Research Article

# Predictivity of the Kinetic Direct Peptide Reactivity Assay (kDPRA) for Sensitizer Potency Assessment and GHS Subclassification

*Andreas Natsch[1], Tina Haupt[1], Britta Wareing[2], Robert Landsiedel[2] and Susanne N. Kolle[2]*

[1]Fragrances S&T, Ingredients Research, Givaudan Schweiz AG, Kemptthal, Switzerland; [2]BASF SE Experimental Toxicology and Ecology, Ludwigshafen, Germany

**Abstract**

Several *in vitro* OECD test guidelines address key events 1-3 of the adverse outcome pathway for skin sensitization, but none are validated for sensitizer potency assessment. The reaction of sensitizing molecules with skin proteins is the molecular initiating event and appears to be rate-limiting, as chemical reactivity strongly correlates with sensitizer potency. The kinetic direct peptide reactivity assay (kDPRA), a modification of the DPRA (OECD TG 442C), allows derivation of rate constants of the depletion of the cysteine-containing model peptide upon reaction with the test item. Its reproducibility was demonstrated in an inter-laboratory study. Here, we present a database of rate constants, expressed as log $k_{max}$, for 180 chemicals to define the prediction threshold to identify strong sensitizers (classified as GHS 1A). A threshold of log $k_{max}$ -2 offers a balanced accuracy of 85% for predicting GHS 1A sensitizers according to the local lymph node assay. The kDPRA is proposed as a stand-alone assay for identification of GHS 1A sensitizers among chemicals identified as sensitizers by other tests or defined approaches. It may also be used for the prediction of sensitizer potency on a continuous scale, ideally in combination with continuous parameters from other *in vitro* assays. We show how the rate constant could be combined with read-outs of other *in vitro* assays in a defined approach. A decision model based on log $k_{max}$ alone has, however, a high predictivity and can be used as stand-alone model for identification of GHS 1A sensitizers among chemicals predicted as sensitizers.

## 1 Introduction

The field of non-animal testing for skin sensitization has rapidly advanced over the past decade, leading to the publication of the adverse outcome pathway (AOP) for skin sensitization by the OECD in 2012, in which the sensitization process has been simplified and described as a series of mechanistic key events (KE) (OECD, 2012).

Three OECD test guidelines were published covering the first three KE (OECD, 2018a,b, 2020), namely KE 1 on covalent binding to proteins (the molecular initiating event, MIE), KE 2 on keratinocyte activation, and KE 3 on activation of dendritic cells. At the time of writing, these three test guidelines cover a total of seven test methods. They were all developed for hazard identification based on a binary prediction model, while the cellular assays also provide concentration-response information. However, reproducibility of the concentration-response data was not evaluated in the peer-review of the validation studies, nor has a formalized way to use this quantitative information reached test guideline status. Nevertheless, multiple studies have investigated application of quantitative *in vitro* data from the validated assays for potency prediction (Hirota et al., 2018; Jaworska et al., 2013, 2015; Natsch et al., 2015; Nukada et al., 2013).

Potency prediction is of importance for two reasons: First, the Global Harmonized System (GHS) has introduced a subclassification of sensitizers into 1A (strong sensitizers) and 1B (other sensitizers). Chemicals with an estimated concentration inducing a 3-fold increase in the stimulation index (EC3) $\leq$ 2% in the local lymph node assay (LLNA) are classified as GHS Cat 1A, or, alternatively, chemicals with an induction dose ($DSA_{05}$) $\leq$ 500 µg/cm$^2$ in humans tests are attributed to this class, with the $DSA_{05}$ defined as the dose per unit area of skin required to sensitize 5% of the panelists in a human repeat insult patch test (HRIPT) or human maximization test (HMT). This differentiation into sub-cat-

Received April 29, 2020; Accepted August 17, 2020;
Epub August 19, 2020; © The Authors, 2020.

**ALTEX 37(4), 652-664**. doi:10.14573/altex.2004292

Correspondence: Andreas Natsch, PhD
Fragrances S&T, Ingredients Research, Givaudan Schweiz AG
Kemptpark 50, 8310 Kemptthal, Switzerland
(andreas.natsch@givaudan.com)

egories leads to different labels, but has no other regulatory consequences so far. However, the regulatory requirement for this subclassification necessitates that the LLNA on mice must still be conducted on many chemicals found positive in *in vitro* assays (Barentsen et al., 2019). It is thus highly desirable to validate a non-animal assay to perform this subclassification for labeling of chemicals. Also, and at least as importantly, a quantitative and continuous prediction of sensitizer potency is required to derive a NESIL (no expected sensitization induction level) as a point of departure (PoD), which is required for quantitative risk assessment (QRA) (Api et al., 2008; Basketter and Safford, 2016).

Reaction of chemicals or their abiotic or biotic transformation products with nucleophilic residues in skin proteins is the MIE of skin sensitization. The modification of the proteins creates novel immunogenic epitopes. These epitopes ultimately trigger activation of epitope-specific T-cells, which are the effector cells that finally elicit allergic contact dermatitis. While multiple steps are involved in the acquisition of skin sensitization (epitope processing, danger signal formation, activation of dendritic cells and keratinocytes, cell migration), the MIE is argued to be of predominant importance (Roberts and Aptula, 2008). Potency is characterized by the steepness of the dose-response curve, and the formation of immunogenic epitopes sets the effective internal dose. Hence, significant correlation between a sensitizer's reactivity and its potency was found in multiple studies (Alvarez-Sánchez et al., 2004; Delaine et al., 2014; Natsch et al., 2011; Niklasson et al., 2011; Roberts and Aptula, 2014). In chemical kinetics, the rate constant k indicates the extent of reaction that occurs in a given time t.

The OECD adopted methods (OECD TG 442C) (OECD, 2020) to characterize the MIE are the direct peptide reactivity assay (DPRA) (Gerberick et al., 2004) and the amino acid derivative reaction assay (ADRA) (Yamamoto et al., 2015). In the classical DPRA, reactivity with cysteine- and lysine-containing synthetic peptides as surrogate nucleophiles is measured. Reactivity is quantified as the relative depletion of the peptide signal after incubation with the test chemical using UV-HPLC analysis. The prediction model of the DPRA is based on the mean of lysine- and cysteine-peptide depletion at a single time point measurement ($\geq$ 24 h) and at one fixed test material concentration (5 mM for the cysteine-containing peptide) (Gerberick et al., 2004). Thus, the DPRA determines the peptide depletion at one late time point and one fixed test substance concentration, and thus chemicals of differing potency may lead to similar depletions (e.g., 100% depletion after a complete reaction).

The kinetic direct peptide reactivity assay (kDPRA) is a modification of the DPRA. In the kDPRA, the reaction kinetics of a test substance towards the same synthetic cysteine-containing peptide as used in the DPRA is evaluated. For this purpose, several concentrations of the test substance are incubated with the synthetic peptide for several incubation times. The residual concentrations of the cysteine peptide after several reaction times are quantified by their reaction with monobromobimane (mBrB). The highly reactive and non-fluorescent mBrB rapidly reacts with unbound cysteine moieties of the model peptide to form a fluorescent complex. The non-depleted peptide can thus be determined fluorometrically, and the matrix of depletion values versus time and concentration can be transformed to a rate constant. The maximum observed logarithmic rate constant observed at any time point (log $k_{max}$, with log k for the different time points derived from the slopes of the natural logarithm of the non-depleted peptide concentrations plotted versus the concentration of the test chemical) is the key parameter reported, noting that this is the maximum observed rate constant obtained with the standardized data interpretation procedure, not necessarily the true maximal rate constant.

The basic method of the kDPRA was described for the first time by Natsch et al. (2007). The approach to use it for rate constant determinations was described by Roberts and Natsch (2009) and later applied to more chemicals from specific domains (Natsch et al., 2011). In 2017, it was applied to a more diverse set of chemicals, and a tentative prediction model was proposed with a threshold of log $k_{max}$ = -1.1 to discriminate between GHS Cat 1A and 1B (Wareing et al., 2017). An international validation study in seven laboratories has recently been completed and has proven a high quantitative reproducibility of the rate constant determinations (Wareing et al., 2020).

Here, we have compiled a database of log $k_{max}$ values on 180 chemicals based on new measurements, published data, and data from the validation study. This database is used to derive an optimal prediction threshold for GHS subclassifications, and it is analyzed for quantitative predictions in relation to LLNA EC3 values alone or in combination with other endpoints measured using OECD approved *in vitro* methods.

## 2 Materials and methods

Test substance chemicals were purchased from Sigma-Aldrich; fragrance chemicals were obtained from Givaudan Schweiz AG. The Cys-peptide (Ac-RFAACAA, MW 750.1) was obtained from Genscript Inc. (Piscataway, NJ, USA). Two batches with a purity of 95.6 and 96.4% were used.

*kDPRA and rate constant determination*

The kDPRA was run according to the standard operating procedure (SOP) defined for the validation study; details of the method are described in the parallel paper on the ring trial validation study (Wareing et al., 2020; training video[1]). Briefly, the kDPRA was run in microtiter plates in a final volume of 160 µL. Test agents were dissolved in acetonitrile at concentrations of 20, 10, 5, 2.5 and 1.25 mM. These solutions (40 µL) were added to individual wells of the microtiter plates containing 120 µL of the Cys-peptide solution (0.66 mM in 100 mM phosphate buffer pH 7.5). Control wells contained solvent and buffer only to determine background fluorescence. The plates were cov-

[1]doi:10.5281/zenodo.3901109

ered with impermeable foil to avoid evaporation and were incubated protected from light at 25°C for different time intervals. The reaction was stopped by addition of mBrB solution (40 µL per well; 3 mM in ACN). The highly reactive and non-fluorescent mBrB rapidly reacts with unbound cysteine moieties of the model peptide to form a fluorescent complex. The remaining non-depleted Cys-peptide concentration can thus be determined by measuring the fluorescence at 390/480nm after 5 min incubation at RT.

Peptide depletion (dp) was expressed as the percentage decrease in concentration of free thiol groups compared to control samples with the peptide in solvent only. The kDPRA is performed in parallel reactions of five final concentrations of the test chemical (5, 2.5, 1.25, 0.625 and 0.3125 mM) and over six different reaction times (10, 30, 90, 150, 210 and 1440 min). If the depletion of the highest tested concentration (5 mM test chemical) surpasses the threshold of 13.89% (threshold used in the DPRA for Cys-only positivity in accordance with OECD 442C) and this depletion is statistically significant versus controls with peptide only, further calculations are performed: The natural logarithm of the non-depleted peptide concentrations is plotted versus the concentration of the test chemical at each time point. If a linear relationship is observed (correlation coefficient > 0.9), the slope of this curve is determined and divided by the incubation time to calculate the rate constant in [$min^{-1}mM^{-1}$] measured at a given time. This value is transformed to the rate constant in [$s^{-1}M^{-1}$], and the logarithm is taken. The maximum value observed at any time point is taken as the log $k_{max}$. In cases of non-linear behavior (reaction does not progress continuously over time or curvature of the plots versus concentration), the SOP of the kDPRA proposes alternative calculations and provides detailed guidelines on how to interpret the data. In such cases, experiments are first repeated to confirm non-linearity is intrinsic to the chemical being assessed and not to experimental error. Reaction rates then can be calculated using Equation1:

$$k = [ln(100/(100-dp))]/[[E]o\ t] \qquad \text{Eq. 1}$$

where dp is depletion in %, E is the concentration of test chemical and t is the incubation time. This is detailed in the SOP and implemented in the evaluation template of the standardized test.

*Chemical selection and literature reference data*

Chemical selection aimed at complementing the database with reference chemicals that had been previously evaluated in various other alternative methods and with presumably well-known *in vivo* reference data, e.g., those in (Hoffmann et al., 2018; ICCVAM, 2011; Urbisch et al., 2015). They were thus selected from the following lists: (i) Chemicals from the ICCVAM list on LLNA potency (ICCVAM, 2011), (ii) performance standard substances selected for OECD 442D (Nrf-2 luciferase test) (OECD, 2018a), OECD 442C (DPRA) (OECD, 2020) and OECD 429 (LLNA) (OECD, 2010), (iii) chemicals from the Cosmetics Europe database (Hoffmann et al., 2018). All these chemicals have been tested in multiple *in vitro* skin sensitization tests and some human evidence, at least from clinical data and expert judgment, is available (Basketter et al., 2014), and finally (iv) additional chemicals from the compilation by Urbisch et al. (2015). Metals and complex mixtures such as essential oils included in the above datasets were not included as they are considered outside the applicability domain of the DPRA and also the kDPRA.

The *in vivo* reference data for the LLNA were taken from the different reference lists cited above. If LLNA data in the Cosmetics Europe database (Hoffmann et al., 2018) were available, these were given precedence, as they are derived from multiple LLNA studies if available, unless there is evidence that more reliable data are reported elsewhere. This LLNA dataset was recently reviewed by the OECD expert group on defined approaches (DAs) to skin sensitization. However, an initial draft of this analysis (OECD, 2019a,b) became public just after the evaluation reported here was finalized. The public draft OECD review has corrected some errors, but it also introduced some new mistakes under further revision, and hence we present the analysis done with the database compiled before the OECD review here. We indicate, however, specific cases where the OECD review came to a different conclusion and have added the preliminary OECD assessment to the database in Table S2[3].

The human data originate from the assessment by ICCVAM when evaluating the LLNA for potency discrimination (ICCVAM, 2011), the RIFM database, and largely also from the Cosmetics Europe database and the corresponding potency assessment by Basketter et al. (2014). As described in that paper, human potency classes 1 and 2 fall into GHS Cat 1A, while 3-6 are considered GHS Cat 1B/NC. However, it should be noted that part of that assessment is based on clinical data and not on predictive human tests. Thus, not all chemicals labeled as Cat 1A in that assessment would be rated as Cat 1A based on human predictive tests (HRIPT and HMT). The database in Table S2[3] also contains data from KeratinoSens®, DPRA and h-CLAT assays where available. These data are primarily drawn from previous data compilations (Jaworska et al., 2015; Natsch et al., 2013, 2015; Urbisch et al., 2015), with additional data added from the testing conducted by Cosmetics Europe (Hoffmann et al., 2018).

*Statistical data evaluation*

A potential classification into GHS Cat 1A was determined for all chemicals based on different thresholds for high reactivity, which were set to log $k_{max}$ = -3.0 to +0.8 in incremental steps of 0.1. For each threshold, the sensitivity, specificity and balanced accuracy were calculated to discriminate GHS Cat 1A chemicals (as defined by an LLNA EC3 ≤ 2%) from all other chemicals. Based on this predictivity for identifying GHS Cat 1A chemicals at different thresholds, receiver operator characteristics (ROC) curves were plotted, and the threshold maximizing balanced accuracy was determined. Additionally, the same analysis for maximized balanced accuracy was also performed to predict human data.

To calculate the correlation between log $k_{max}$ and the LLNA EC3 values on a continuous scale, EC3 of non-sensitizers (i.e., SI 3 not reached at maximum test concentrations) were set to 100 for the part of the analysis including the negative chemicals, although

**Tab. 1: Chemicals tested independently in previous experiments prior to SOP definition and in the validation ring trial according to the SOP. Log $k_{max}$ values in $s^{-1}M^{-1}$ are shown.**

| Chemicals | CAS | BASF historical (Wareing et al., 2017) | Givaudan historical (Natsch et al., 2011; Roberts and Natsch, 2009)[b] | Ring trial[a] |
|---|---|---|---|---|
| Cinnamic aldehyde | 104-55-2 | -1.33 | -1.60 / -1.66 | -1.35 |
| 2,4-Dinitrochlorobenzene | 97-00-7 | -0.61 | -0.65 | -0.56 |
| Benzylidene acetone | 122-57-6 | -1.77 | -1.86 / -1.89 | -1.85 |
| Diethylmaleate | 141-05-9 | -1.16 | -1.09 / -1.03 | -1.21 |
| Ethylene glycol dimethacrylate | 97-90-5 | -2.34 | -2.24 / -2.11 | -2.44 |
| Oxazolone | 15646-46-5 | -0.19 | -0.43 | -0.14 |

[a] Average of 7 laboratories testing according to the SOP; [b] Rate constants were derived from the full depletion matrix as described in Roberts and Natsch (2009).

some may have been tested only up to 20-25%. LLNA EC3 values were transformed to pEC3 values according to Equation 2:

$$pEC3 = log\left(\frac{Mol.weight}{EC3}\right) \qquad \text{Eq. 2}$$

For regression analysis, the numerical data from KeratinoSens® (EC1.5 and EC3 for 1.5- and 3-fold luciferase induction, IC50 for 50% cytotoxicity, all in µM) were transformed to logarithmic values. If no induction was reached and in case no toxicity was observed up to the maximum test concentration of 2000 µM, a default value of 4000 µM was used. The numerical data from h-CLAT (EC150 for 1.5-fold CD86 induction, EC200 for 2-fold CD54 induction, MIT, minimum induction threshold of EC150 and EC200, and CV75 for 25% cytotoxicity, all in µg/mL) were transformed to logarithmic values. If no induction was reached and in case no toxicity was observed up to a maximum test concentration of 5000 µg/mL, a default value of 5000 µg/mL was used for numerical analysis. For the DPRA, the single time point depletion values are transformed to logarithmic rate constants according to Equation 3 (Roberts and Natsch, 2009):

$$\log k = \log \frac{\ln \frac{100}{100\text{-}dp}}{[E]_0 t} \qquad \text{Eq. 3}$$

where dp is the peptide depletion, [E] the test substance concentration in mM (5 mM) and t the incubation time (1440 min). All logarithmic data from LLNA and *in vitro* assays were then used for single or multiple linear regression analysis in MiniTab software (MiniTab® 18.1; ©2017 MiniTab).

## 3 Results

### 3.1 A database of log $k_{max}$ values

The database includes all data from the validation study (Wareing et al., 2020) and all the data measured in the previous three publications (Natsch et al., 2011; Roberts and Natsch, 2009; Wareing et al., 2017) (Tab. S2[3]). Chemicals with weak peptide depletion in these previous publications were retested according to the SOP, which contains a number of rules to exclude, based on a more rigorous statistical analysis, random positive results from small fluctuations in the depletion data (DB-ALM protocol no. 217, in preparation). Some previously tested chemicals with significant reactivity were included in the validation study, and the results are compared here to the published, historical data. The method as used in the previous publications is identical in terms of volume of the reaction, composition of the reaction, and the fluorescent read-out, and therefore it is fully compatible with the method as described in the SOP, while the definition of the plate setup as well as the automatic evaluation and statistical analysis of the data was introduced with the SOP (DB-ALM protocol no. 217, in preparation). Table 1 summarizes data for six chemicals with historical results that were repeated according to the SOP in the validation study. Highly similar results were obtained, which indicate that it is justified that the published historical results are integrated into the database.

Additional chemicals, selected from different reference lists as described in Section 2, were tested according to the SOP. This complete database now contains log $k_{max}$ data on a total of 180 chemicals with LLNA and 123 chemicals with human potency category attribution. Six chemicals were excluded from evaluation due to strong interference (reaction with mBrB, quenching or autofluorescence, Tab. S1[2]). The full database including the human and LLNA reference data and quantitative data from other *in vitro* assays is given in Table S2[3].

### 3.2 ROC analysis to derive optimal prediction threshold and predictivity for identification of GHS Cat 1A sensitizers

Different proposals for DAs for skin sensitization hazard identification are currently being evaluated for their inclusion into an OECD guideline (OECD, 2019a), i.e., for discrimination of 1B/1A chemicals from non-sensitizing chemicals. Thus, a predictive threshold from a validated method is now needed to identi-

[2] doi:10.14573/altex.2004292s1

[3] doi:10.14573/altex.2004292s2

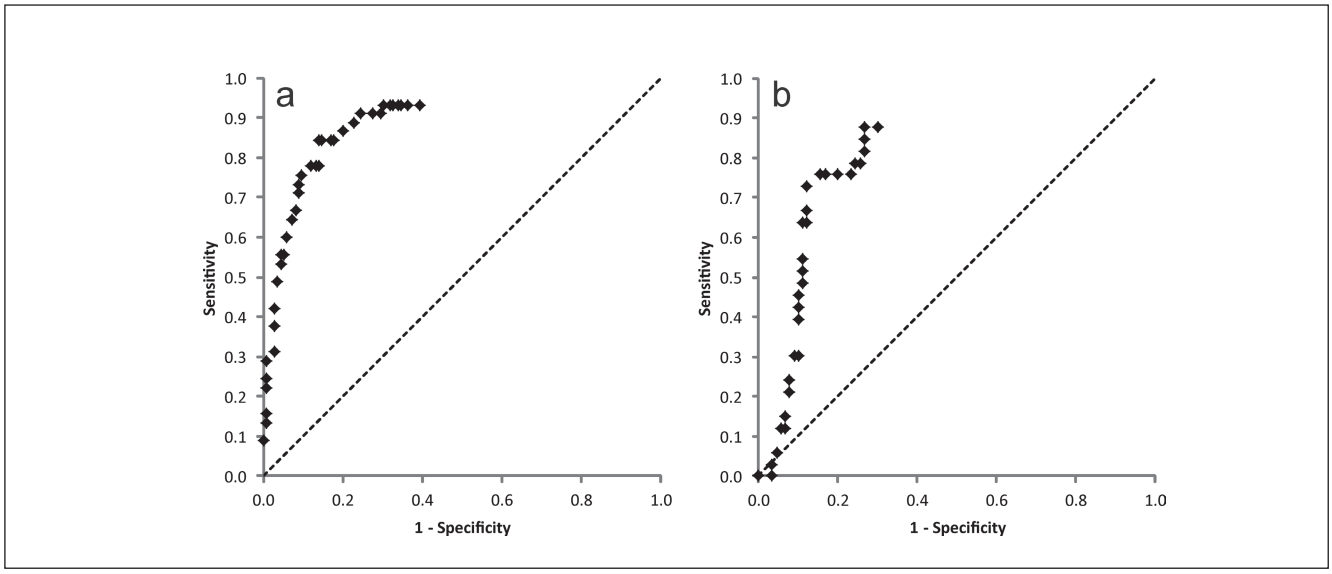


**Fig. 1: ROC curves for different log $k_{max}$ threshold values to discriminate GHS Cat 1A from 1B/NC chemicals as determined by LLNA data (a) and human data (b)**

**Tab. 2: Predictivity of different log $k_{max}$ threshold values to predict GHS Cat 1A vs. LLNA and human data**
The threshold range from log $k_{max}$ = -1.5 to -2.5 is shown.[a]

| Threshold | LLNA | | | Human | | |
|---|---|---|---|---|---|---|
| | Sensitivity | Specificity | Balanced accuracy | Sensitivity | Specificity | Balanced accuracy |
| -1.5 | 73.3% | 91.1% | 82.2% | 42.4% | 90.0% | 66.2% |
| -1.6 | 75.6% | 90.4% | 83.0% | 45.5% | 90.0% | 67.7% |
| -1.7 | 77.8% | 88.1% | 83.0% | 48.5% | 88.9% | 68.7% |
| -1.8 | 77.8% | 86.7% | 82.2% | 51.5% | 88.9% | 70.2% |
| -1.9 | 77.8% | 85.9% | 81.9% | 54.5% | 88.9% | 71.7% |
| -2.0 | 84.4% | 85.9% | 85.2% | 63.6% | 88.9% | 76.3% |
| -2.1 | 84.4% | 85.2% | 84.8% | 63.6% | 87.8% | 75.7% |
| -2.2 | 84.4% | 83.0% | 83.7% | 66.7% | 87.8% | 77.2% |
| -2.3 | 84.4% | 82.2% | 83.3% | 66.7% | 87.8% | 77.2% |
| -2.4 | 86.7% | 80.0% | 83.3% | 72.7% | 87.8% | 80.3% |
| -2.5 | 88.9% | 77.0% | 83.0% | 75.8% | 84.4% | 80.1% |

[a] A binary classification into GHS 1A vs. 1B/NC chemicals for each incremental 0.1 step for log $k_{max}$ was performed here and corresponding Cooper statistics were calculated.

fy GHS Cat 1A sensitizers in a binary prediction model. The data from the extended database were separated into GHS Cat 1A versus GHS Cat 1B/NC chemicals, and the sensitivity and specificity for identifying 1A chemicals was then calculated versus LLNA data or, separately, versus human data for different thresholds of log $k_{max}$ in incremental steps of 0.1 between -3 and +0.8.

Figure 1 shows results as a ROC curve for LLNA data and human reference data plotting sensitivity versus [1 - specificity]. The predictivity for all different thresholds underlying these curves is summarized in Table S3[2], and Table 2 shows results for the more relevant log $k_{max}$ threshold range between -1.5 to -2.5.

The predictivity analysis versus LLNA data indicates that a log $k_{max}$ = -2.0 maximizes balanced accuracy (85.2%) with a specificity of 85.9% and a sensitivity of 84.4%. This threshold is more conservative and more predictive compared to the tentative threshold of log $k_{max}$ = -1.1 proposed initially, which was based on only 38 substances (Wareing et al., 2017). This difference may be due to a bias in the much smaller database evaluated previously.

**Tab. 3: LLNA GHS Cat 1A sensitizers that are underpredicted when applying a threshold of log $k_{max}$ = -2.0**

| Name | CAS | log $k_{max}$ | LLNA EC3 (%) | Human GHS Cat |
|---|---|---|---|---|
| 4-Phenylenediamine | 106-50-3 | -2.81 | 0.15 | 1A |
| Glutaric aldehyde | 111-30-8 | -3.50 | 0.09 | 1A |
| 2-Aminophenol | 95-55-6 | -2.46 | 0.45 | 1A |
| Hexyl salicylate | 6259-76-3 | -3.50 | 0.18[a] | (1B/NC)[b] |
| Bisphenol A-diglycidyl ether | 1675-54-3 | -2.53 | 1.5 | 1B |
| 2,4-Dinitrobenzenesulfonic acid | 885-62-1 | -2.30 | 1.9 | no data |

[a] LLNA GHS 1B based on OECD review. [b] Attributed 1B in Basketter compilation, but human HRIPT indicates no sensitization when tested up to 35,400 $\mu g/cm^2$ and there are no data from the clinic indicating it to be a sensitizer; thus there is no evidence that it is a human sensitizer.

**Tab. 4: LLNA Cat 1B sensitizers that are overpredicted by applying a threshold of log $k_{max}$ = -2.0**

| Name | CAS | log $k_{max}$ | LLNA EC3 (%) | Human GHS Cat |
|---|---|---|---|---|
| Ethyl 2-nonynoate | 111-80-8 | -1.66 | 2.5 | 1A |
| *trans*-2-Hexenal | 6728-26-3 | -0.47 | 4.05 | 1A |
| Benzylidene acetone | 122-57-6 | -1.85 | 3.7 | 1A |
| Diethylmaleate[a] | 141-05-9 | -1.21 | 4.7 | 1A |
| Benzosiothiazolinone | 2634-33-5 | -0.12 | 4.8 | 1A |
| Safranal | 116-26-7 | -1.74 | 7.5 | 1A |
| 2-Mercaptobenzothiazole[a] | 149-30-4 | -0.15 | 2.6 | 1B |
| Tetramethyldiuram disulfide | 137-26-8 | 0.74 | 2.93 | 1B |
| Abietic acid | 514-10-3 | -0.55 | 11 | 1B |
| Imidazolidinyl urea | 39236-46-9 | -1.11 | 24 | 1B |
| Ethyl acrylate | 140-88-5 | -0.97 | 32.75 | 1B |
| 4-Vinylcyclohex-1-ene-carbaldehyde | 1049017-68-6 | -1.77 | 4 | No data |
| α-Damascone | 24720-09-0 | -1.64 | 3.3 | No data |
| 2,4-Heptadienal | 5910-85-0 | -1.52 | 4 | No data |
| Bergamal | 22418-66-2 | -1.64 | 11 | No data |
| 2-Decenal | 3913-71-1 | -1.03 | 2.5 | No data |
| Oxalic acid anhydrous | 144-62-7 | -1.01 | 15 | No data |
| Butyl acrylate | 141-32-2 | -0.83 | 20 | No data |
| 1,2,4-Benzenetricarboxylic anhydride | 552-30-7 | -0.13 | 9.2 | No data |

[a] LLNA GHS 1A based on OECD review

Table 3 summarizes the underpredicted chemicals versus an LLNA-based classification. Two of these are pre-haptens, which require oxidation to become reactive (4-phenylenediamine and 2-aminophenol). There is no evidence that hexyl salicylate is a human sensitizer (tested up to 35,400 $\mu g/cm^2$ in human predictive tests), and it was rated 1B in the LLNA in the recent data review conducted by the OECD group. Two chemicals (bisphenol A-diglycidyl ether and 2,4-dinitrobenzenesulfonic acid) are very close to the prediction threshold of both the kDPRA and the LLNA; it is by definition difficult to reliably attribute chemicals to classes if the data are close to a binary prediction threshold. Glutaric aldehyde is a chemical predominately reactive by a Schiff-base formation mechanism (and hence reacting selectively with amine groups) in the classical DPRA protocol (Gerberick et al., 2004). It is probably the most important sensitizer next to 4-phenylenediamine that is underpredicted by the kDPRA in the current dataset.

Table 4 lists the overpredicted chemicals as compared to the LLNA when applying the threshold log $k_{max}$ = -2.0. It is interesting that 11 of the 19 cases are in the LLNA EC3 range of 2-5%, thus close to the GHS Cat 1A threshold. Furthermore, the list con-

**Tab. 5: Summary of the predictivity using the threshold of log $k_{max}$ = -2**
Predictivity data are given in percent and n correct predictions / n chemicals in the group are given in brackets

| | Sensitivity (%) | Specificity (%) | Balanced accuracy (%) | n tested |
|---|---|---|---|---|
| kDPRA vs LLNA | 84.4% (38/45) | 85.9% (116/135) | 85.2% | 180 |
| kDPRA vs LLNA for a set of substances with human data | 76.9% (20/26) | 88.7% (86/97) | 82.8% | 123 |
| kDPRA vs human | 63.6% (21/33) | 88.9% (80/90) | 76.3% | 123 |
| *LLNA vs human* | *54.5% (18/33)* | *91.1% (82/90)* | *72.8%* | *123* |

tains a number of chemicals with clear sensitization risks based on clinical data (e.g., the important glove allergens 2-mercaptobenzothiazole (Warburton et al., 2015) and tetramethyldiuram disulfide (Warshaw et al., 2013, 2017)) and 6 human Cat 1A chemicals. Thus, for at least eight of the 19 false-positives versus the LLNA-based class attribution, a significant human sensitization potential is reported. The apparent rapid initial depletion of oxalic acid cannot be explained, however. The list also includes two strongly reactive acrylates, which are highly volatile (ethyl acrylate and butyl acrylate) and evaporate rapidly under LLNA conditions (Natsch et al., 2015) but are likely to be strong sensitizers when applied under occluded conditions as is done in human predictive tests and as may occur under in-use scenarios.

The analysis versus human data indicates that a log $k_{max}$ = -2.0 yields a balanced accuracy of 76% with a specificity of 89% and a sensitivity of 64%. A further gain in sensitivity (73%) and balanced accuracy (80%) could be achieved using a threshold of log $k_{max}$ = -2.4; however, this would then reduce the accuracy versus LLNA data (84% instead of 85% balanced accuracy).

To evaluate whether this predictivity for human data by the kDPRA is sufficiently good, a comparison of the LLNA versus human data can be made: The LLNA has a sensitivity of 55% and a balanced accuracy of 73% for this dataset when predicting the human potency categorization for the same chemicals and based on the same reference data. Thus, with the threshold of log $k_{max}$ = -2.0, the kDPRA is still slightly superior to the LLNA in predicting human potency in this dataset. Predictivity for human and LLNA data and of the LLNA versus human data is summarized in Table 5.

Table S4[2] lists the false-negative human GHS Cat 1A sensitizers based on a threshold of log $k_{max}$ = -2.0, and a discussion is provided to interrogate the predictivity and to check whether this threshold is sufficiently protective. The analysis indicates that only few chemicals are very clear-cut human GHS Cat 1A sensitizers. In terms of specific domains that tend to be underpredicted, phenolic pre-haptens (chemicals that can be oxidized to reactive catechols or quinones, e.g., phenylenediamine and 2-aminophenol), pro-haptens which can be transformed to potent allergens reacting by Schiff-base formation (diethylenetriamine and 3-dimethylaminopropylamine), and potent Schiff-base forming chemicals (glutaric aldehyde) are underpredicted for their human sensitization potency.

### 3.3 Correlation of log $k_{max}$ and continuous parameters from other *in vitro* assays to potency in the LLNA

Besides the identification of GHS Cat 1A sensitizers, the kDPRA may be used for the prediction of sensitizer potency on a more granular, continuous scale in an integrated testing strategy (ITS) or DA. Therefore, the correlation of the log $k_{max}$ and other parameters measured in *in vitro* assays versus LLNA EC3 values was analyzed with single or multiple linear regressions to inform on (i) the predictive power of single parameters, (ii) a potential more predictive linear combination, and (iii) data redundancy when combining multiple parameters.

In principle, such correlations can be made for all data including sensitizers and non-sensitizers (Natsch et al., 2015). The result will then be influenced by the potential of an assay to predict hazard and by its potential to predict potency. In parallel, it may be more interesting to analyze only chemicals with an LLNA EC3 below a given threshold. Here we chose to focus, in a separate analysis, on chemicals with LLNA EC3 < 30%: These are the chemicals with, in general, a *bona fide* positive LLNA result and sensitization potential as revealed by the LLNA. Chemicals with an EC3 of 30-100% are in a grey zone: The majority of chemicals were only tested up to 25% when the LLNA was validated (Kolle et al., 2020), and many non-sensitizers were only tested up to such intermediary concentrations in original databases. At higher test concentrations, irritancy becomes an important confounding factor in the LLNA (see also Section 4). Hence, we cannot firmly conclude on the relevance of LLNA results with an EC3 above 30%, other than that they indicate a very weak/non-sensitizer status.

Two datasets were evaluated:

a) Set I: Chemicals with data for log $k_{max}$ AND for KeratinoSens® (n = 173; EC3 < 30% n = 121)
b) Set II: Chemicals with data for log $k_{max}$ AND KeratinoSens® AND h-CLAT AND DPRA (n = 154; EC3 < 30% n = 107)

For comparing predictivity of kDPRA and KeratinoSens® and their combination, the more comprehensive Set I was used. To evaluate predictivity in an additional combination with h-CLAT and, optionally, DPRA, the smaller Set II was used.

*Correlation of individual parameters*

First, single linear regression of each *in vitro* parameter versus the pEC3 from the LLNA was calculated. As shown in Table 6, the strongest correlation is observed for the full Set I for the log $k_{max}$ from the kDPRA ($r^2$ = 0.51). On the same set, the parameters from KeratinoSens® have a range of $r^2$ = 0.29-0.35. Limiting the evaluation to positives with EC3 < 30%, the correlation is reduced: $r^2$ = 0.40 for the kDPRA and $r^2$ = 0.13-0.17 for KeratinoSens®.

For the smaller Set II, also including h-CLAT data, the correlation to potency in general is slightly weaker: $r^2$ = 0.45 and 0.32 for log $k_{max}$ compared to all LLNA values or the chemicals with

**Tab. 6: $R^2$ coefficient for linear regression of logarithmic *in vitro* parameters vs pEC3**

| | *In vitro* parameter | Set I: (n = 173) | Set II: (n = 154) | Set I: EC3 < 30% (n = 121) | Set II: EC3 < 30% (n = 107) |
|---|---|---|---|---|---|
| kDPRA | log $k_{max}$ | 0.51 | 0.45 | 0.40 | 0.32 |
| KeratinoSens® | EC1.5 | 0.29 | 0.27 | 0.13 | 0.11 |
| | EC3 | 0.35 | 0.35 | 0.17 | 0.16 |
| | IC50 | 0.34 | 0.34 | 0.14 | 0.14 |
| h-CLAT | EC150 (CD 86) | | 0.28 | | 0.17 |
| | EC200 (CD54) | | 0.16 | | 0.04 |
| | MIT[a] | | 0.36 | | 0.20 |
| | CV75 | | 0.43 | | 0.21 |
| DPRA | kCys | | 0.33 | | 0.19 |
| | kLys | | 0.16 | | 0.17 |
| | Mean depletion | | 0.33 | | 0.20 |

[a] The MIT is a parameter that is not directly measured but a mathematical combination of two parameters, i.e. the minimal induction threshold (the lower concentration of the experimentally derived values of EC150 and EC200). As this is widely used in publications on h-CLAT, it is also provided here.

EC3 < 30%, respectively. For the parameters from the cellular assays, a range of $r^2 = 0.16$-$0.43$ is observed for all chemicals and 0.11-0.20 for chemicals with EC3 < 30%. Based on this analysis, log $k_{max}$ as a single parameter has the strongest correlation to potency among all parameters investigated in both datasets, and this is in particular true when focusing on chemicals with EC3 < 30%, i.e., those where the potential for hazard identification of the *in vitro* tests no longer contributes to the overall predictivity of potency.

All correlations shown in Table 6 are statistically highly significant at $p \leq 0.0005$ (with the exception of EC200 / Set II / EC < 30%, where $p = 0.039$). Table S8[2] additionally lists all the F values for these statistical comparisons. It has to be emphasized that these are very general comparisons, as they integrate chemicals from all chemical domains and focus on predictivity of potency in the LLNA with associated limitations regarding variability and potential to predict human sensitization potential. Also, *in vitro* assays may have a stronger correlation to potency when used in specific mechanistic domains (Natsch et al., 2015), and these benefits are not taken into account by these statistical comparisons on the entire database.

*Continuous models with multiple regressions*

Next, we performed multiple regressions of several *in vitro* parameters versus the pEC3 from the LLNA. This allows estimating whether a model incorporating several parameters or tests will significantly improve potency prediction and gives a first indication of data redundancy. Results of the key linear combinations are shown in Table 7, and more detailed information on the regression equations is provided in the supplementary file[2].

For the larger Set I, adding KeratinoSens® parameters (EC3 and IC50) to log $k_{max}$ improves $r^2$ from 0.51 to 0.61 for all chemicals and from 0.40 to 0.45 for those with EC3 < 30%. For Set II, this same improvement is from 0.45 to 0.57 for all chemicals and from 0.32 to 0.38 for those with EC3 < 30%. Similarly, when adding h-CLAT (MIT and CV75) to $k_{max}$, the $r^2$ is improved from 0.45 to 0.59 for all chemicals and from 0.32 to 0.40 for those with EC3 < 30%. Thus, adding one cellular assay improves predictivity, and this effect is more pronounced when integrating all chemicals and less so when only those with EC3 < 30% are considered. If both cell-based assays are added to log $k_{max}$, the further increase in correlation is only marginal (row 4 in Tab. 7) compared to using only one cellular assay, which indicates strong data redundancy between the cellular assays for predicting potency, as has already been observed before (Natsch et al., 2015).

In an additional evaluation, we stratified the dataset into chemicals that are positive in the LLNA between 10% and 100% and those that are positive with an EC3 < 10%. When conducting the regression analysis combining log $k_{max}$ with either the h-CLAT parameters or the KeratinoSens parameters, a very different result is obtained on the two subsets of data. For chemicals with an EC3 between 10 and 100%, cytotoxicity (either from IC50 KeratinoSens or CV75 from h-CLAT) is the key parameter correlating to LLNA potency, while log $k_{max}$ is the key predictor if the EC3 is below 10% (Tab. S9[2]). This at least partly explains why adding the parameters from the cell-based assays improves the overall correlation for the full potency range, despite the fact that log $k_{max}$ appears sufficient to predict the GHS 1A potency class.

When combining all *in vitro* parameters from classical DPRA, h-CLAT and KeratinoSens®, the quantitative correlation to potency for the chemicals with EC3 < 30% is weaker ($r^2 = 0.27$) as compared to log $k_{max}$ alone ($r^2 = 0.32$) or the combination of log $k_{max}$ and one cellular assay ($r^2 = 0.38$-$0.40$). This is a further indication that log $k_{max}$ is the strongest contributing parameter for potency of all the validated *in vitro* assays and an important contribution to an optimized linear model, which is in line with the

**Tab. 7: $R^2$ coefficient for multiple linear regression of logarithmic *in vitro* parameters vs pEC3**

| | Set I: (n = 173) | Set II: (n = 154) | Set I: EC3 < 30% (n = 121) | Set II: EC3 < 30% (n = 107) |
|---|---|---|---|---|
| log $k_{max}$ | 0.51 | 0.45 | 0.40 | 0.32 |
| KS[a] + log $k_{max}$ | 0.61 | 0.57 | 0.45 | 0.38 |
| h-CLAT[b] + log $k_{max}$ | | 0.59 | | 0.40 |
| h-CLAT + KS + log $k_{max}$ | | 0.60 | | 0.41 |
| h-CLAT + KS + DPRA[c] | | 0.54 | | 0.27 |
| h-CLAT + KS | | 0.51 | | 0.27 |

[a] KS, KeratinoSens®. For KS, the parameters IC50 for 50% cytotoxicity and EC3 for 3-fold luciferase induction were used. [b] For h-CLAT, MIT, i.e., the minimal value of EC150 and EC200 and CV75, concentration for 25% reduction in viability, was used. [c] For DPRA, log Cys k measured in the classical DPRA was used.

**Tab. 8: Predictivity (in %) of linear regression models for sub-classification of chemicals with EC3 < 30% into GHS 1A and 1B chemicals**

| | n | Regression equation[a] | Sensitivity[b] (%) | Specificity (%) | Balanced accuracy (%) |
|---|---|---|---|---|---|
| *Set I* | | | | | |
| log $k_{max}$ threshold -2 | 125 | | 84.4 | 77.8 | 81.1 |
| log $k_{max}$ regression model[d] | 125 | Eq. S1[1] | 82.2[c] | 75.3 | 78.8 |
| KS + log $k_{max}$ | 125 | Eq. S2[1] | 82.2 | 71.6 | 76.9 |
| *Set II* | | | | | |
| $k_{max}$ threshold -2 | 106 | | 81.1 | 77.1 | 79.1 |
| KS + log $k_{max}$ | 106 | Eq. S4[1] | 78.4 | 75.7 | 77.0 |
| h-CLAT + log $k_{max}$ | 106 | Eq. S5[1] | 81.1 | 77.1 | 79.1 |
| **h-CLAT + KS + log $k_{max}$** | 106 | Eq. S6[1] | 75.7 | 80.0 | 77.8 |
| h-CLAT + KS | 106 | Eq. S7[1] | 67.6 | 78.6 | 73.1 |
| h-CLAT + KS + DPRA | 106 | Eq. S8[1] | 91.6 | 62.9 | 77.3 |

[a] The underlying regression analysis is shown in the training video[1]. [b] Here, prediction statistics are given for identifying GHS 1A sensitizers within the chemicals with an LLNA EC3 < 30%. [c] For each chemical, the predicted EC3 was calculated with the regression models summarized in Table 7 column 4 and 5 and in the supplementary file[1], and chemicals with predicted EC3 < 2% were assigned to GHS 1A. [d] Note: For verification of the analysis, the average molecular weight of 185 g/mol for the chemicals with an EC3 < 30% and available $k_{max}$ values was used and entered along an EC3 = 2% into the regression equation with $k_{max}$ as single predictor (Eq. S1: pEC3 = 2.652 + 0.3491 × log $k_{max}$), and the equation was solved for log $k_{max}$. A log $k_{max}$ = -1.96 was obtained as decision threshold indicating that we obtain a very similar prediction threshold with the regression analysis and with the ROC analysis. KS, KeratinoSens®

view that protein modification as MIE is an important rate-limiting step for the acquisition of skin sensitization.

*Subclassification of chemicals into GHS Cat 1A based on multiple regression*

Quantitative models integrating multiple parameters are of key interest for quantitative risk assessment (Leontaridou et al., 2016) and especially when evaluating potency in specific chemical domains (Natsch et al., 2015). But could a model with multiple parameters also better predict and separate GHS Cat 1A chemicals from GHS Cat 1B/NC chemicals as compared to kDPRA alone for regulatory classification? We thus used the models in Table 7 (i.e., the regression equations underlying the reported $r^2$ values, which are given in the supplementary file[2]) and calculated the predicted LLNA EC3 for each chemical according to these equations. Chemicals with a predicted EC3 < 2% were then assigned a GHS 1A subcategory, which was compared to the *in vivo* value. We show the results in Table 8, here only for the chemicals with LLNA EC3 < 30%. The same analysis was also done with all chemicals, in that case specificity and balanced accuracy is

higher as all the chemicals with weak/no response in LLNA and *in vitro* tests are included, boosting specificity, while sensitivity, by definition, remains the same.

In the first row in Table 8, the predictivity of the kDPRA when applying the simple prediction model with a threshold of log $k_{max}$ > -2 for GHS Cat 1A attribution ("log $k_{max}$ threshold -2") is repeated for these subsets of chemicals. A regression model with log $k_{max}$ as single predictor is calculated and used for GHS Cat 1A attribution (row 2 in Tab. 8) in order to compare like for like when making the comparison to multiple regression models. As shown in Table 8, sensitivity for identifying GHS Cat 1A sensitizers is at 85% with the threshold prediction model and slightly lower (84%) with the regression model using $k_{max}$ alone (see also detailed analysis of these two approaches for subclassification in Tab. S5-S7[2]). By adding further parameters from Keratino-Sens® and/or h-CLAT to the regression model, sensitivity and balanced accuracy are not enhanced. A model with h-CLAT, KeratinoSens® and classical DPRA data (Lys- and Cys-depletion transformed to rate constants) can reach higher sensitivity but, not surprisingly, at significant costs regarding specificity.

This analysis thus further indicates that, overall and for this relatively large database, a decision model for GHS Cat 1A attribution based on log $k_{max}$ alone has a high predictivity and can be used as stand-alone model for the identification of GHS Cat 1A sensitizers.

## 4 Discussion

This study provides detailed information on the predictivity of the kinetic rate constant of the validated kDPRA for sensitizer potency assessment, both for subclassification according to the GHS and as important input parameter for deriving a continuous PoD for risk assessment. Statistical analysis clearly indicates that the rate constant is the best single predictor for potency of any validated *in vitro* test. Still, it is important to carefully discuss its limitations.

First of all, the assay only measures reactivity with the Cys-peptide. This may limit applicability/predictivity, as some sensitizers exclusively react with lysine residues, such as some acyl-halides, phenol-esters or aldehydes. However, only few Cat 1A sensitizers with this reactivity pattern are known. Glutaric aldehyde is such a case. Only few other Cat 1A sensitizers with pre-dominant lysine reactivity, such as the o-hydroxy benzaldehydes atranol and chloratranol (Natsch et al., 2012), are currently known. The analysis of a large database (Urbisch et al., 2015) on Cys- and Lys-depletion by skin sensitizers indicates that 88% of skin sensitizers positive in the DPRA react more strongly with the Cys-peptide, and among the GHS Cat 1A chemicals according to the LLNA, only glyoxal, glutaric aldehyde and phthalic anhydride deplete the Lys-peptide more strongly after 24 h (Urbisch et al., 2015), although anhydrides have a strong Cys-reactivity if evaluated at earlier timepoints and are thus correctly identified by the kDPRA. Considering exclusive and strong Lys-reactivity in the DPRA (with a threshold of Lys-depletion yet to be defined) as criterion for a 1A classification in a tiered strategy is an option to reduce this uncertainty.

Pre-haptens that are activated to strong haptens, such as 4-phenylenediamine, are further cases that may be underestimated in the kDPRA – the time needed for them to oxidize will reduce the apparent kinetic rate constant of the reaction with the test peptide. Thus, chemicals that are spontaneously but not instantly transformed to very reactive species may lead to some underestimation of the sensitization potential if the lag-period for oxidation is in the range of hours. Thus, in a regulatory setting, aromatic amines, catechols or hydroquinones may be considered outside of the applicability domain of the test or may require further data to confirm their weak reactivity even under oxidizing conditions before attributing them to class 1B in case the measured log $k_{max}$ is > -2.

*In vitro* investigations (Patlewicz et al., 2016; Urbisch et al., 2016) using compounds requiring molecular transformation to attain a sensitizing potential have shown that most pre-haptens can readily be detected in the DPRA. Moreover, many pro-haptens are also activated by non-enzymatic oxidation (and therefore are both pre- and pro-haptens). As discussed widely before, the DPRA does not contain a metabolic system (Urbisch et al., 2016), and thus prediction of strict pro-haptens requiring metabolic activation (i.e., not acting as direct haptens nor pre-haptens) in theory is a significant limitation for the (k)DPRA. From the database evaluated so far, two compounds were indeed identified that are potential pro-haptens and clearly false-negative in human GHS 1A classification (3-dimethylaminopropylamine and diethylenetriamine). These are two out of the ten clear human GHS Cat 1A false-negatives. However, both chemicals are rated as 1B in the LLNA, and actually none of the six false-negatives versus LLNA (Tab. 3) is a presumed strict pro-hapten. The good predictivity to identify GHS Cat 1A sensitizers indeed indicates that the limitation for pro-haptens is not a dramatic shortcoming. Is there a scientific explanation for this observation? Indeed, most *bona fide* pro-haptens, for which activation by metabolic systems is well described and understood, are weak to moderate sensitizers in the LLNA (e.g., dihydroeugenol, eugenol, cinnamic alcohol, ethylene diamine, geraniol). Since the skin is a rather poor metabolic organ, slow metabolic activation may often be a rate-limiting step for sensitization by pro-haptens, thus rendering them less potent allergens. This may explain why an assay without metabolic activation may recognize most strong sensitizers, which is not always the case when it comes to detecting moderate or weak allergens.

The LLNA is measured in %, thus in a weight per volume metric, while the rate constant is measured in molar terms. This may appear as a discrepancy, as we use a fixed threshold for log $k_{max}$ to predict a regulatory threshold of 2% in the LLNA, not taking molecular weight into account. We thus performed an additional analysis, whereby the rate constant is transformed to a [$\%^{-1}s^{-1}$] value, and we then again performed the ROC analysis for an optimal prediction threshold. Alternatively, we also used a prediction based on Equation S1[2], which gives a predicted EC3 value based on a molar correlation (pEC3 versus log $k_{max}$). As shown in the supplementary file[2], both approaches agree for 174 of the 180 chemicals with the approach using a threshold of $k_{max}$ > -2.0, and the balanced accuracy is slightly reduced to 84.4 and 83% with these ap-

proaches. Therefore, using the simple cut-off of log $k_{max}$ > -2 as a regulatory threshold is warranted for predictive purposes, but we would propose to calculate EC3 values for modeling approaches for risk assessment using molar values for the EC3 (i.e., pEC3 as done in the regression analysis presented here).

There are now multiple validated *in vitro* tests, and several DAs have been proposed for sensitizer hazard assessment based on these *in vitro* tests. So how could the kDPRA best be used in a regulatory context? The most straightforward approach is to use one of the DAs that are in advanced validation at the OECD (2019a) for hazard identification. In case a chemical is predicted as a sensitizer, the kDPRA could be conducted and chemicals with a log $k_{max}$ > -2.0 would be attributed to GHS Cat 1A. Optionally, chemicals with exclusive strong lysine reactivity (strong Lys-peptide depletion, which is not due to co-elution, in absence of Cys-depletion in DPRA) could also be attributed to GHS Cat 1A in case reactivity surpasses a yet to be defined threshold. However, this would be an additional correction only required for few chemicals.

There is already one proposal for identification of GHS Cat 1A under discussion at OECD level, namely a modification of the ITS proposed by Nukada et al. (2013). However, the ITS approach is based on concentration-response data from the h-CLAT and reactivity classes of the DPRA, and to our knowledge the reproducibility of these quantitative data so far has not been documented. In addition, it uses an *in silico* input, and it is not clear to which extent the underlying *in silico* tool was trained on the chemicals used to assess its performance (Kolle et al., 2020). Thus, the approach of integrating kDPRA has the advantage that it is based on an experimentally determined and validated quantitative data input.

Based on the analysis in Table 8, subclassification based on kDPRA as single input maximizes predictivity, and no further potency data from the other assays is required for optimal GHS subclassification of those chemicals that have been identified as sensitizers by any other test or DA. It is also interesting to compare the balanced accuracy for class attribution to the variability of the LLNA itself to predict 1A sensitizers. A chance of 73% was reported to correctly identify a 1A sensitizer with the LLNA (OECD, 2019b) in three consecutive tests, thus the chance for a correct assignment in a single LLNA is $p = 0.93$ ($0.93^3 = 0.73$). In light of this intrinsic variability of the LLNA itself, a sensitivity significantly above 90% to predict the LLNA subclassification outcome cannot be expected of any test or DA, hence the reported sensitivity of 84.4% achieved by the kDPRA here appears adequate.

Besides subclassification into GHS subcategories, deriving a NESIL for risk assessment is of key importance. It has previously been shown how kinetic rate constants can be applied to the prediction of skin sensitization potency (Natsch et al., 2011, 2015, 2018; Niklasson et al., 2011; Roberts and Aptula, 2014). Here we further show that of all the measured parameters in the validated *in vitro* assays, log $k_{max}$ from the kDPRA is the parameter with the strongest correlation to sensitizer potency in the LLNA, and Table S9[2] further shows its importance especially to predict potency of moderate and strong sensitizers. Thus log $k_{max}$ from the kDPRA appears to be an important input parameter for a DA for potency assessment on a continuous scale, and it can be combined with read-outs from other *in vitro* assays, structural information on chemical domain, and other reactivity parameters in a more comprehensive analysis of the sensitization risk, as has been shown before for a similar assay based on the Cor1-C420 peptide (Natsch et al., 2015, 2018).

The LLNA is not a direct sensitization assay, as it only measures the proxy of cell proliferation in the lymph node upon chemical administration to mouse ears (Kolle et al., 2020). It was long known that the LLNA is prone to false-positives, especially for cytotoxic irritants (Ball et al., 2011) employed at high test concentrations. The analysis of our database as presented in Table S9[2] indicates that cytotoxicity, which is closely related to skin irritation potential (Kandárová et al., 2006), is an important predictor for EC3 values observed at high test concentrations in the LLNA, further illustrating this limitation of the LLNA in discriminating irritants and sensitizers when conducted at high test concentrations. On the other hand, our data corroborate the correlation between LLNA and reactivity for EC3 values below 10%, thus further confirming the link between the MIE of skin sensitization and cell proliferation in the LLNA.

Here we have built the database in a way that data are now available for many chemicals from tests performed according to OECD 442D and 442E along with the kinetic rate constant from the kDPRA. This will facilitate refinement of existing approaches and development of new approaches. A multiple linear regression as shown in Table 7 and associated supplementary information[2] is just one example of such a predictive DA – other mathematical ways to integrate the information such as neuronal nets, Bayesian nets or other non-linear models might even be superior. A positive kDPRA result (i.e., reactivity above the cysteine-only prediction model of the DPRA) could be used to replace the DPRA in new and existing DAs, while it may need further discussion whether a reactivity assay on the Lysine-peptide may still be needed in case of negative results.

The discussion above summarizes the limitations of the current protocol from the standpoint of predictivity. The standardized protocol also comes with a number of simplifications in order to allow deriving rate constants for diverse chemicals in a standardized setting as needed for regulatory toxicology. Thus, many chemicals may not react linearly over time with a different apparent rate constant over time under our experimental conditions or not be fully soluble in the reaction medium and thus precipitate out of solution in the course of the experiment. Other chemicals may form semi-stable peptide adducts that reach an equilibrium or are prone to a reverse reaction over time. All these effects may lead to a reaction that does not proceed to completion and that may lead to non-linearity (i.e., the rate determined at different points in time is different), which is not fully reflected in the way the data are treated. By deriving maximum rate constants in the current protocol, mainly the initial rate of reaction determines $k_{max}$ for most chemicals. Analysis of our comprehensive database indicates that this approach already has good predictivity, but it remains an open question whether a different protocol or an array of protocols for different types of chemicals accounting

for such non-linear effects would determine rate constants with an even better correlation with LLNA potency.

Finally, the assay measures depletion of the peptide and not formation of peptide-adducts, which is the ultimate MIE for sensitization to occur. As shown in detailed case studies (Natsch et al., 2018), an additional evaluation of adduct formation can further advance risk assessment for specific chemicals. From a conceptual point of view, the approach proposed by Sanderson et al. (2016) would be most preferable, as it directly measures rate constants of adduct formation by real-time NMR analysis. However, we currently do not see how that approach could be adapted to screen a large library of chemicals and to arrive at an OECD test guideline.

Further improvements of the assay may include, e.g., following the reaction with an alternative fluorescent probe with different excitation/emission wavelengths to test chemicals that cause fluorescence interference. Furthermore, a pre-incubation step of the chemical in the test buffer to allow for spontaneous oxidation may help to better identify strong pre-haptens (confirmed by preliminary experiments for PPD; data not shown).

The kDPRA validation, which is described in the parallel paper (Wareing et al., 2020) underwent an international peer-review with experts nominated by ECVAM, NICEATM and JaCVAM. The peer-review panel concluded that "*the data provided by the test method developers is sufficient and adequately supports the scientific validity of the kDPRA for the identification of UN GHS Subcat. 1A*." It will be further discussed by the OECD for integration into TG 442C.

## References


Alvarez-Sánchez, R., Basketter, D., Pease, C. et al. (2004). Covalent binding of the 13C-labeled skin sensitizers 5-chloro-2-methylisothiazol-3-one (MCI) and 2-methylisothiazol-3-one (MI) to a model peptide and glutathione. *Bioorg Med Chem Lett 14*, 365-368. doi:10.1016/j.bmcl.2003.11.002

Api, A. M., Basketter, D. A., Cadby, P. A. et al. (2008). Dermal sensitization quantitative risk assessment (QRA) for fragrance ingredients. *Regul Toxicol Pharmacol 52*, 3-23. doi:10.1016/j.yrtph.2007.10.008

Ball, N., Cagen, S., Carrillo, J. C. et al. (2011). Evaluating the sensitization potential of surfactants: Integrating data from the local lymph node assay, guinea pig maximization test, and in vitro methods in a weight-of-evidence approach. *Regul Toxicol Pharmacol 60*, 389-400. doi:10.1016/j.yrtph.2011.05.007

Barentsen, H. M., Jonis, S. U., Pelgrom, S. et al. (2019). REACH alternative testing strategy for skin sensitization in practice: Fact or fiction? *Regul Toxicol Pharmacol 106*, 292-302. doi:10.1016/j.yrtph.2019.05.019

Basketter, D. A., Alepee, N., Ashikaga, T. et al. (2014). Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis 25*, 11-21. doi:10.1097/DER.0000000000000003

Basketter, D. and Safford, B. (2016). Skin sensitization quantitative risk assessment: A review of underlying assumptions. *Regul Toxicol Pharmacol 74*, 105-116. doi:10.1016/j.yrtph.2015.11.013

Delaine, T., Ponting, D. J., Niklasson, I. B. et al. (2014). Epoxyalcohols: Bioactivation and conjugation required for skin sensitization. *Chem Res Toxicol 27*, 1860-1870. doi:10.1021/tx500297d

Gerberick, G. F., Vassallo, J. D., Bailey, R. E. et al. (2004). Development of a peptide reactivity assay for screening contact allergens. *Toxicol Sci 81*, 332-343. doi:10.1093/toxsci/kfh213

Hirota, M., Ashikaga, T. and Kouzuki, H. (2018). Development of an artificial neural network model for risk assessment of skin sensitization using human cell line activation test, direct peptide reactivity assay, KeratinoSens and in silico structure alert parameter. *J Appl Toxicol 38*, 514-526. doi:10.1002/jat.3558

Hoffmann, S., Kleinstreuer, N., Alepee, N. et al. (2018). Non-animal methods to predict skin sensitization (I): The Cosmetics Europe database. *Crit Rev Toxicol 48,* 344-358. doi:10.1080/10408444.2018.1429385

ICCVAM (2011). ICCVAM Test Method Evaluation Report. Usefulness and Limitations of the Murine Local Lymph Node Assay for Potency Categorization of Chemicals Causing Allergic Contact Dermatitis in Humans. *NIH Publication No. 11-7709*. https://ntp.niehs.nih.gov/iccvam/docs/immunotox_docs/llna-pot/tmer.pdf

Jaworska, J., Dancik, Y., Kern, P. et al. (2013). Bayesian integrated testing strategy to assess skin sensitization potency: From theory to practice. *J Appl Toxicol 33*, 1353-1364. doi:10.1002/jat.2869

Jaworska, J. S., Natsch, A., Ryan, C. et al. (2015). Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: A decision support system for quantitative weight of evidence and adaptive testing strategy. *Arch Toxicol 89*, 2355-2383. doi:10.1007/s00204-015-1634-2

Kandárová, H., Liebsch, M., Schmidt, E. et al. (2006). Assessment of the skin irritation potential of chemicals by using the SkinEthic reconstructed human epidermal model and the common skin irritation protocol evaluated in the ECVAM skin irritation validation study. *Altern Lab Anim 34*, 393-406. doi:10.1177/026119290603400407

Kolle, S. N., Landsiedel, R. and Natsch, A. (2020). Replacing the refinement for skin sensitization testing: Considerations to the implementation of adverse outcome pathway (AOP)-based defined approaches (DA) in OECD guidelines. *Regul Toxicol Pharmacol 115*, 104713. doi:10.1016/j.yrtph.2020.104713

Leontaridou, M., Gabbert, S., Van Ierland, E. C. et al. (2016). Evaluation of non-animal methods for assessing skin sensitisation hazard: A Bayesian value-of-information analysis. *Altern Lab Anim 44*, 255-269. doi:10.1177/026119291604400309

Natsch, A., Gfeller, H., Rothaupt, M. et al. (2007). Utility and limitations of a peptide reactivity assay to predict fragrance allergens in vitro. *Toxicol In Vitro 21*, 1220-1226. doi:10.1016/j.tiv.2007.03.016

Natsch, A., Haupt, T. and Laue, H. (2011). Relating skin sensitizing potency to chemical reactivity: Reactive Michael acceptors inhibit NF-κB signaling and are less sensitizing than S(N)Ar- and S(N)2-reactive chemicals. *Chem Res Toxicol 24*, 2018-2027. doi:10.1021/tx2003678

Natsch, A., Gfeller, H., Haupt, T. et al. (2012). Chemical reactivity and skin sensitization potential for benzaldehydes: Can Schiff

base formation explain everything? *Chem Res Toxicol 25*, 2203-2215. doi:10.1021/tx300278t

Natsch, A., Ryan, C. A., Foertsch, L. et al. (2013). A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. *J Appl Toxicol 33*, 1337-1352. doi:10.1002/jat.2868

Natsch, A., Emter, R., Gfeller, H. et al. (2015). Predicting skin sensitizer potency based on in vitro data from KeratinoSens and kinetic peptide binding: Global versus domain-based assessment. *Toxicol Sci 143*, 319-332. doi:10.1093/toxsci/kfu229

Natsch, A., Emter, R., Haupt, T. et al. (2018). Deriving a no expected sensitization induction level for fragrance ingredients without animal testing: An integrated approach applied to specific case studies. *Toxicol Sci 165*, 170-185. doi:10.1093/toxsci/kfy135

Niklasson, I. B., Delaine, T., Luthman, K. et al. (2011). Impact of a heteroatom in a structure-activity relationship study on analogues of phenyl glycidyl ether (PGE) from epoxy resin systems. *Chem Res Toxicol 24*, 542-548. doi:10.1021/tx100417r

Nukada, Y., Miyazawa, M., Kazutoshi, S. et al. (2013). Data integration of non-animal tests for the development of a test battery to predict the skin sensitizing potential and potency of chemicals. *Toxicol In Vitro 27*, 609-618. doi:10.1016/j.tiv.2012.11.006

OECD (2010). Test No. 429: Skin Sensitization: Local Lymph Node Assay. *OECD Guidelines for the Testing of Chemicals, Section 4*. OECD Publishing, Paris. doi:10.1787/9789264071100-en

OECD (2012). The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins, Part 1: Scientific Evidence. *OECD Series on Testing and Assessment, No. 168*. OECD Publishing, Paris. doi:10.1787/9789264221444-en

OECD (2018a). Test No. 442D: In Vitro Skin Sensitisation Assays Addressing the AOP Key Event on Keratinocyte Activation. *OECD Guidelines for the Testing of Chemicals, Section 4*, OECD Publishing, Paris. doi:10.1787/9789264229822-en

OECD (2018b). Test No. 442E: In Vitro Skin Sensitisation Assays Addressing the Key Event on Activation of Dendritic Cells on the Adverse Outcome Pathway for Skin Sensitisation. *OECD Guidelines for the Testing of Chemicals, Section 4*, OECD Publishing, Paris. doi:10.1787/9789264264359-en

OECD (2019a). Draft OECD Guideline Defined Approaches for Skin Sensitisation. *OECD Series on Testing and Assessment*. https://www.oecd.org/env/ehs/testing/GL%20DASS_22Sep2019v2.pdf

OECD (2019b). Supporting document for evaluation and review of draft Guideline (GL) for Defined Approaches (DAs) for Skin Sensitisation. *Series on Testing and Assessment*. https://www.oecd.org/env/ehs/testing/latestdocuments/DAGL%20supporting%20document_8Oct2018_v2_Clean.pdf

OECD (2020). 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, Paris. doi:10.1787/9789264229709-en

Patlewicz, G., Casati, S., Basketter, D. A. et al. (2016). Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? *Regul Toxicol Pharmacol 82*, 147-155. doi:10.1016/j.yrtph.2016.08.007

Roberts, D. W. and Aptula, A. O. (2008). Determinants of skin sensitisation potential. *J Appl Toxicol 28*, 377-387. doi:10.1002/jat.1289

Roberts, D. W. and Natsch, A. (2009). High throughput kinetic profiling approach for covalent binding to peptides: Application to skin sensitization potency of Michael acceptor electrophiles. *Chem Res Toxicol 22*, 592-603. doi:10.1021/tx800431x

Roberts, D. W. and Aptula, A. O. (2014). Electrophilic reactivity and skin sensitization potency of SNAr electrophiles. *Chem Res Toxicol 27*, 240-246. doi:10.1021/tx400355n

Sanderson, P. N., Simpson, W., Cubberley, R. et al. (2016). Mechanistic understanding of molecular initiating events (MIEs) using NMR spectroscopy. *Toxicol Res (Cambridge) 5*, 34-44. doi:10.1039/c5tx00246j

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

Urbisch, D., Becker, M., Honarvar, N. et al. (2016). Assessment of pre- and pro-haptens using nonanimal test methods for skin sensitization. *Chem Res Toxicol 29*, 901-913. doi:10.1021/acs.chemrestox.6b00055

Warburton, K. L., Bauer, A., Chowdhury, M. M. U. et al. (2015). ESSCA results with the baseline series, 2009-2012: Rubber allergens. *Contact Dermatitis 73*, 305-312. doi:10.1111/cod.12454

Wareing, B., Urbisch, D., Kolle, S. N. et al. (2017). Prediction of skin sensitization potency sub-categories using peptide reactivity data. *Toxicol In Vitro 45*, 134-145. doi:10.1016/j.tiv.2017.08.015

Wareing, B., Kolle, S. N., Birk, B. et al. (2020). The kinetic direct peptide reactivity assay (kDPRA): Intra- and inter-laboratory reproducibility in a seven-laboratory ring trial. *ALTEX 37*, 639-651. doi:10.14573/altex.2004291

Warshaw, E. M., Kwon, G. P., Mathias, C. G. T. et al. (2013). Occupationally related contact dermatitis in North American food service workers referred for patch testing, 1994 to 2010. *Dermatitis 24*, 22-28. doi:10.1097/DER.0b013e31827b14e1

Warshaw, E. M., Hagen, S. L., DeKoven, J. G. et al. (2017). Occupational contact dermatitis in North American production workers referred for patch testing: Retrospective analysis of cross-sectional data from the North American contact dermatitis group 1998 to 2014. *D ermatitis 28*, 183-194. doi:10.1097/DER.0000000000000277

Yamamoto, Y., Tahara, H., Usami, R. et al. (2015). A novel in chemico method to detect skin sensitizers in highly diluted reaction conditions. *J Appl Toxicol 35*, 1348-1360. doi:10.1002/jat.3139

## Conflict of interest

The authors declare that they have no conflicts of interest.

## Acknowledgements

This work was entirely funded by Givaudan Schweiz SA and BASF SE.