Off to a Good Start? Review of the Predictivity of Reactivity Methods Modelling the Molecular Initiating Event of Skin Sensitization

Nathalie Alépée1, Fleur Tourneix1, Akanksha Singh2, Nadège Ade3 and Sébastien Grégoire1

1 L’Oréal, Research & Innovation, Aulnay-sous-Bois, France; 2 L’Oréal, Research & Innovation, Bengaluru, India; 3 EpiSkin SA, Lyon, France

Abstract
The assessment of skin sensitizing properties of chemicals has moved away from animal methods to new approach methodolo-
gies (NAM), guided by qualitative mechanistic understanding operationalized in an adverse outcome pathway (AOP). As with any AOP, the molecular initiating event (MIE) of covalent binding of a chemical to skin proteins is particularly important. This MIE has been modelled by several test methods by measuring the reaction of a test chemical with model peptides in chemico. To better understand the similarities and differences, a data repository with publicly available data for the direct peptide reactivity assay (DPRA), amino acid derivative reactivity assay (ADRA) and kinetic DPRA (kDPRA), as well as the peroxidase peptide reactivity assay (PPRA) was assembled. The repository comprises 260 chemicals with animal and human reference data, data on four relevant physicochemical properties, and between 161 to 242 test chemical results per test method. First, an overview of the experimental conditions of the four test methods was compiled allowing to readily compare them. Second, data analyses demonstrated that the test methods’ predictivity was consistently reduced for poorly water-soluble chemicals and that the DPRA and ADRA can be used interchangeably. It also revealed new categorization thresholds for the DPRA and ADRA that are potentially relevant for strategic uses. In summary, a detailed assessment of reactivity test methods is provided, highlighting their potential and limitations. The results presented are intended to stimulate scientific discussion around test methods modelling the MIE of the skin sensitization AOP.

1 Introduction
The assessment of skin sensitizing properties of chemicals has moved away from animal methods to new approach methodolo-
gies (NAM). This development has been guided and facilitated by a qualitative understanding of the toxicological events leading to the induction of skin sensitization. This mechanistic knowledge has been operationalized by structuring it in an adverse outcome pathway (AOP) (OECD, 2014).

The AOP’s molecular initiating event (MIE), or key event 1 (KE1), i.e., the initial interaction between a chemical and a biotar-get that can be causally linked to an outcome via a pathway (Allen et al., 2014, 2016), is covalent binding of a chemical to skin protein (Natsch and Emter, 2017). While there is high certainty associated with the biological plausibility and the essentiality of the MIE, uncertainty is introduced when modelling the MIE. As proteomic approaches, e.g., in skin models, are technically extremely challenging and complex, the experimental approach taken was to model the peptide reaction in chemico using various synthetic peptides, of which those containing cysteine and lysine emerged as the most promising (Gerberick et al., 2004, 2007). The approach described by Gerberick et al. (2007) was subsequently referred to as the direct peptide reactivity assay (DPRA). The DPRA successfully underwent formal validation by the European Centre for the Validation of the Alternative Methods (ECVAM) of the European Commission from 2009 to 2013, which is extensively documented in the “Tracking system for alternative methods towards regulatory acceptance”1. In 2022, the Organisation for Economic Cooperation and Development (OECD) adopted the DPRA as test guideline (TG) 442C for the purpose of discriminating sensitiz-
ers from non-sensitizers (OECD, 2023a). The DPRA provided the role model for later developments that led to major amendments to the guideline, i.e., the addition of the amino acid derivative reactivity assay (ADRA), a test method very similar to the DPRA and for the same regulatory purpose, and the addition of the kinetic DPRA (kDPRA), a test method optimized to discriminate UN Globally Harmonised System (GHS) subcategory 1A skin sensitizers from UN GHS subcategory 1B skin sensitizers and chemicals assigned “no category” according to UN GHS (Fujita et al., 2014; Wareing et al., 2017).

Furthermore, modifications of the DPRA were developed to address potential shortcomings. Of those developments, we have included the peroxidase peptide reactivity assay (PPRA), which uses horseradish peroxidase to potentially activate chemicals needing oxidative or metabolic activation and for which a substantial amount of data is available (Gerberick et al., 2009; Hoffmann et al., 2022; Ryan et al., 2020). However, the PPRA has not yet been proposed for an evaluation in a multicentric validation study nor for inclusion in OECD TG 442C. Another evolution, which is not addressed here, is the development of high-throughput DPRA versions (Cho et al., 2019; Wei et al., 2021). These developments came with variations of the experimental approach. While the reactivity of the test chemical with model peptides remains the fundamental principle, the experimental conditions of the test methods differ substantially. Therefore, one aim of this review was to summarize the experimental approaches for the DPRA, ADRA, kDPRA and PPRA, providing a comprehensive overview and highlighting similarities and differences for users.

Another aim was to approach the uncertainties associated with the peptide reactivity test methods, such as model uncertainty, which, at least to some extent, causes a lack of a quantitative interpretability and limitations of the applicability domain as well as aspects of data interpretation. Addressing such uncertainties presents challenges. Besides expert knowledge, it requires empirical evidence to identify issues and to explore potential solutions to reduce or to (semi-)quantify uncertainties. Therefore, one aim of our work was to compile a central repository for peptide reactivity test method data. We used it to investigate general reactivity data patterns related to physicochemical properties of the chemicals in the repository to better inform their use, e.g., for application in defined approaches (DA) and next generation risk assessments (NGRA) (Gilmour et al., 2020).

Despite those uncertainties, the DPRA is generally considered valuable in the assessment of skin sensitization. This becomes specifically evident in its ubiquitous use in DA. It is not only an integral part of the hazard and UN GHS potency classification DA included in OECD Guideline 497 (OECD, 2021a), but also in many other DA, some of which aim at risk assessment (Kleinsteuer et al., 2018). As the ADRA is similar to the DPRA, we also investigated whether there are data differences that might suggest different uses, or whether the DPRA and ADRA can be used interchangeably.

Similar to Natsch and Ente (2017), our intention is to continue and stimulate discussions in the scientific and risk assessment communities about the reactivity test methods informing the MIE of the skin sensitization AOP.

2 Methods

Reactivity test methods

Four reactivity test methods, for which a substantial amount of data is available, were considered. These comprise the three test methods described in OECD TG 442C, i.e., the DPRA, the ADRA, and the kDPRA (OECD, 2023a), as well as the PPRA. All test methods have been well described and the reader interested in experimental details is referred to the following main references: DPRA (Gerberick et al., 2007), ADRA (Fujita et al., 2014), kDPRA (Wareing et al., 2017), and PPRA (Ryan et al., 2020).

In brief, all test methods measure the relative depletion of one (cysteine) or two (cysteine and lysine) model peptides caused by a test chemical after 24 hours or at several time points using one or more test concentrations and employing different analytical techniques. The data are interpreted with test method specific prediction models that have been optimized for the intended (regulatory) purpose. We compiled the essential experimental information, such as the identity and concentrations of the nucleophiles, the analytical approach, and the prediction model, in a tabular format to provide a comprehensive overview. This table facilitates comparison of the four test methods and allows to identify similarities and differences.

Data repository

The collection of data for the four reactivity methods resulted in inclusion of 260 chemicals, which are identified by name, CAS registry number, and SMILES. In addition to the primary output data for the four test methods we retrieved from Hoffmann et al. (2022), Imamura et al. (2021) and Natsch and Gerberick (2022), the data repository includes skin sensitization binary or categorical predictions (for DPRA and PPRA only) of skin sensitization potential and categories according to the standard prediction models of the reactivity tests (Gerberick et al., 2007; OECD, 2023a; Ryan et al., 2020).

For reference, data from the standard in vivo test, the local lymph node assay (LLNA), and human data were added, which were retrieved from Hoffmann et al. (2022) and the OECD reference database (OECD, 2021b). From the same databases, information on the simple transformation (pre-hapten) and metabol-

---

**Abbreviations**

ADRA, amino acid derivative reactivity assay; AOP, adverse outcome pathway; Cys, cysteine peptide; Cys+, cysteine depletion in the presence of horseradish peroxidase; DA, defined approach; DPRA, direct peptide reactivity assay; EC50 value, effective concentration inducing a stimulation index of 3; FN, false negative; FP, false positive; KE, key event; kDPRA, kinetic DPRA; LLNA, local lymph node assay; MW, molecular weight; NAM, new approach methodologies; NGRA, next generation risk assessment; NS, non-sensitizer; PM, prediction model; MIE, molecular initiating event; OECD, Organization for Economic Co-operation and Development; PPRA, peroxidase peptide reactivity assay; S, sensitizer; SI, stimulation index; TN, true negative; TP, true positive; UN GHS, United Nations Globally Harmonised System; WS, water solubility
Tab. 1: Comparison of the experimental conditions of the DPRA, kDPRA, PPRA and ADRA

<table>
<thead>
<tr>
<th>Assay</th>
<th>Principle incl. endpoint and incubation time</th>
<th>Nucleophile (concentration)</th>
<th>Chemical/ nucleophile ratio</th>
<th>Solvents</th>
<th>Analytical technique</th>
<th>Prediction</th>
<th>Technical limitations</th>
<th>Chemistry-related limitations</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DPRA</strong></td>
<td>% depletion of peptide (Cys/Lys) after 24 h incubation with test chemical</td>
<td>Ac-Ala-Phe-Ala- Ala/ Cys-Ala-Ala- COOH (500 μM) and Ac-Ala-Phe-Ala- Ala- Lys-Ala-Ala- COOH (500 μM)</td>
<td>1:10 (Cys) and 1:50 (Lys)</td>
<td>Acetonitrile, water, 1:1 mixture water: acetonitrile, isopropanol, acetone or 1:1 mixture acetone: acetonitrile, 1:9 dimethyl sulfoxide: acetonitrile</td>
<td>HPLC-UV (absorbance at 220 nm)</td>
<td>– Cys and Lys Depletion (%): NS/NR &lt; 6.38 % and S/R otherwise – Cys depletion only (%): NS/NR &lt; 13.89 % and S/R otherwise – prediction model for 4 categories is also available</td>
<td>– Cys peptide dimerization – coelution – solubility at 100 mM concentration</td>
<td>– Pro-haptens – Metals and inorganics – Insoluble non-reactives – Mixtures with unknown composition – Oxidants</td>
<td>Gerberick et al. (2007) and OECD (2023a)</td>
</tr>
<tr>
<td><strong>kDPRA</strong></td>
<td>% depletion of peptide (Cys) as a function of time (10, 30, 90, 210, 1440 min.) (LogKmax) and test chemical concentration</td>
<td>Ac-Ala-Phe-Ala- Ala/ Cys-Ala-Ala- COOH (500 μM)</td>
<td>1.0, 0.625, 1.125, 1.25, 1.5, 1.10</td>
<td>Acetonitrile, pH 7.5 phosphate buffer</td>
<td>Fluorometer (λex = 390 nm; λem = 480 nm)</td>
<td>LogKmax &gt; -2 Cat1A LogKmax &lt; -2 Cat1B/not classified</td>
<td>– Cys peptide dimerization – fluorescence quenching – auto fluorescence – solubility</td>
<td>– Pro-haptens – Metals and inorganics – Insoluble non-reactives – Mixtures with unknown composition – Hydroquinones, catechols and aromatic amines – Thiols or thiol-releasers – Exclusively Lys-reactive chemicals – Oxidants</td>
<td>Wareing et al. (2017) and OECD (2023a)</td>
</tr>
<tr>
<td><strong>PPRA</strong></td>
<td>% depletion of peptide Cys ± HRP/P and Lys after 24 h incubation with test chemical (DPmax)</td>
<td>Ac-Ala-Phe-Ala- Ala/ Cys-Ala-Ala- COOH (20 μM) and Ac-Ala-Phe-Ala- Ala- Lys-Ala-Ala- COOH (5 μM)</td>
<td>1.0, 0.08, 1.0, 0.4, 1.2, 1.10 (Cys and Lys)</td>
<td>Acetonitrile, methanol, ethanol, isopropanol, water, dimethyl sulfoxide, and acetone</td>
<td>HPLC-UV (absorbance at 220 nm)</td>
<td>DPmax ≤ 15.1% (Cys ±), ≤ 10% (Lys) = minimally reactive; DPmax &gt; 25% (Cys ±), &gt; 20% (Lys), EC25 ≤ 0.1 mM = reactive; otherwise: interpret data prediction model for 3 categories is also available</td>
<td>– Cys peptide dimerization without HRP/P – coelution – solubility</td>
<td>– Metals and inorganics – Insoluble non-reactives – Mixtures with unknown composition</td>
<td>Ryan et al. (2020)</td>
</tr>
</tbody>
</table>
ic (pro-hapten) activation potential of chemicals was obtained, mainly based on information from Urbisch et al. (2015) and Patlewicz et al. (2016).

Potentially relevant physicochemical properties were selected: molecular weight (MW), water solubility (WS), octanol-water partition coefficient (KOW), and volatility classes. MW was determined from the structural information. For salts, the MW was based on the free acid/base. The WS, expressed as mol per liter (M) was determined using the commercial software VolSurf+, version 9.38. The KOW was estimated using the clogP version 5.2 commercial software from Biobyte and is expressed as the logarithm (logKOW). The volatility class prediction was based on the freely available EPI Suite version 4.1. vapor pressure prediction. To assign chemicals to volatility classes, the approach proposed by Spicer et al. (2002) was modified based on the observation that mass balance does not seem to be affected for chemicals having vapor pressure lower than 1.3*10^{-4} mmHg (Grégoire et al., 2019). The following volatility classes were defined:

- very volatile with vapor pressure higher than 1 mmHg (class 3)
- volatile with vapor pressure between 1*10^{-2} and 1 mmHg (class 2)
- semi-volatile with vapor pressure between 10^{-2} and 10^{-4} mmHg (class 1)
- non-volatile with vapor pressure below 10^{-4} mmHg (class 0)

Data analysis

The data were primarily analyzed descriptively. Descriptive, correlation and significance testing statistics as well as most graphical data representation were conducted with GraphPad Prism, version 9.4.1. The overlap of chemicals for which reactivity data were available was graphically summarized using a free Venn diagram application^2. Receiver-operation characteristic (ROC) curves were used to compare the predictivity of test methods.

3 Results

3.1 Comparison of the reactivity test methods’ experimental conditions

The main experimental conditions of the four in chemico test methods were compared (Tab. 1). All measure the depletion of one or two synthetic peptides (nucleophiles) as an indicator of a chemical’s potential to covalently bind to skin proteins. The test methods DPRA, kDPRA, and PPRA use the same synthetic cysteine peptide. However, the DPRA and the PPRA also include the same synthetic lysine peptide. The ADRA uses amino acid derivative NAC and NAL as the test nucleophiles.

The DPRA, ADRA, and PPRA measure depletion at one time-point only (24 hours), while the kDPRA determines the depletion at six timepoints. In contrast to DPRA and ADRA, which test only one concentration, the kDPRA and the PPRA use five, respectively four concentrations. The PPRA is the only test method that includes an activation of test chemical by measuring cysteine depletion in the presence of horseradish peroxidase, denoted as Cys+. The chemical/nucleophile ratio ranges from 1:50 to 1:1, being lowest for the ADRA and DPRA lysine peptide.

Remaining peptide is determined either by HPLC-UV or by fluorometer. The use of different nucleophiles allowed to establish both analytical techniques for the ADRA.

---

^2 [http://www.interactivenn.net](http://www.interactivenn.net)
3.2 The database

Drawing from the various data sources, a total of 260 chemicals, each with data from at least one reactivity test method, were included in the data repository. We used a Venn diagram to illustrate how many chemicals have been tested in the individual test methods and how the chemicals overlap for the four test methods (Fig. 1). The most data was available for the DPRA (for 242 chemicals), the least for the ADRA (161 chemicals). Figure 1 shows that 91 chemicals had data for all test methods. They are marked in Table S1 and are in the following referred to as the “core chemical set”. The second largest group were 57 chemicals tested in the DPRA, ADRA and kDPRA, followed by 43 tested in the DPRA and PPRA, which were mainly cosmetic ingredients and structurally different from the other chemicals (Hoffmann et al., 2022).

When compared pairwise, the overlap was smallest for ADRA and PPRA (92 chemicals) and largest for DPRA and kDPRA (184 chemicals). LLNA data were available for 256 chemicals, which comprised 72 non-sensitizers (UN GHS No Cat.), 121 weak to moderate sensitizers (UN GHS Cat. 1B), and 63 strong to extreme sensitizers (UN GHS Cat. 1A). The core chemical set consisted of 26 UN GHS Cat. 1A, 48 UN GHS Cat. 1B and 17 non-sensitizers. For four chemicals, no LLNA data but a guinea pig maximization test (GPMT) was available. Among the 260 chemicals, 30 were identified as pre-haptens and/or pro-haptens.

Depending on the complexity of the experimental design, the data are processed differently. The DPRA and ADRA, with a single measurement per peptide, do not require any data summary. Among the parameters determined by the PPRA, the maximum depletion across all tested ratios per peptide and with activation is used for further interpretation. The kDPRA calculates the reaction kinetic constants for the timepoints and uses the logarithm of the maximum (logKmax) for data interpretation. Binary prediction models (PM) are available for all four test methods. While the PM of the DPRA, ADRA, and PPRA are intended to discriminate non-sensitizers (UN GHS No Cat.) from sensitizers (UN GHS Cat. 1B and 1A), the kDPRA PM was developed to discriminate UN GHS Cat. 1A from the other two categories. In addition, categorical PM are available for the DPRA and PPRA.

The test methods have basically similar technical and chemistry-related limitations (Tab. 1). However, each method has specific advantages. The DPRA is the best-established, has the longest use history, and has been explored in combination with in vitro test methods addressing AOP key events 2 and 3. PPRA is particularly suited to test chemicals that can be activated. The ADRA avoids the issue of coelution by measuring depletion via fluorescence and is less affected by poor solubility due to the low amount of test chemical used.

Fig. 1: Representation of the number of test chemicals tested per test method

The intersection in the center indicates that 91 chemicals were tested in all test methods (core chemical set).
3.3 Analyses of physicochemical properties
The chemicals included a broad spectrum of physicochemical properties, as shown in Figure S1. An analysis of pairwise inter-dependencies of the four properties by non-parametric correlation according to Spearman, which was chosen due to particular properties in the distributions, showed – as expected – a high negative correlation (correlation coefficient of -0.854) of WS and logKOW. WS was negatively correlated and logKOW positively correlated with MW. WS and logKOW were at best weakly correlated with volatility class, while MW and volatility class were clearly negatively correlated (correlation coefficient of -0.659).

Next, we analyzed any potential relation between the individual physicochemical properties and the LLNA result expressed as sensitizer and non-sensitizer, as UN GHS categories, and as EC3 values (see Tab. S1 for data). Apparent tendencies were kept in mind for subsequent analyses comparing the reactivity test method results with the LLNA data. Figure 2 demonstrates the distribution of the individual physicochemical properties by LLNA result expressed as sensitizer and non-sensitizer. While there were no statistically significant differences (Mann-Whitney test, significance level 0.05) for MW and volatility, WS was significantly higher and logKOW significantly lower for non-sensitizers. The results for UN GHS categories and EC3 values are shown in Figures S2 and S3, respectively.

3.4 Global predictivity of test methods
We analyzed the predictivity of the LLNA result of the four test methods using their standard binary prediction models. We calculated the predictive parameters specificity, sensitivity, accuracy, and balanced accuracy using all available data for a given test method and using the core chemical set, which were tested in all four test methods. Due to the low number of LLNA negative chemicals (n = 17) in the core chemical set, the test methods’ specificities should be interpreted with care.

Due to the differences in the prediction model purposes, the predictivity of the kDPRA, which discriminates UN GHS Cat. 1A sensitizers from the other two UN GHS categories, is not comparable to the other test methods’ predictivity where prediction models discriminate sensitizers from non-sensitizers.

Table 2 summarizes the predictive parameters. While the predictivity for all available data suggests that the ADRA has the highest predictivity, the predictivity of the ADRA and DPRA are very similar for the core chemical set. Also, when focusing on the 159 chemicals tested in both test methods, the predictivity of the DPRA and ADRA are almost identical with the standard prediction models, with 82.4% of the chemicals classified identically. The high similarity of the predictivity of dichotomized reference results for the common test chemicals was confirmed by ROC-curves (Fig. S4). This suggests that the differences in the predictivity for all chemicals is due to differences in the respective chemical sets. One major difference is that cosmetic ingredients that had been previously underrepresented were tested in the DPRA to explore its applicability (Hoffmann et al., 2022), but not in the ADRA. This difference has also contributed to the substantial difference in DPRA specificities between the core and all available chemicals. Another aspect to be considered in this regard is that 12 of the 17 NS in the core chemical set were part of the DPRA test set (Gerberick et al., 2007).

In comparison, the PPRA has a generally lower predictivity. An increase in sensitivity comes at the cost of a greater decrease in specificity. The predictive parameters of the kDPRA were very similar regardless of the chemical set.

Focusing on pre- and pro-haptens, the test methods’ predictivity of pre-/pro-haptens ranged from 66.7% (20/30) for the DPRA, to 86.1% (25/29) for the ADRA, and 92.0% (23/25) for...
the PPRA. The kDPRA correctly predicted 83.3% (10/12) UN GHS Cat. 1A pre- and pro-haptens. These values deviated from the sensitivity values of Table 2 in one or the other direction, but not to an extent that would suggest substantial systematic differences.

### 3.5 Predictivity associated with the water-solubility of chemicals

Triggered by the discussion on potential underprediction of poorly water-soluble chemicals, indicated by a logKOW > 3.5 (Hoffmann et al., 2022; OECD, 2023b; Takenouchi et al., 2013), the association of mispredicted chemicals, i.e., false negatives (FN) and false positives (FP) compared to the animal reference data, with the four individual physical-chemical properties was analyzed systematically for the four test methods based on all available chemicals. The statistical comparison of the chemical groups, i.e., true negatives (TN), true positives (TP), FN and FP showed a significantly higher logKOW of DPRA FN compared to FP and of PPRA FN compared to all other groups (Fig. 3, left side).

In addition, the predictive capacity parameters were calculated depending on the logKOW (Fig. 3, right side). While “all” presents the value as included in Table 2, more and more chemicals were removed from the calculation according to their logKOW, so that the dots on the right are based on chemicals with a logKOW > 6 only. It was observed that the DPRA and ADRA sensitivity, and therefore the accuracy, decreased for chemicals with logKOW around 3 (and higher). The PPRA showed a similar pattern which started around a logKOW of 2. Also, the kDPRA sensitivity dropped, but sample sizes were too small for any meaningful evaluation, e.g., there were only three UN GHS Cat. 1A chemicals with a logKOW > 3. This demonstrates the general problem that sample sizes, regardless of test method, shrink the higher the logKOW, so that the predictivity for chemicals with logKOW > 5 is of limited interpretability. The complexity of the issue is further increased when considering the poor specificity of LLNA for predicting human skin sensitizers for logKOW > 3 (Fig. 4), as alluded to by Natsch et al. (2021, 2023) and Hoffmann et al. (2022). Figure 4 displays – applying the same approach for the right-hand side of Figure 3 – the LLNA predictivity of known human sensitizers and non-sensitizers, depending on the logKOW of the chemicals. The specificity of the LLNA dropped considerably for chemicals with a logKOW > 3, indicating that the LLNA tends to overpredict such chemicals and potentially explaining the above-described decrease in sensitivity of the test methods.

For comparison, the same analysis as conducted for the logKOW (Fig. 3) is presented for the volatility class in Figure 5. It does not indicate any statistical significance nor pattern between mispredictions and volatility class. Corresponding figures for MW and WS, which show a similar pattern as observed for the logKOW for poorly water-soluble chemicals (due to the high correlation with the logKOW), are shown in Figure S5.

### 3.6 GHS predictivity: Identification of new thresholds

Finally, the predictivity of the very similar test methods DPRA and ADRA was analyzed in further detail, i.e., beyond dichotomous predictions. For this purpose, the LLNA EC3 values (in %) were plotted against the mean depletion values (Fig. 6). Chemicals not reaching an EC3 were considered as non-sensitizer, and a value of 100% was assigned. The four chemicals (HC violet (Cys depletion: 17.7%), oxazolone (47.2%), 2,4,6-trinitrobenzene sulfonic acid (99.94%), and jasmine absolute (Grandiflorum) (16.97%)) that co-eluted with lysine in the DPRA were assigned a mean lysine depletion value of 16.6%.

Mean depletion values and LLNA EC3 were moderately negatively correlated for both test methods, which indicated that a
Fig. 3: Predictivity of the four test methods depending on the logKOW of all available chemicals

The dot plots on the left-hand side show the logKOW distributions of false positives (FP) as circles, false negatives (FN) in grey, and correct classification (TP and TN) in black. The right-hand side shows the dependence of the predictivity parameters accuracy (black dots), specificity (white dots) and sensitivity (grey dots) on the logKOW, calculated for all chemicals (left side) up to those chemicals with a logKOW of > 6 (Kruskal-Wallis test; *, p < 0.05; **, p < 0.01; ***, p < 0.001).
led to the establishment of OECD GD 497, which includes two DA, one focused on skin sensitization hazard assessment and the other focused on classification according to the UN GHS, for which two versions using different in silico tools are available (OECD, 2021a).

Despite this progress enabling NAM-based regulatory classification decision making, several scientific issues related to the individual test methods and their combination remain. These include:

− the uncertainty originating from how well the individual test methods model the respective key events as, for example, discussed for key event 1 by Natsch and Emter (2017);
− a lack of understanding of the mutual dependence of the test methods covering the same key event addressed, for example, for some test methods by Hoffmann et al. (2022);
− use of prediction models developed for individual test methods for the combination of test methods in DA;
− a lack of qualitative mechanistic understanding, as summarized by Paini et al. (2022);
− discussions on how to assess test methods and DA by comparing them to animal or human data or both (Hoffmann et al., 2008; Irizar et al., 2022; Natsch et al., 2021; OECD, 2021b);
− their use for risk assessment purposes (Gilmour et al., 2020).

The here presented comparison of the experimental set-ups, the data repository, and respective analyses were intended to contribute to the scientific discussion on some of these issues with respect to the molecular initiating event (key event 1) of the skin sensitization AOP, i.e., covalent binding of a chemical to skin protein.

Fig. 4: LLNA predictivity of human skin sensitizers and non-sensitizers depending on the logKOW of 149 chemicals (all), for which human data were available, expressed as accuracy (black dots), specificity (white dots), and sensitivity (grey dots)

4 Discussion

Significant progress has been made toward NAM-based skin sensitization hazard assessment and classification and labeling. This has resulted in the adoption of the frequently updated OECD TG 442C, 442D and 442E (OECD 2023a-c), which feature in chemico or in vitro test methods modelling the first three key events of the skin sensitization AOP (OECD, 2014). These meaningful quantitative interpretation of the reactivity test method is not obvious. However, exploring the plots with a focus on UN GHS categories suggested an alternative threshold potentially useful for skin sensitization hazard and risk assessment. A DPRA threshold of 18% mean depletion, for example, is associated with a predictivity of chemical being either LLNA non-sensitizer (NS) or Cat. 1B sensitizer. Only 3 of 52 Cat. 1A chemicals (benzalkonium chloride, hexyl salicylate and nonanyl chloride) would be missed, resulting in a sensitivity for Cat. 1A vs Cat. 1B/no category of 94.2% (49/52) and a negative predictive value of 97.3% (110/113), i.e., among all chemicals with a mean depletion < 18% predicted as Cat. 1B/no category, 2.7% would be Cat. 1A. A similar performance (missing one (hexyl salicylate) of 43 Cat. 1A chemical) would be obtained for the ADRA with a threshold of 7.5%.

Another potentially useful ADRA threshold would be 70% mean depletion: All chemicals with a higher mean depletion are Cat. 1A sensitizers, resulting in a positive predictive value for identifying Cat. 1A chemicals of 100%. In practice, such chemicals could be de-prioritized or even excluded from further investigations, e.g., for uses with consumer exposure. In addition, it would be interesting to explore such thresholds in the context of possibly tiered defined approaches.
Fig. 5: Predictivity of the four test methods depending on the volatility class of all available chemicals

The dot plots on the left-hand side show the volatility class distributions of false positives (FP) as circles, false negatives (FN) in grey, and correct classification (TP and TN) in black. The right-hand side shows the dependence of the predictivity parameters accuracy (black dots), specificity (white dots) and sensitivity (grey dots) on the volatility class, calculated for all chemicals (left side) up to those chemicals assigned to volatility class 3, i.e., > 2.
These differences need to be considered when analyzing the predictivity of the individual test methods. For example, the predictivity of the ADRA seemed to be superior to the DPRA when considering all available data for each test method. However, when focusing on the chemicals tested using both methods, the predictivity was almost identical. This result confirmed that the ADRA and DPRA do not only have a very similar test definition, but also provide very similar results. From this it was concluded that DPRA and ADRA can be used interchangeably, for example in DA.

Excluding the kDPRA due to the low number of UN GHS Cat. 1A chemicals, a general tendency towards increased FN with increasing logKOW was observed when comparing to LLNA results, confirming earlier results reported for the DPRA.

These differences need to be considered when analyzing the predictivity of the individual test methods. For example, the predictivity of the ADRA seemed to be superior to the DPRA when considering all available data for each test method. However, when focusing on the chemicals tested using both methods, the predictivity was almost identical. This result confirmed that the ADRA and DPRA do not only have a very similar test definition, but also provide very similar results. From this it was concluded that DPRA and ADRA can be used interchangeably, for example in DA.

In addition, publicly available data for the four reactivity in chemico test methods have been compiled to address some specific questions, but also as a resource for interested researchers. The repository includes 260 chemicals and a total of 770 test method results, ranging from 161 to 242 per test method. While this represents a wealth of data, the core chemical set of 91 chemicals with data from all test methods might need to be expanded to approach questions around applicability and predictivity, primarily as it currently includes only 17 non-sensitizers.
and PPRA (Hoffmann et al., 2022). Considering that the LLNA tends to overpredict high logKOW chemicals as compared to human data, the analysis, which was somewhat limited by the relatively low number of chemicals, suggested that reactivity test methods may be more human relevant than the LLNA for chemicals with a logKOW > 3. Investigation of the relation of mean depletion values and LLNA EC3 values and UN GHS categories for the DPRA and the ADRA showed that the extent of correlation is likely too limited for quantitative predictions.

However, this analysis revealed new categorization thresholds that could be relevant for strategic uses. For both the DPRA and ADRA, thresholds at the lower response spectrum (18% for the DPRA and 7.5% for the ADRA) were identified that have a high sensitivity for discriminating UN GHS Cat. 1A chemicals from the rest, resulting in very high negative predictive values, i.e., among all chemicals with a mean depletion below the respective threshold, less than 3% would be Cat. 1A. In addition, the 11 chemicals with a mean depletion of at least 70% in the ADRA were Cat. 1A sensitizers, resulting in a positive predictive value of 100%. In practice, such high ADRA mean depletion values could serve as a stop criterion early in the development of specific chemical uses or as an early decision point in a tiered risk assessment approach. It is acknowledged that these identified thresholds were driven by the data. However, the established thresholds of the DPRA and ADRA were also determined based on the available data at the time, but from a substantially lower number of chemicals. For example, the DPRA was developed with 56 chemicals (and tested with 26 chemicals) using a classification tree approach (Gerberick et al., 2007). Confidence in the ADRA threshold of 70% would likely need to be increased by expanding the underlying database.

In summary, a detailed assessment of reactivity test methods is provided here, highlighting the potential, but also some limitations, of the test methods. The results are intended to stimulate scientific discussion around the test methods modelling the molecular initiating event of the skin sensitization AOP. In particular, the results inform the applicability of test methods covering the same key event, allowing to identify similarities and differences that can guide the choice of the most appropriate test method for a given chemical. In a hazard and classification and labelling context, for example, the interchangeability of the DPRA and ADRA is supported. This strongly suggests that for the “2 out of 3” defined approach the predictive performance will be very similar regardless of whether the DPRA or the ADRA is used (OECD, 2021b). Such interchangeability – an aspect that is currently under general discussion at OECD – may be adopted/approved in the context of the OECD Guideline 497 revision. It also means that the added value of conducting both the DPRA and ADRA in parallel will be limited. In the worst case, inherent test method variability and the binary interpretation based on existing PM may lead to inconsistent predictions, potentially decreasing confidence. In the broader risk assessment context, such insights are valuable for the development of integrated approaches to testing and assessment (IATA) for skin sensitization risk assessment purposes.

References
Kleinsteuver, N. C., Hoffmann, S., Alépée, N. et al. (2018). Non-animal methods to predict skin sensitization (II): An assessment of


OECD (2023a). Test No. 442C: In Chemico Skin Sensitisation: Assays addressing the Adverse Outcome Pathway key event on covalent binding to proteins. OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing. doi:10.1787/9789264229709-en


Conflict of interest
The authors declare that they have no conflicts of interest.

Data availability
The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.3,4

Acknowledgements
The authors wish to thank Sebastian Hoffmann (seh consulting + services), and Ann Detroyer and Laurent Nardelli (L’Oréal Research and Innovation, Aulnay-sous-Bois, France) for their contributions to this project.