



## Research Article

# Integrated Skin Sensitization Assessment Based on OECD Methods (III): Adding Human Data to the Potency Assessment

Andreas Natsch

Fragrances S&T, Ingredients Research, Givaudan Schweiz AG, Kempthal, Switzerland

### Abstract

Skin sensitizer potency assessment based on new approach methodologies is key to deriving a point of departure (PoD) for risk assessment. Regression models to predict a PoD based on OECD-validated *in vitro* tests and trained on local lymph node assay (LLNA) data were previously presented, and results from human tests were recently compiled. To integrate both data sources, the Reference Chemical Potency List (RCPL), which provides potency values (PV) for 33 chemicals integrating LLNA and human data in a structured weight-of-evidence approach, was developed. When calculating regression models vs PV or LLNA data, different weights for the input parameters were noted. As the RCPL is based on too few chemicals to train robust statistical models, the list of human data was extended to a larger set of PV ( $n = 139$ ) with associated *in vitro* data. This database was used to retrain the regression models and to compare regression models trained vs (i) LLNA, (ii) PV or (iii) human DSA04 values. Using the PV as a target, predictive models of similar predictivity to the LLNA-based models were obtained, which mainly differ in a lesser weight of cytotoxicity and a higher weight of cell activation and reactivity parameters. Analysis of the human DSA04 dataset indicates a similar pattern but also shows that the human dataset is too small and biased to be a key dataset for potency prediction. Hence, an enlarged set of PV values appears to be a complementary tool to train predictive models next to an LLNA-only database.

## 1 Introduction

Development of predictive tools for the skin sensitization endpoint has been a key focus in the area of new approach methodologies (NAM) to replace animal tests, culminating in the publication by the OECD of test guidelines (TGs) based on three key events of the sensitization adverse outcome pathway (AOP) (OECD, 2014). So far, these TGs cover nine different test protocols (Natsch and Gerberick, 2022a; OECD, 2018a,b, 2023). This was followed by the first OECD TG implementing defined approaches (DA), which combine several test methods to an integrated assessment of the skin sensitization endpoint (OECD, 2021a). These DA use the methods that were first implemented as OECD TGs, namely the direct peptide reactivity assay (DPRA) (Gerberick et al., 2004), the KeratinoSens assay (KS) (Emter et al., 2010), and the human cell line activation assay (h-CLAT) (Sakaguchi et al., 2006). These individual methods and the DA primarily yield a binary output to determine the sensitization hazard of chemicals. In addition, the DA “integrated testing strategy

(ITS)” (Takenouchi et al., 2015) and the kinetic direct peptide reactivity assay (kDPRA) (Natsch et al., 2020; Wareing et al., 2020) were accepted by the OECD as approaches to discriminate the Globally Harmonized System (GHS) potency classes 1A and 1B. However, next to discriminating potency classes, sensitization potency information is key for deriving a point of departure (PoD) for quantitative risk assessment (QRA) (Api et al., 2008) on a continuous scale. The assays used in the DA as well as the kDPRA yield continuous quantitative data in addition to the binary or categorical outcome. These data can be used to derive a PoD for risk assessment, and several models have been proposed to integrate such data (Gilmour et al., 2022; Jaworska et al., 2015; Natsch et al., 2018; Reynolds et al., 2022). In our most recent analysis, data from OECD-validated methods were combined in multiple regression models to predict a PoD (Natsch and Gerberick, 2022a). These models combine quantitative results from the KS and/or the h-CLAT with the kinetic rate constants from the kDPRA and predict an estimate of the local lymph node assay (LLNA) EC<sub>3</sub> value on a continuous scale. We further demon-

Received February 8, 2023; Accepted April 17, 2023;  
Epub April 17, 2023; © The Authors, 2023.

ALTEX 40(4), 571-583. doi:10.14573/altex.2302081

Correspondence: Andreas Natsch, PhD  
Givaudan Schweiz AG,  
Kemptpark 50, 8310 Kempthal, Switzerland  
(andreas.natsch@givaudan.com)

This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is appropriately cited.



strated how this assessment can then be combined with the DA “2 out of 3” as implemented in TG 497 (Natsch and Gerberick, 2022b), as the same test methods used for the regulatory classification are applied for the PoD assessment. These models, similar to the detailed Bayesian models developed by Jaworska et al. (2013, 2015), were trained on EC3 values from the LLNA as the *in vivo* measure of sensitization potency. The EC3 gives the concentration for a three-fold stimulation of cell proliferation in the draining lymph nodes of the treated mouse ear.

The LLNA offers the most comprehensive source for *in vivo* potency data, and it is also the only data source with dose-response data, while tests in humans or on guinea pigs are mostly done at only one test concentration. Hence, the LLNA data appear most attractive to develop potency predictions in terms of data quality and size of the available database. However, the LLNA measures overall cell proliferation in the local lymph node as a proxy for the induction of skin sensitization and not the proliferation of hapten-specific T-cells, which would be the true endpoint we would like to predict. In addition, we might be faced with a species difference when using mice instead of human data. Nevertheless, this effect is generally considered minor for the sensitization endpoint as all key human sensitizers can sensitize laboratory animals. Furthermore, a correlation between human potency expressed as DSA05 and potency in mice expressed as LLNA EC3 with a slope close to one and a y-intercept close to zero has been reported, indicating that there is no fundamental species difference here (Api et al., 2015; Bil et al., 2017).

Next to LLNA data, results from human tests have been compiled, e.g., by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) (ICCVAM, 2008, 2011) and the OECD expert group (OECD, 2021d). These data mostly originate from the human repeat insult patch test (HRIPT) (Politano and Api, 2008) or the human maximization test (HMT) (Kligman, 1966). The human tests truly measure sensitization in the species of interest, as they test for clinical hypersensitivity reactions after an induction phase with repeated exposure. However, as these human data lack dose-response information and come from much less standardized protocols, it remains an open question which data source is optimal to develop predictive models for sensitization potency.

Next to human and LLNA data, data from guinea pig maximization (GPMT) and Buehler tests are available. They were used as key information in the validation of the LLNA (Dean et al., 2001). While it is difficult to derive potency information from guinea pig data, they may be consulted in a weight of evidence (WoE) on whether a chemical is a relevant skin sensitizer (Emter et al.,

2010), and they have the advantage that they also truly measure elicitation of skin sensitization after an induction phase.

Recently, the Reference Chemical Potency List (RCPL), which presents potency values (PV) for 33 chemicals in an attempt to integrate both LLNA and human data in a structured WoE approach following discrete workflows, was published (Irizar et al., 2022). For studies with positive cases of sensitization, the human data were expressed as DSA04 values, i.e., the estimated dose per surface area (DSA, expressed in  $\mu\text{g}/\text{cm}^2$ ) of a chemical that would be required to induce skin sensitization in 4% of exposed subjects under the conditions of an HRIPT or HMT. The DSA04 value was selected in preference to any other DSA value as it was found to correlate best with LLNA EC3 values on a set of 20 chemicals and thus should be a metric directly comparable to the EC3 expressed as dose per area. Although the evidence to give preference to the DSA04 over, e.g., the DSA05 (as used in ICCVAM studies) (ICCVAM, 2011) or the DSA02 is limited, this work is based on the DSA04 to build on the analysis made for the RCPL. The aim of the RCPL is to serve as benchmark to evaluate the ability of NAM (and quantitative DA based on NAM data) to rank the sensitizer potency of chemicals and to compare different approaches for PoD determination. However, the RCPL may be based on too few chemicals to train robust statistical models for potency determination, especially if they are based on multiple input parameters as the higher the number of input variables, the greater the risk of obtaining over-fitted models. In addition, a key focus of the RCPL is to assess potency prediction for fragrance allergens, and thus the chemical selection was intentionally biased towards fragrance ingredients. Therefore, there is a risk of limited coverage of the chemical space for any model trained on this small dataset.

Here we first developed an extended WoE list of chemicals with DSA04 values or other robust sensitization evidence and derived a PV for these chemicals according to the RCPL workflows. Based on the same *in vitro* input data from the previous, LLNA-focused analysis (Natsch and Gerberick, 2022a,b), regression models were trained vs (i) LLNA EC3, (ii) PV values, and (iii) human DSA04 values in the different data subsets in order to compare predictivity for different *in vivo* targets.

## 2 Materials and methods

### *In vitro database, data normalization and regression analysis*

This paper is based on the database of *in vitro* and LLNA data published previously as ESM1<sup>1</sup> by Natsch and Gerberick

<sup>1</sup> doi:10.14573/altex.2201141s1

### Abbreviations

AOP, adverse outcome pathway; DA, defined approach; DASS, defined approaches for skin sensitization; DNCB, 2,4-dinitro-chlorobenzene; DPRA, direct peptide reactivity assay; DSA, dose per surface area; ECHA, European Chemicals Agency; EC3, estimated concentration for threefold cell proliferation; GHS, Globally Harmonized System; GPMT, guinea pig maximization test; h-CLAT, human cell line activation test; HMT, human maximization test; HRIPT, human repeat insult patch test; IATA, integrated approaches to testing and assessment; ICCVAM, Interagency Coordinating Committee on the Validation of Alternative Methods; IFRA, International Fragrance Association; ITS, integrated testing strategy; kDPRA, kinetic direct peptide reactivity assay; KS, KeratinoSens™; LLNA, local lymph node assay; NAM, new approach methodology; NESIL, no expected sensitization induction level; NOEL, no observed effect level; PoD, point of departure; PV, potency values; OECD, Organisation for Economic Co-operation and Development; QRA, quantitative risk assessment; RCPL, Reference Chemical Potency List; RIFM, Research Institute for Fragrance Materials; SDS, sodium dodecyl sulphate; TG, test guideline; WoE, weight of evidence

(2022a), which contained also much of the data analysis conducted by the OECD Defined Approaches for Skin Sensitization (DASS) expert group (OECD, 2021b,c,d). The curated LLNA EC3 values compiled by the DASS expert group were used, except in those instances where additional data, e.g., from the European Chemicals Agency (ECHA), provided a more robust value or if no EC3 was available in the DASS database. For human data, the values recently curated for the RCPL were given preference, for a larger number of chemicals the data curated by the OECD expert groups were used, and in a few instances data gaps were filled with data from Bil et al. (2017) (which are based on the ICCVAM data compilation) or from the Research Institute for Fragrance Materials (RIFM) database. Thus, the majority of *in vivo* data have previously been carefully curated.

All data normalizations and all regression analyses were performed as described in detail before (Natsch and Gerberick, 2022a). In the previous work, three key and most predictive models were developed: (i) EQ1 combines KS and kDPRA data, (ii) EQ4 combines h-CLAT and kDPRA data, and (iii) EQ5 combines all three data sources. All models also contain a volatility parameter for highly volatile chemicals. For chemicals lacking Cys-peptide reactivity, an additional model based on KS and h-CLAT data (EQ6) was derived. In the current analysis, these four regression equations were recalculated for the different data subsets using the LLNA EC3, the PV or the DSA04 as *in vivo* target.

The input parameters derived from the different *in vitro* tests are Log  $k_{\max \text{ norm}}$ , the logarithmic normalized rate constant for peptide depletion in the kDPRA; Log  $EC1.5_{\text{norm}}$ , the concentration to induce the luciferase gene in KS 1.5-fold; Log  $IC50_{\text{norm}}$ , the concentration to reduce cell viability by 50% in the KS; Log  $MIT_{\text{norm}}$ , the normalized lower value of the concentration to induce the cell surface markers CD54 by 200% or CD86 by 150% in the h-CLAT; Log  $CV75_{\text{norm}}$ , the concentration to reduce cell viability by 25% in the h-CLAT, and Log  $VP_{\text{norm}}$ , describing volatility of the most volatile chemicals. All concentrations are expressed in molarity, and all logarithmic values were linearly normalized to 0 for chemicals with no reactivity, cell activation or cytotoxicity as described before (Natsch and Gerberick, 2022a) and as reported in the previously published database. The normalized *in vitro* data are reiterated in ESM1<sup>2</sup>; for source data consult the original dataset.

#### *An extended dataset with potency values and DSA04 values*

The concept of PVs used in the RCPL is to (a) select chemicals for which a solid base of *in vivo* data are available, ideally human data next to LLNA data, to (b) collect LLNA EC3 data and quantitative human potency data in the form of DSA04 and no observed effect level (NOEL) values, and to (c) derive a potency value using a defined WoE workflow integrating both the human and LLNA data, whereas these workflows are based on a general preference for positive human sensitization data if available. The current RCPL contains a total of 33 chemicals. While this number of chemicals is deemed sufficient to compare different pre-

dictive approaches for their ability to predict and rank sensitizer potency of chemicals (which is the original goal of the RCPL), it is a relatively small dataset to (re)train any predictive model with multiple input parameters. Here we present an extended set of chemicals and apply the same principles used to derive the PV as were used for the RCPL.

For step (a), the list of chemicals from our prior publication (Natsch and Gerberick, 2022a) and the list of chemicals with LLNA, human, and *in vitro* data as compiled by the OECD expert group (OECD, 2021b,c,d) were used as a starting point. For each chemical it was assessed whether, next to the LLNA, relevant human data was available and, if not, whether there were GPMT data from the LLNA validation (Haneke et al., 2001), from the database compiled by ICCVAM (2008), or from the RIFM or ECHA database, the idea being that congruent LLNA and guinea pig data can contribute to the WoE for the sensitization status of a chemical. Chemicals were selected for inclusion in the assessment if they were (i) in the RCPL list; (ii) rated negative for human sensitization in the DASS database; (iii) had negative LLNA data supported by negative human HRIPT/HMT data from studies at a high test concentration, other convincing negative human data or negative guinea pig data; (iv) a DSA04 could be derived from positive human data; (v) they had a positive LLNA study supported by either positive human data (positive HRIPT, positive HMT, frequent positive reactions in the clinic) or positive guinea pig studies (in the ICCVAM, RIFM or ECHA database); or (vi) multiple positive LLNA studies. Chemicals that had only a single LLNA study in the absence of other data or with divergent results from guinea pig studies were excluded. With these criteria, chemicals with a less certain status of sensitization hazard were excluded and the dataset was enriched for chemicals with available human data. The rationale for inclusion in the list is given separately for each chemical in the database in the supporting information (ESM1<sup>2</sup>).

For step (b) – collection of human data – all individual human studies collected by the DASS expert group (OECD, 2021d) were used. In addition, the data collected from the ICCVAM database and summarized by Bil et al. (2017) were consulted, and a few data gaps could be filled with data from the RIFM database (e.g., in Na et al., 2021) not yet included in the DASS database (Tab. 1). For the data-rich chemicals in the RCPL list, preference was given to HRIPT studies, as these do not include irritation by SDS and often were conducted on a larger number of subjects. However, here all available positive studies in the DASS human database were used, and if multiple studies with positive cases of sensitization were available for one chemical, a DSA04 was derived for each study. The DSA04 was calculated from the DSA1<sup>+</sup> value in the DASS database, which reports the extrapolated concentration for 1 positive reaction, but which is not normalized to the number of panelists. The DSA04 is an extrapolation from the frequency of positive reactions at the tested concentration to derive the concentration at which 4% reaction frequency would be expected. In the absence of knowledge on the true dose-response

<sup>2</sup> doi:10.14573/altex.2302081s1



it is assumed that the incidence is doubled by doubling the test concentration, thus the DSA04 is:

$$DSA04 = DSA1^+ \times n \div 25,$$

with DSA1<sup>+</sup> as reported in the DASS database and n the number of study subjects in a given study.

In another expression, i.e., if DSA1<sup>+</sup> was not already given in the DASS database, the DSA04 was calculated as:

$$DSA04 = \text{Test dose } (\mu\text{g}/\text{cm}^2) \times \frac{4\%}{\text{Incidence } (\%)}$$

If multiple studies were available, the weighted geometric mean was calculated from the logarithmic weighted geometric mean, giving studies with fewer panelists a lesser weight:

$$\text{Log (Weighted Geomean DSA04)} = \sum_{i=1}^j \frac{n_i}{n} \times \text{Log DSA04}_i,$$

whereby j is the number of studies, n<sub>i</sub> the number of subjects in a given study and n the total number of subjects in all studies considered.

This approach of weighing all the studies available in the DASS database led overall to similar DSA04 values as those derived by the RCPL authors (Fig. 1a) and those calculated based on the collection of DSA05 values in Bil et al. (2017) (Fig. 1b). Of course, these different data collections are largely based on the same primary studies, but the good correlation indicates that pooling the detailed studies collected by the DASS expert team using weighted geometric means overall confirms other assessments and is a valid way for deriving DSA04 values by integrating all collected evidence.

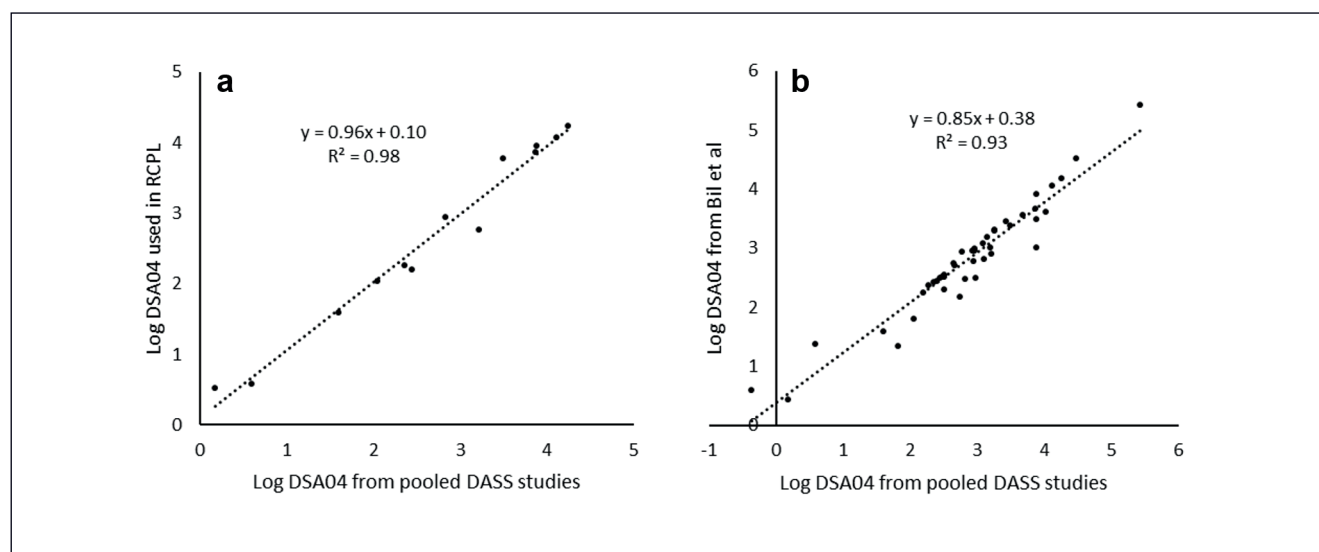
For step (c), i.e., deriving the PV, the workflows from Irizar et al. (2022) were applied. These workflows give preference to human data in case the LLNA value is not close to the human value. In case a PV was available from the RCPL list, this value was chosen by default, as these values had been carefully curated. To compare LLNA and DSA04 values, LLNA EC3 values in percent are multiplied with 250 to be expressed in  $\mu\text{g}/\text{cm}^2$  (Griem et al., 2003). The PV is thus expressed in  $\mu\text{g}/\text{cm}^2$  and may come from either the transformed LLNA EC3, the DSA04 or from the human NOEL. The rationale and source data for the selected PV is given for each chemical in the database in the supporting information (ESM1<sup>2</sup>).

In total, comprehensive *in vitro* data are available for 31 chemicals with existing PV values from Irizar et al. (2022) (no data are available for furaneol and 5-chloro-2-methyl-4-isothiazolinone, which had only been tested in all *in vitro* tests as the mixture Kathon CG). In addition, for 108 chemicals with comprehensive *in vitro* data a new PV was derived. Table 2 summarizes the database used for the current analysis and from which data source the PV were derived. For non-sensitizers the default PV is set to 25,000, i.e., an LLNA EC3 value of 100%. Based on the RCPL workflows, expert judgment is needed to derive the PV in some specific cases, but in most cases the workflows led to unambiguous decisions.

To apply the same regression analysis as done previously using pEC3 values, a pPV value was calculated similar to the pEC3 which is then in the same scale and can be directly compared:

$$pPV = \text{Log} \frac{MW}{PV \div 250}$$

For non-sensitizers, the pPV is set to 0, i.e., the value for a chemical with MW of 100 and a PV of 25,000. Similarly, a pDSA04 val-



**Fig. 1:** DSA04 values derived from the DASS database and using the weighted geometric mean in cases where multiple studies were available as compared to the DSA04 values used in (a) the RCPL list or (b) DSA04 values calculated from the ICCVAM DSA05 values reported in Bil et al. (2017)

**Tab. 1: Summary of different datasets with *in vivo* potency information and available *in vitro* data in the three OECD tests**

	Chemicals (n)	Sensitizers (n)	very weak / non-sensitizers (n)	Availability of <i>in vitro</i> data <sup>b</sup>	Comments
DASS LLNA EC3 database <sup>a</sup>	154	116	38	116 S and 33 NS <sup>c</sup>	88 S and 16 NS included in extended PV list derived here; 29 are also in the RCPL
DASS human potency database	63	52	11	50 S and 11 NS	45 S and 11 NS included in extended PV list; 24 are also in the RCPL
RCPL	33	28	5	26 S and 5 NS	26 S and 5 NS included in extended PV list; these 31 chemicals were used for the preliminary analysis in Tab. 3
Dataset Natsch and Gerberick, 2022a	188	135	53	135 S and 53 NS	This list was used to train published models including h-CLAT data (EQ 4 and 5); additional 15 chemicals in list used to train EQ 1 (n = 203)
<b>Extended PV list (this work; see Tab. 2)</b>	<b>139</b>	<b>101</b>	<b>38</b>	<b>101 S and 38 NS</b>	<b>This list was used for the key analysis in Tab. 4, 108 additional chemicals with PV and <i>in vitro</i> data as compared to RCPL with <i>in vitro</i> data (n = 31)</b>
Human DSA04 dataset (this work)	62	51	11 <sup>d</sup>	51 S and 11 NS	This list was used for analysis in Tab. 6; 46 S and 11 NS are from the DASS human potency database; 87-85-5 and 107-15-3 are from Bill et al. (2017); 1205-17-0, 18127-01-0, 24720-09-0 from RIFM DB; no potency is given in DASS DB for 119-84-6 but source data given to derive DSA04

<sup>a</sup> 16 additional chemicals are positive in the LLNA but lack EC3 values in the DASS DB; <sup>b</sup> Chemicals with kDPRA, KS, and h-CLAT data were considered for inclusion in the WoE list; <sup>c</sup> S, sensitizers; NS, non-sensitizers; <sup>d</sup> Three of the human sensitizers have DSA04 values > 25,000, hence their potency is set equal to the non-sensitizers; here they are counted with the sensitizers.

ue was calculated to directly use DSA04 values for the analysis.

$$pDSA04 = \text{Log} \frac{MW}{DSA04 \div 250}$$

An overview of the different datasets relevant for potency assessments put together by the OECD DASS in our previous analysis and for this work is given in Table 1, while Table 2 gives more details on the database used in the current analysis.

### 3 Results

In the analysis below, predictive regression equations were calculated in the form of, e.g., EQ5 from the previous publication given here as an example:

$$\text{EQ5: } pEC3 = \mathbf{0.20} \text{ (constant)} + \mathbf{0.34} \times \text{Log } k_{\max \text{ norm}} + \mathbf{0.09} \times \text{Log } EC1.5_{\text{norm}} + \mathbf{0.11} \times \text{Log } IC50_{\text{norm}} + \mathbf{0.20} \times \text{Log } MIT_{\text{norm}} + \mathbf{0.21} \times \text{Log } CV75_{\text{norm}} - \mathbf{0.19} \times \text{Log } VP_{\text{norm}}$$

For ease of reading, the new regression equations in the text are not explicitly spelled out as above but rather the regression coefficients (shown in bold, underlined text) defining these equations are given in tabular form. The analysis is performed for different combinations of input parameters:

- KS and kDPRA (EQ1 in the previous publication)
  - KS and h-CLAT (EQ4 in the previous publication)
  - KS, h-CLAT and kDPRA (EQ5 in the previous publication)
- In addition, the results for the different versions of EQ6 integrating KS and h-CLAT data are shown in the supporting information (ESM2<sup>3</sup>).

<sup>3</sup> doi:10.14573/altex.2302081s2




**Tab. 2: Summary of the extended list of chemicals with PV values developed for this work**

Contained in	WoE sensitizer status	PV value is derived from	n
RCPL	S	DSA04	11
RCPL	S	EC3 (close to DSA04) <sup>a</sup>	3
RCPL	S	EC3	10
RCPL	S	human NOEL	3
Extended list	S	DSA04	18
Extended list	S	EC3 (close to DSA04) <sup>a</sup>	9
Extended list	S	EC3	42
Extended list	S	EC3 (lower than DSA04) <sup>b</sup>	3
Extended list	S	human NOEL <sup>c</sup>	3
RCPL	NS	LLNA and congruent human/GP data	3
RCPL	NS	human NS	2
Extended list	NS	LLNA and congruent human/GP data	28
Extended list	NS	human NS	4
<b>Total</b>			<b>139</b>

<sup>a</sup> If EC 3 is less than 2-fold different from DSA04, EC3 is taken as PV value in RCPL workflow. <sup>b</sup> For three chemicals the DSA04 may underestimate the human sensitization potential, and therefore according to workflow #2 the lower EC3 is used as PV (phenyl benzoate, tetramethylthiuram disulphide, 2-mercaptobenzothiazol). <sup>c</sup> For 3 chemicals the selected PV is the human NOEL ( $\alpha$ -amylcinnamic aldehyde, resorcinol,  $\alpha$ -methylcinnamaldehyde; workflow #3)

The calculations were performed for different target data:

- LLNA pEC3
- Potency values expressed as pPV or
- human data expressed as pDSA04.

The coefficients of the regression equation from the previous work on predicting the pEC3 value on the comprehensive database ( $n = 188-203$ ) are given in the Tables first in each case for comparison. First the equation for the pEC3 target in each data subset is calculated and compared to the published equation calculated vs the pEC3. With this first comparison, an estimation can be made to which extent the specific smaller data subset affects the outcome when using the same target (namely pEC3) as in the previous analysis. This is important in order not to confuse changes observed when analyzing the different targets (pPV or pDSA04 vs pEC3) with changes due to the different set of chemicals.

Below, the same analysis is shown first for the published RCPL, then (and as the key analysis) for the extended PV list developed within this work, and finally on the list of chemicals with human sensitization data (DSA04 and human non-sensitizers).

The different equations are labeled with lowercase letters added to the numbers from the previous publications to differentiate equations on different combinations of input parameters, different targets, and different datasets.

### 3.1 Regression analysis vs the RCPL

The starting point of this investigation is based on the published list of 33 PV values, for 31 of which *in vitro* data were available. Table 3 indicates the results of the regression analysis on this

dataset. For all combinations of *in vitro* input parameters, models with similar or higher  $R^2$  values as for the published models were obtained. There are three important observations from this initial analysis:

First, comparing the equations for the pEC3 on these 31 chemicals with the published equations indicates a clearly different outcome: In all three combinations of input parameters (EQ1, EQ4 and EQ5) a lesser weight for the reactivity parameter  $k_{max}$  and a higher weight for both cytotoxicity parameters and the KS EC1.5 was found (see, e.g., EQ1a vs EQ1). This indicates that based on this small data subset, a clearly different predictive model for the LLNA would be derived, indicating that this small set of chemicals is not representative of the whole chemical dataset with LLNA and *in vitro* data available, and it is therefore not recommended to use the RCPL to train predictive models. This is not at all surprising given the difference in size ( $n = 31$  vs  $n = 188-203$ ).

Next, comparing the equations calculated for pEC3 and pPV for the 31 chemicals, an interesting observation is made in all combinations of input parameters (EQ1, EQ4 and EQ5): For predicting the pPV, the coefficients (and hence the contribution to sensitizing potency) of  $k_{max}$ , KS EC1.5, and/or h-CLAT MIT increase strongly in parallel to a large drop in the coefficients of both cytotoxicity parameters as compared to the situation when predicting the pEC3 (see, e.g., EQ1a vs EQ1b). This observation may indicate an important mechanistic difference when integrating the human data in the PV and actually triggered my curiosity to evaluate predictive models on a larger dataset of PV values as presented below.

**Tab.3: Regression coefficients and statistics for predictive models trained on the RCPL dataset**

Model	Constant	$k_{max}$	EC1.5	IC50	MIT	CV75	$VP_{norm}$	$R^2$	$n^d$	EQ <sup>e</sup>
<b>EQ1 (KS+kDPRA)<sup>a</sup></b>	0.42	0.40	0.15	0.36			-0.21	62%	203	EQ1
vs pEC3 <sup>b</sup>	0.30 <sup>c</sup>	<b>0.22</b>	<b>0.27</b>	<b>0.54</b>			0.07	63%	31	EQ1a
vs pPV	0.3	<b>0.47</b>	<b>0.43</b>	<b>0.08</b>			0.46	62%	31	EQ1b
<b>EQ4 (h-CLAT+kDPRA)</b>	0.18	0.36			0.21	0.35	-0.19	61%	188	EQ4
vs pEC3	-0.17	<b>0.24</b>			<b>0.18</b>	<b>0.78</b>	-0.30	70%	31	EQ4a
vs pPV	0.00	<b>0.47</b>			<b>0.40</b>	<b>0.25</b>	0.15	66%	31	EQ4b
<b>EQ5 (KS+h-CLAT+kDPRA)</b>	0.20	0.34	0.09	0.11	0.20	0.21	-0.19	62%	188	EQ5
vs pEC3	-0.39	<b>0.125</b>	0.30	<b>-0.1</b>	<b>0.18</b>	<b>0.70</b>	-0.26	75%	31	EQ5a
vs pPV	-0.43	<b>0.32</b>	0.44	<b>-0.4</b>	<b>0.39</b>	<b>0.35</b>	0.12	74%	31	EQ5b

<sup>a</sup> The published equations are shown in bold. <sup>b</sup> The same analysis vs pEC3 as for the published EQ1 but only for the 31 RCPL chemicals.

<sup>c</sup> Regression coefficients for the different parameters are given that fully define the regression equations as shown in the text for EQ5. New regression models were calculated for the specific dataset and different *in vivo* targets (PV or EC3 values). <sup>d</sup> Number of chemicals used to train the model. <sup>e</sup> Numbers and lowercase letters to differentiate the new equations. EQ1, EQ4 and EQ5 are the published equations.

Third, the highest  $R^2$  (74-75%) is found for the models integrating all data in EQ5. This is higher than in all published models and is true for both the pEC3 and the pPV. However, this may already partly be due to an over-fitted model. Thus, while the overall statistics are still highly significant, the coefficient is still statistically significant in EQ5b only for the KS EC1.5 and the MIT, and all the coefficients have a high standard deviation (data not shown).

### 3.2 Regression analysis vs the extended list of PV values

In Table 4, the same analysis as above is presented for the extended list ( $n = 139$ ) with PV values (see Section 2 and ESM1<sup>2</sup>). In all cases, highly significant regression models with  $R^2$  values similar to those of the published models were derived. There are two important findings from this analysis:

First, the regression models developed to predict the pEC3 based on these 139 chemicals are almost identical to the published models developed on  $n = 188-203$  chemicals (see, e.g., EQ1c vs EQ1). This indicates that this set of chemicals can be considered representative of the full database, and we would not expect a major database bias for models developed for this smaller dataset.

Secondly, and maybe most importantly, the outcome is clearly different when evaluating against pPV values as compared to the evaluation vs the pEC3 target. More specifically, in all three input parameter combinations (EQ1, EQ4 and EQ5), the weight of the reactivity parameter, the KS EC1.5, and/or the h-CLAT MIT is increasing in parallel to a large drop in the weight of the cytotoxicity coefficients (see, e.g., EQ1d vs EQ1c). This confirms the observation reported above on the much smaller RCPL set. Actually, the cytotoxicity parameters become statistically non-significant when evaluating vs the PV, but they are highly significant in EQ1 and EQ4 when evaluating vs the LLNA (Tab. 5). Also, the volatility parameter becomes non-significant. Thus, in order to predict the pPV, a simplified equation (EQ5e) based only on

KS EC1.5, h-CLAT MIT, and  $k_{max}$  and completely excluding cytotoxicity can be derived for the extended PV list which has an identical  $R^2 = 63\%$ :

$$\text{EQ5e: pEC3} = 0.21 + 0.38 \times \text{Log } k_{max \text{ norm}} + 0.31 \times \text{Log MIT}_{\text{norm}} + 0.21 \times \text{Log EC1.5}_{\text{norm}}$$

Constant	Standard error = 0.11, $p = 0.050$
Log $k_{max \text{ norm}}$	Standard error = 0.06, $p < 0.0005$
Log $EC1.5_{\text{norm}}$	Standard error = 0.07, $p = 0.002$
Log $MIT_{\text{norm}}$	Standard error = 0.07, $p < 0.0005$

$$n = 139; S = 0.74; R^2 = 63.4\%; R^2(\text{adj}) = 62.6\%$$

The same analysis on KS and h-CLAT data only (EQ6) is shown in ESM2<sup>3</sup> and led to completely congruent conclusions.

### 3.3 Regression analysis vs the list with DSA04 values

Having collected the DSA04 values along with the *in vitro* data raises an obvious question: Given that the models differ significantly whether we predict the EC3 or the PV – and given that the PV is still significantly influenced by the large number of EC3 values used to derive PV values next to the human data – would we arrive at more human-relevant models when using only the human DSA04 values and the information on the human non-sensitizers? Thus, the same analysis as above was repeated a third time, this time only including the subset of data with human DSA04 values or a negative human sensitization status ( $n = 62$ ). This analysis is shown in Table 6, and similar comparisons as above can be made, with three key observations:

First, comparing the models for predicting EC3 (e.g., EQ1f vs EQ1) with the published models indicates that we obtain significantly different LLNA models on this set of chemicals. Thus, on this dataset the reactivity parameter has almost no weight to predict the LLNA EC3 and the coefficient for the MIT is reduced


**Tab. 4: Regression coefficients and statistics for predictive models on the extended PV list**

Model	Constant	$k_{max}$	EC1.5	IC50	MIT	CV75	VP <sub>norm</sub>	R <sup>2</sup>	N	EQ
<b>EQ1 (KS+kDPRA)<sup>a</sup></b>	0.42	0.40	0.15	0.36			-0.21	62%	203	EQ1
vs pEC3	0.42	<b>0.39</b>	<b>0.16</b>	<b>0.39</b>			-0.16	61%	139	EQ1c
vs pPV	0.42	<b>0.45</b>	<b>0.30</b>	<b>0.08</b>			-0.08	59%	139	EQ1d
<b>EQ4 (h-CLAT+kDPRA)</b>	0.18	0.36			0.21	0.35	-0.19	61%	188	EQ4
vs pEC3	0.14	<b>0.36</b>			<b>0.18</b>	<b>0.43</b>	-0.15	65%	139	EQ4c
vs pPV	0.28	<b>0.45</b>			<b>0.36</b>	<b>0.07</b>	-0.04	61%	139	EQ4d
<b>EQ5 (KS+h-CLAT+kDPRA)</b>	0.20	0.34	0.09	0.11	0.20	0.21	-0.19	62%	188	EQ5
vs pEC3	0.15	<b>0.34</b>	<b>0.09</b>	<b>0.03</b>	<b>0.18</b>	<b>0.33</b>	-0.15	65%	139	EQ5c
vs pPV	0.24	0.38	<b>0.26</b>	<b>-0.17</b>	<b>0.36</b>	<b>0.01</b>	-0.06	64%	139	EQ5d

<sup>a</sup> See footnotes of Table 3.

**Tab. 5: *p*-values for the cytotoxicity parameters in the evaluation for the extended PV list**

Target	IC50 in EQ1	CV75 in EQ4	IC50 in EQ5	CV75 in EQ5
vs pEC3	< 0.0005	< 0.0005	0.83	<b>0.045</b>
vs pPV	0.57	0.57	0.24	0.95

too, while a higher weight for both cytotoxicity parameters and KS EC1.5 is found. This indicates that this chemical set, similar to the RCPL, is not representative of the full database. The observation that cytotoxicity becomes a dominant predictor for pEC3 and reactivity appears almost irrelevant is contrary to our knowledge on the AOP and further underlines the notion that this appears to be a biased data subset.

Next, when comparing the models for the pEC3 target vs the models for the pDSA04 target on these  $n = 62$  chemicals, a difference in the weight of the parameters is noted as in the above analysis comparing the targets pEC3 and pPV. Thus, the weight of the reactivity parameter – although remaining low – is increased, the coefficients for the EC1.5 and the MIT are strongly increased in all three cases (EQ1, EQ4 and EQ5), and again, the weight of the cytotoxicity parameters is strongly reduced and becomes insignificant or even negative in all cases (see, e.g., EQ1g vs EQ1f).

Third, we note overall low R<sup>2</sup> values (32–45%) when fitting models to pDSA04 values, while the R<sup>2</sup> is only marginally reduced with this dataset when evaluating against pEC3 (53–60%) as compared to the published models on the full dataset (R<sup>2</sup> = 61–62%). This might offer a rather disturbing truth, namely that the validated *in vitro* methods are rather poor predictors of the human sensitization potency as compared to their ability to predict the LLNA. However, this conclusion may be a shortcut: Thus for the same dataset also a correlation analysis of the pDSA04 against the LLNA pEC3 was made, indicating that actually the prediction of the human data by the LLNA (which is considered

the gold standard to predict human potency) only yields an R<sup>2</sup> = 43% in this curated dataset, and thus prediction of human potency by animal data is not better than the prediction of human potency by *in vitro* data:

$$pDSA04 = 0.4 + 0.73 \times pEC3$$

$$\begin{array}{ll} \text{Constant} & \text{Standard error} = 0.17, p = 0.017 \\ pEC3 & \text{Standard error} = 0.11, p < 0.0005 \end{array}$$

$$n = 62; S = 0.78; R^2 = 43\%; R^2(\text{adj}) = 42.0$$

### 3.4 Predictive differences of the PV-derived models as compared to the EC3-derived models

Among the three sets of chemicals, the analysis on the extended PV list is of highest relevance: This dataset appears representative of the larger dataset analyzed earlier. At the same time, this dataset leads to the same conclusion as the analysis on the other subsets, namely that integrating human data yields different predictive equations, with less weight for cytotoxicity parameters and higher weight for reactivity and cell activation in both KS and h-CLAT. Thus, the key question is how the prediction will be changed for individual chemicals by, e.g., applying EQ5e based on kDPRA, KS and h-CLAT and trained on PV values instead of the published EQ5. To illustrate the difference, the predicted potency using these two models for all 139 chemicals with PV values is plotted in Figure 2.

Even though the new regression equation EQ5e, trained on the PV values and hence excluding the cytotoxicity parameters, is significantly different from the published EQ5, the predictions with these two equations correlate well (R<sup>2</sup> = 0.94; slope = 1.02 and y-intercept = 0.04, indicating the regression line is very close to the line of identity). This indicates that this additional analysis integrating the human evidence in the WoE PV values leads to a surprisingly small change in the prediction of skin sensitizer potency for individual chemicals. Still, on a more granular scale, interesting differences can be noted: (i) for the chemicals for



**Tab. 6: Regression coefficients and statistics for predictive models on the list with human data**

Model	Constant	k <sub>max</sub>	EC1.5	IC50	MIT	CV75	VP <sub>norm</sub>	R <sup>2</sup>	N	EQ
<b>EQ1 (KS+kDPRA)<sup>a</sup></b>	0.42	<b>0.40</b>	0.15	<b>0.36</b>			-0.21	62%	203	EQ1
vs pEC3	0.27	<b>0.00</b>	0.30	<b>0.59</b>			-0.04	57%	62	EQ1f
vs pDSA04	0.47	0.09	0.48	<b>0.07</b>			-0.02	40%	62	EQ1g
<b>EQ4 (h-CLAT+kDPRA)</b>	0.18	<b>0.36</b>			0.21	0.35	-0.19	61%	188	EQ4
vs pEC3	0.16	<b>0.12</b>			0.11	<b>0.59</b>	-0.10	53%	62	EQ4f
vs pDSA04	0.55	0.29			0.42	<b>-0.12</b>	-0.01	32%	62	EQ4g
<b>EQ5 (KS+h-CLAT+kDPRA)</b>	0.20	<b>0.34</b>	0.09	0.11	0.20	0.21	-0.19	62%	188	EQ5
vs pEC3	0.07	<b>-0.02</b>	0.26	<b>0.26</b>	0.06	<b>0.35</b>	-0.06	60%	62	EQ5g
vs pDSA04	0.34	0.04	0.44	<b>-0.06</b>	0.35	-0.15	0.01	45%	62	EQ5h

<sup>a</sup> See footnotes of Table 3.

**Tab. 7: Chemicals whose potency is predicted more than 3.3-fold higher or lower by the PV-derived model EQ5e as compared to the published LLNA-based model EQ5**

Name	EC3 (µg/cm <sup>2</sup> )	DSA04 (µg/cm <sup>2</sup> )	PV (µg/cm <sup>2</sup> )	Predicted EC3 EQ5 (µg/cm <sup>2</sup> )	Predicted PV EQ5e (µg/cm <sup>2</sup> )
Sodium dodecyl sulfate	925	NC (6,207)	25,000	10,391	> 25,000 (44,000)
Ethyl acrylate	8,188	1,231	1,231	641	120
5-Methyl-2,3-hexanedione	6,500	2,644	2,644	5,405	1,603
Ethylenediamine	550	586	550	9,540	2,448
Glyoxal	350	249	350	4,109	621
Formaldehyde	950	553	553	247	32
3-Dimethyl-amino-1-propylamine	713	no data	713	8,010	2,389

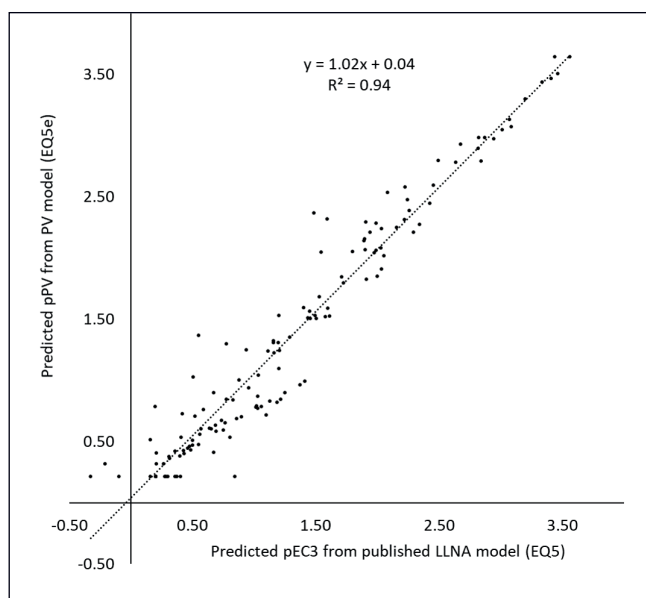
which a strong sensitization potency is predicted (pPV or pEC3 is > 2.3 corresponding to an EC3 of around < 1%), the predictions are almost equivalent, and all data points are close to the line of identity; (ii) for moderate sensitizers (pPV / pEC3 ≈ 1.3-2.3 corresponding to an EC3 of ≈ 1-10%), there is a tendency that more points are above the regression line, indicating that the new equation EQ5e is slightly more conservative for moderate sensitizers; (iii) the highest scattering on both sides of the regression line is observed for the weak sensitizers and non-sensitizers (pPV / pEC3 < 1.3), with a tendency that more points are below the regression line, indicating that EQ5, which includes the cytotoxicity parameters, is more conservative and predicts a higher sensitization potency for these weak and non-sensitizing chemicals. The actual predictions by both EQ5 and EQ5e for all individual chemicals along with the PV and EC3 values is given in the supporting information, and here only the cases with the highest discrepancy (> 3.3-fold difference, i.e., > 0.5 Log-unit) between

the two assessments are summarized in Table 7. This list includes only six chemicals. The omission of the cytotoxicity parameters in EQ5e clearly renders the LLNA false-positive surfactant SDS a predicted non-sensitizer. On the other hand, two weak and four moderate sensitizers in the LLNA are predicted as more potent by the PV model. This is better aligned with the *in vivo* data in four cases but leads to over-prediction in the case of formaldehyde and ethyl acrylate.

### 3.5 Application in practice

For a straightforward prediction of the sensitization potency, our previous publication provided a calculation spreadsheet to directly derive a PoD from the quantitative *in vitro* data. Here an updated prediction spreadsheet is provided in ESM3<sup>4</sup>, including the additional regression equations derived from the extended PV list. Users will thus have multiple PoD calculated based on LLNA or PV trained models and based on the different input pa-

<sup>4</sup> doi:10.14573/altex.2302081s3



**Fig. 2:** Predicted skin sensitizer potency based on kDPRA, KS, and h-CLAT with the published EQ5 trained on LLNA data vs the predictions with EQ5e trained on the PV values and hence excluding the cytotoxicity parameters

rameters. How to best select the most relevant result (EQ1, EQ4 EQ5 or EQ6) depending on (partial) data availability has been discussed in detail in our previous publications (Natsch and Gerberick, 2022a,b). The additional, PV-derived PoD values may be used in a similar way. In a conservative assessment, users may select the lower PoD from either the LLNA- or the PV-based models, unless there is a mechanistic reason to prefer one over the other. Thus, e.g., in the case of a strongly cytotoxic molecule with only a weak sensitization alert and weak experimental reactivity, preference may be given to the models trained on the PV as this may avoid a bias introduced by cytotoxicity, since cytotoxicity may rather reflect the irritant potential of that molecule. *Vice versa*, in case of strong activation of cellular markers significantly below the cytotoxic level, and especially in the presence of structural alerts for reactivity and/or experimental reactivity, the more conservative PoD derived from the PV models may be preferred. However, as shown above, the difference in most cases will be rather small. The prediction spreadsheet also provides two case studies on 2,4-dinitro-chlorobenzene (DNCB) and cinnamic aldehyde, and the results for these two chemicals are shown in Table 8. For these data-rich molecules with congruent *in vitro* results (kDPRA, KS and h-CLAT are all positive), the PoD values derived by all published models but also by the new PV-derived models are very close, indicating that the assessment based on different subsets of input parameters and based on different targets (LLNA or PV) is highly robust. All the models under-predict the PV of DNCB, which is based on a strong sensitization potency observed in humans, while they are more aligned with the LLNA potency of cinnamic aldehyde, which has a relatively high PV based on human predictive tests.

**Tab. 8:** Case studies DNCB and cinnamic aldehyde predicted by multiple published models and new models trained on the PV values

		DNCB	Cinnamic aldehyde
<b>EC3</b> (DSA in $\mu\text{g}/\text{cm}^2$ )	LLNA EC3	13	250
	EQ1	45	254
	EQ4	48	201
	<b>EQ5<sup>a</sup></b>	<b>41</b>	<b>199</b>
	EQ6	65	278
<b>PV</b> (DSA in $\mu\text{g}/\text{cm}^2$ )	PV value	3	885
	EQ1d	57	196
	EQ4d	45	168
	EQ5d	52	140
	EQ6d	83	195
	<b>EQ5e<sup>a</sup></b>	<b>43</b>	<b>160</b>

<sup>a</sup> EQ5 is highlighted based on evidence from all three tests.

#### 4 Discussion

The goal of NAM-based *in vitro* testing for skin sensitization is to provide the necessary tools for risk assessment, i.e., to be able to derive safe use levels for chemicals in the absence of animal data. Such a risk assessment needs to be human-relevant, and one may therefore start with the naïve assumption that the assessment ultimately should be trained on human data only. However, as indicated in Section 1, there are pros and cons of using human data in terms of data availability, data quality, and availability of dose-response information. Thus, the present analysis is focused on an empirical and statistical assessment of the use of human data to determine the best way to integrate them into the assessment, and it builds on the work on the RCPL list (Irizar et al., 2022).

The data analysis follows our previous work on skin sensitization potency assessment exclusively based on OECD-accepted *in vitro* assays and focused on the LLNA as target (Natsch and Gerberick, 2022a,b), which itself was a continuation of a previous analysis using a different peptide reactivity assay (Natsch et al., 2018).

There are four key learnings from this in-depth analysis on different *in vivo* targets and different chemical sets which will be discussed separately:

- (i) Adding human data in the form of PV to an extended dataset appears a better option than training models exclusively on the available human data due to bias and limitations in the human dataset.
- (ii) Integrating human data into the assessment leads to models with a lesser weighting of cytotoxicity, which – based on a

purely statistical analysis – points to important mechanistic differences in human and LLNA data.

- (iii) While the human data affect the weight of the input parameters, there is a surprisingly small effect on the individual predictions, indicating a high robustness of the regression models for application in practice.
- (iv) The extended RCPL list may be used to train models integrating human data in a more comprehensive database.

#### *(i) Human data as target*

The analysis of the dataset ( $n = 62$ ) with human data only showed that this set of chemicals is biased. Thus, when ignoring the human data and just predicting the pEC<sub>3</sub> with models trained on this subset, the models deviate strongly from the models on the extended PV list ( $n = 139$ ) and from the published models (Natsch and Gerberick, 2022a) ( $n = 188$ -203) and also on the larger dataset included in supporting information (ESM4<sup>5</sup>) to the previous analysis (Natsch and Gerberick, 2022a) ( $n = 317$ , EQ22). Statistical analysis of all those larger datasets points to a very consistent and strong contribution of the Cys-peptide reaction rate to the potency next to the contributions from KS and/or h-CLAT, in perfect alignment with the sensitization AOP (OECD, 2014). All these larger datasets led to almost identical regression coefficients in an analysis vs the pEC<sub>3</sub>. However, the contribution of reactivity is lost when predicting the pEC<sub>3</sub> for the human dataset only (see EQ1f and EQ5g in Tab. 6). This shows that – completely independent of the human data and the human data quality – the collection of chemicals for which we can make this human assessment is a biased set and not a representative subset of the database. This finding may not only be relevant to the modeling approach by the regression models, but also to all other attempts to derive human-relevant potency models relying on this human dataset.

Next to this dataset bias (which is independent of the quality of the human data), we also may have a certain reservation about the quality of the human data. The rather poor prediction (as expressed by low  $R^2$ ) of the human sensitization potency by both the LLNA and the *in vitro* data may indicate that the quantitative human data may not be ideal in terms of data quality, which may be related to the limited dose-response information and the historical nature of most of the test results coming from not fully standardized protocols. This further indicates that these data should be used in a WoE and not as sole predictive target. Thus, using these data in the form of a WoE PV as proposed by the RCPL workflows (Irizar et al., 2022) appears a balanced way forward for integration of the human evidence.

#### *(ii) Importance of cytotoxicity parameters*

From a mechanistic angle, the most interesting finding is that moving from an LLNA-only based assessment to the PV or the DSA04 as a target leads to an almost complete loss of the influence of the cytotoxicity parameters. The same effect is observed on all three sets of chemicals (Tab. 3, 4 and 6) and on all four combinations of input parameters (including EQ6, see ESM2<sup>3</sup>),

and it is thus highly consistent. The resulting modification of the regression model in EQ5e with the omission of the cytotoxicity parameters has the strongest impact on the prediction of the weak/non-sensitizers, and it points to an important mechanistic difference between LLNA and human potency data, with cytotoxicity / irritation being a confounder for the weak / non-sensitizers in the LLNA. Interestingly, in a Bayesian statistics approach for predicting LLNA potency classes, the mutual information of the cytotoxicity parameters was strongest for the prediction of non-sensitizers followed by weak sensitizers, whereas they had the least mutual information with the strong sensitizers (Jaworska et al., 2015), which is in line with our observations. Skin irritation as a confounding factor in the LLNA had been discussed in detail when assessing the LLNA for hazard assessment (Ball et al., 2011; Garcia et al., 2010; Kreiling et al., 2008; Natsch et al., 2023), and the relationship between irritation and cytotoxicity is well established (Muller-Decker et al., 1994). The statistical analysis provided here independently again indicates an important effect of cytotoxicity to predict a weak sensitization potency in the LLNA, which may not perfectly mimic the human situation in all cases. Cytotoxicity and resulting irritation may be a key factor causing false-positives in the LLNA, especially for chemicals only positive at high concentrations, and this was recently discussed in detail (Natsch et al., 2023). On the other hand, a positive LLNA result cannot be questioned based only on cytotoxicity as also true skin sensitizers are often irritants. Identifying LLNA false positives thus requires a WoE assessment including cytotoxicity, lack of structural alert, and lack of human sensitization data on the target chemical or related analogues.

#### *(iii) Effect of integrating human data on predicted potency*

Maybe the most important question is the practical one – Does the integration of human data significantly change the risk assessment we would make based on LLNA-trained models such as the ones we provided before (Natsch and Gerberick, 2022a,b)? This question may also relate to any existing or future model solely based on LLNA potency data. The comparison of the predictions obtained from the different equations EQ5 and EQ5e indicates: The effect is surprisingly small, and overall, predictions are quite similar. This indicates that the models derived from LLNA data only are quite robust, but we could now provide additional models in the calculation spreadsheet (ESM3<sup>4</sup>) which can be used in a WoE approach, taking into account the learning that for some cytotoxic compounds (which may be over-predicted by the LLNA models) or for compounds strongly activating the cellular markers in the absence of cytotoxicity (which may be under-predicted by the LLNA model) it may make sense to use this additional information in a WoE and select the PoD accordingly.

#### *(iv) The extended PV list as a tool for modeling*

Finally, this work provides an extended list of chemicals with PV assessed according to the general RCPL principles. This list is not meant to replace the established lists for an assessment of predic-

<sup>5</sup> doi:10.14573/altex.2201141s4



tive capacity for hazard assessment such as the DASS database but rather for modeling potency. It may be further refined, but it could be a useful tool going forward since training potency models exclusively against human data appears no ideal solution (see point (i) above). On the other hand, it is not recommended to train models on the narrow set of chemicals of the RCPL: Thus, EQ5b trained on the RCPL has the highest predictivity for the RCPL ( $R^2 = 74\%$ ; Tab. 3), while published EQ5 has an  $R^2 = 67\%$  for this set of chemicals (data not shown). However, EQ5b trained on the PV list has an  $R^2$  of 61% on the entire dataset (data not shown) and performs worse than EQ5d trained on the full dataset ( $R^2 = 64\%$ ; Tab. 4), indicating that training on the RCPL will not lead to the best model to represent the complete dataset. This is because the chemical set of the RCPL is not representative of the full dataset.

## 5 Outlook

Based on the shortcomings of the human dataset, a key question remains: Should further models for potency assessment always include human data? Based on the current analysis, there are important mechanistic learnings from the human data in terms of weighing irritation and cytotoxicity, yet the human data do not lead to a potency model with a substantially different final outcome – which is reassuring for risk assessors. Future attempts may thus apply the extended list of PV to test the robustness of models trained entirely on LLNA data in order to not miss mechanistic aspects, yet the limitations of the human data highlighted here also indicate that the human database should not be overemphasized when predicting sensitizer potency.

## References

- 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
- Api, A. M., Basketter, D. and Lalko, J. (2015). Correlation between experimental human and murine skin sensitization induction thresholds. *Cutan Ocul Toxicol* 34, 298-302. doi:10.3109/15569527.2014.979425
- 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
- Bil, W., Schuur, A. G., Ezendam, J. et al. (2017). Probabilistic derivation of the interspecies assessment factor for skin sensitization. *Regul Toxicol Pharmacol* 88, 34-44. doi:10.1016/j.yrtph.2017.05.015
- Dean, J. H., Twerdok, L. E., Tice, R. R. et al. (2001). ICCVAM evaluation of the murine local lymph node assay. Conclusions and recommendations of an independent scientific peer review panel. *Regul Toxicol Pharmacol* 34, 258-273. doi:10.1006/rtp.2001.1497
- Emter, R., Ellis, G. and Natsch, A. (2010). Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. *Toxicol Appl Pharmacol* 245, 281-290. doi:10.1016/j.taap.2010.03.009
- Garcia, C., Ball, N., Cagen, S. et al. (2010). Comparative testing for the identification of skin-sensitizing potentials of nonionic sugar lipid surfactants. *Regul Toxicol Pharmacol* 58, 301-307. doi:10.1016/j.yrtph.2010.06.016
- 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
- Gilmour, N., Reynolds, J., Przybylak, K. et al. (2022). Next generation risk assessment for skin allergy: Decision making using new approach methodologies. *Regul Toxicol Pharmacol* 131, 105159. doi:10.1016/j.yrtph.2022.105159
- Griem, P., Goebel, C. and Scheffler, H. (2003). Proposal for a risk assessment methodology for skin sensitization based on sensitization potency data. *Regul Toxicol Pharmacol* 38, 269-290. doi:10.1016/j.yrtph.2003.07.001
- Haneke, K. E., Tice, R. R., Carson, B. L. et al. (2001). ICCVAM evaluation of the murine local lymph node assay. Data analyses completed by the national toxicology program interagency center for the evaluation of alternative toxicological methods. *Regul Toxicol Pharmacol* 34, 274-286. doi:10.1006/rtp.2001.1498
- ICCVAM (2008). ICCVAM Database. [https://ntp.niehs.nih.gov/sites/default/files/iccvam/docs/immunotox\\_docs/llna-pot/tmer.pdf](https://ntp.niehs.nih.gov/sites/default/files/iccvam/docs/immunotox_docs/llna-pot/tmer.pdf) (accessed December 2008)
- 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 Number 11-7709*. <https://bit.ly/3HQ9hGN> (accessed 23.01.2014)
- Irizar, A., Bender, H., Griem, P. et al. (2022). Reference chemical potency list (RCPL): A new tool for evaluating the accuracy of skin sensitization potency measurements by new approach methodologies (NAMs). *Regul Toxicol Pharmacol* 134, 105244. doi:10.1016/j.yrtph.2022.105244
- 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
- Kligman, A. M. (1966). The identification of contact allergens by human assay. 3. The maximization test: A procedure for screening and rating contact sensitizers. *J Invest Dermatol* 47, 393-409. doi:10.1038/jid.1966.160
- Kreiling, R., Hollnagel, H. M., Hareng, L. et al. (2008). Comparison of the skin sensitizing potential of unsaturated compounds as assessed by the murine local lymph node assay (LLNA) and the guinea pig maximization test (GPMT). *Food Chem Toxicol* 46, 1896-1904. doi:10.1016/j.fct.2008.01.019
- Muller-Decker, K., Furstenberger, G. and Marks, F. (1994). Keratinocyte-derived proinflammatory key mediators and cell viability as in vitro parameters of irritancy: A possible alternative to the



- Draize skin irritation test. *Toxicol Appl Pharmacol* 127, 99-108. doi:10.1006/taap.1994.1144
- Na, M., Ritacco, G., O'Brien, D. et al. (2021). Fragrance skin sensitization evaluation and human testing: 30-year experience. *Dermatitis* 32, 339-352. doi:10.1097/DER.0000000000000684
- 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
- Natsch, A., Haupt, T., Wareing, B. et al. (2020). Predictivity of the kinetic direct peptide reactivity assay (kDPRA) for sensitizer potency assessment and GHS subclassification. *ALTEX* 37, 652-664. doi:10.14573/altex.2004292
- Natsch, A. and Gerberick, G. F. (2022a). Integrated skin sensitization assessment based on OECD methods (I): Deriving a point of departure for risk assessment. *ALTEX* 39, 636-646. doi:10.14573/altex.2201141
- Natsch, A. and Gerberick, G. F. (2022b). Integrated skin sensitization assessment based on OECD methods (II): Hazard and potency by combining kinetic peptide reactivity and the "2 out of 3" defined approach. *ALTEX* 39, 647-655. doi:10.14573/altex.2201142
- Natsch, A., Kleinstreuer, N. and Asturiol, D. (2023). Reduced specificity for the local lymph node assay for lipophilic chemicals: Implications for the validation of new approach methods for skin sensitization. *Regul Toxicol Pharmacol* 138, 105333. doi:10.1016/j.yrtph.2023.105333
- OECD (2014). 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: ARE-Nrf2 Luciferase Test Method. *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 (2021a). Guideline No. 497: Defined Approaches on Skin Sensitisation. *OECD Guidelines for the Testing of Chemicals, Section 4*. OECD Publishing, Paris. doi:10.1787/b92879a4-en
- OECD (2021b). Supporting Document to the Guideline (GL) on Defined Approaches (DAs) for Skin Sensitisation – Annex 2. *Series on Testing and Assessment, No. 336*. OECD Publishing, Paris. <https://www.oecd.org/chemicalsafety/testing/Annex-2-in-vitro-in-silico-in-vivo-defined-approaches.xlsx>
- OECD (2021c). Supporting Document to the Guideline (GL) on Defined Approaches (DAs) for Skin Sensitisation – Annex 3. *Series on Testing and Assessment, No. 336*. OECD Publishing, Paris. [https://one.oecd.org/document/ENV/CBC/MONO\(2021\)11/ann3/en/pdf](https://one.oecd.org/document/ENV/CBC/MONO(2021)11/ann3/en/pdf)
- OECD (2021d). Supporting Document to the Guideline (GL) on Defined Approaches (DAs) for Skin Sensitisation – Annex 4. *Series on Testing and Assessment, No. 336*. OECD Publishing, Paris. [https://one.oecd.org/document/ENV/CBC/MONO\(2021\)11/ann4/en/pdf](https://one.oecd.org/document/ENV/CBC/MONO(2021)11/ann4/en/pdf)
- OECD (2023). 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
- Politano, V. T. and Api, A. M. (2008). The research institute for fragrance materials' human repeated insult patch test protocol. *Regul Toxicol Pharmacol* 52, 35-38. doi:10.1016/j.yrtph.2007.11.004
- Reynolds, J., Gilmour, N., Baltazar, M. T. et al. (2022). Decision making in next generation risk assessment for skin allergy: Using historical clinical experience to benchmark risk. *Regul Toxicol Pharmacol* 134, 105219. doi:10.1016/j.yrtph.2022.105219
- Sakaguchi, H., Ashikaga, T., Miyazawa, M. et al. (2006). Development of an in vitro skin sensitization test using human cell lines; human cell line activation test (h-CLAT) II. An inter-laboratory study of the h-CLAT. *Toxicol In Vitro* 20, 774-784. doi:10.1016/j.tiv.2005.10.014
- Takenouchi, O., Fukui, S., Okamoto, K. et al. (2015). Test battery with the human cell line activation test, direct peptide reactivity assay and DEREK based on a 139 chemical data set for predicting skin sensitizing potential and potency of chemicals. *J Appl Toxicol* 35, 1318-1332. doi:10.1002/jat.3127
- Wareing, B., Kollé, 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

#### Conflict of interest

The author declares no competing interests.

#### Data availability

All data of this publication are publicly available, and all models and tests used are freely available.

#### Electronic supplementary material

ESM1<sup>2</sup> gives the database on the extended PV list with LLNA data, human data, rationale why chemical was included, rationale for the PV selected and the PV, additional background *in vivo* data considered, and finally, the key *in vitro* data used. It further gives predicted potency by EQ5 and EQ5e for all chemicals. ESM2<sup>3</sup> shows the analysis integrating KS and h-CLAT data, without kDPRA. ESM3<sup>4</sup> is an updated prediction spreadsheet for practical application to predict potency with both the LLNA models and the PV-derived models.

#### Acknowledgments

I am grateful for the useful discussions with and input from Frank Gerberick and Amaia Irizar.