 provides the testing sequence starting with h-CLAT and comparison of predictions with OECD curated LLNA values.

ESM3 discusses > 5-fold misprediction vs. LLNA outcome also considering other (e.g., human) evidence

ESM4 discusses GHS-misclassifications also considering other (e.g., human) evidence.

References


08444.2018.1429385


OECD (2021b). Test No. 442C: In Chemico Skin Sensitisation Assays addressing the Adverse Outcome Pathway, key event on covalent binding to proteins. OECD Testing Guidelines. doi:10.1787/9789264229709-en


Conflict of interest

The authors declare no competing interests.

Data availability

All data of this publication are made publicly available, and all models and tests used are freely available. All input data are available in ESM1 to this manuscript on the first datasheet. All individual predictions are included in ESM1 on the second spreadsheet.