

A quantitative *in silico* model for predicting skin sensitization using a nearest neighbours approach within expert-derived structure–activity alert spaces

Steven J. Canipa^a, Martyn L. Chilton^a, Rachel Hemingway^a,
Donna S. Macmillan^{a*} , Alun Myden^a, Jeffrey P. Plante^a,
Rachael E. Tennant^a, Jonathan D. Vessey^a, Thomas Steger-Hartmann^b,
Janet Gould^c, Jedd Hillegass^c, Sylvain Etter^d, Benjamin P. C. Smith^e,
Angela White^f, Paul Sterchele^g, Ann De Smedt^h, Devin O'Brienⁱ
and Rahul Parakhiaⁱ

ABSTRACT: Dermal contact with chemicals may lead to an inflammatory reaction known as allergic contact dermatitis. Consequently, it is important to assess new and existing chemicals for their skin sensitizing potential and to mitigate exposure accordingly. There is an urgent need to develop quantitative non-animal methods to better predict the potency of potential sensitizers, driven largely by European Union (EU) Regulation 1223/2009, which forbids the use of animal tests for cosmetic ingredients sold in the EU. A Nearest Neighbours *in silico* model was developed using an in-house dataset of 1096 murine local lymph node (LLNA) studies. The EC3 value (the effective concentration of the test substance producing a threefold increase in the stimulation index compared to controls) of a given chemical was predicted using the weighted average of EC3 values of up to 10 most similar compounds within the same mechanistic space (as defined by activating the same Derek skin sensitization alert). The model was validated using previously unseen internal ($n = 45$) and external ($n = 103$) data and accuracy of predictions assessed using a threefold error, fivefold error, European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) and Globally Harmonized System of Classification and Labelling of Chemicals (GHS) classifications. In particular, the model predicts the GHS skin sensitization category of compounds well, predicting 64% of chemicals in an external test set within the correct category. Of the remaining chemicals in the previously unseen dataset, 25% were over-predicted (GHS 1A predicted: GHS 1B experimentally) and 11% were under-predicted (GHS 1B predicted: GHS 1A experimentally). Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: skin sensitization; Derek Nexus; local lymph node assay; LLNA; QSAR

Introduction

Skin sensitization

Contact dermatitis is a common inflammatory skin disease in industrialized countries with a significant socioeconomic impact (Diepgen and Kanerva, 2006). It is, therefore, important to ensure new and existing chemicals are appropriately assessed for their potential to cause skin sensitization and to mitigate exposure accordingly. To evaluate the sensitising hazard presented by a chemical, several animal tests have been developed and two tests, in particular, have gained widespread acceptance and have been validated by the Organization for Economic Co-operation and Development (OECD): the Guinea Pig Maximization Test (GPMT) (OECD, 1992) and the murine Local Lymph Node Assay (LLNA) (OECD, 2010). Both tests provide a fair assessment of skin sensitizing potential and have been successfully used to assist chemical risk assessment for several decades. However, as some chemicals are intrinsically more potent than others (concentration required to induce skin sensitization may vary by many orders of magnitude),

*Correspondence to: Donna S. Macmillan, Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, UK.
E-mail: donna.macmillan@lhasalimited.org

^aLhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, UK

^bBayer Pharma, Muellerstrasse 178, Wedding, Berlin, D-13342, Germany

^cBristol-Myers Squibb, 1 Squibb Drive, New Brunswick, NJ, 08903, USA

^dFirmenich S.A., Rue de la Bergère 7, Meyrin 2, CH-1217, Switzerland

^eFirmenich Inc., POB 5880, Princeton, NJ, 08543, USA

^fGlaxosmithKline, Park Road, Ware, SG12 0DP, UK

^gInternational Flavors & Fragrances Inc., 800 Rose Lane, Union Beach, NJ, 07735, USA

^hJanssen Research and Development, Turnhoutseweg 30, Beerse, B-2340, Belgium

ⁱRIFM, 50 Tice Boulevard, Woodcliff Lake, NJ, 07677, USA

differentiating only between sensitizing and non-sensitizing chemicals can lead to sub-optimal risk assessments, which in some cases may be over-protective or too lax.

In this respect, the LLNA offers a significant advantage over the GPMT. While the GPMT is typically used only to identify chemicals that represent a sensitizing hazard, the LLNA provides information on the potency of an identified sensitizer, thereby permitting an assessment of both hazard and potency. Other significant advantages of LLNA vs. GPMT are the reduced number of animals needed and the improved animal welfare (no adjuvants, no elicitation). Although modifications to the GPMT have been suggested to provide a more quantitative assessment of sensitising potential, these have yet to be fully validated or widely used (van Och, 2001).

Potency is inferred from the LLNA using the test concentration required to cause a threefold increase in the stimulation index (SI) relative to vehicle controls (Basketter et al., 2000) which is referred to as the EC3 value. Put simply, the lower the EC3 value, the more potent the sensitizer as a lower concentration is required to induce skin sensitization. EC3 values have proven reliable measures of potency and several studies have shown a good correlation between EC3 values and known sensitization potency in humans (Basketter et al., 2001, 2005; Gerberick et al., 2001). To aid chemical risk assessment, the European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) has proposed a classification scheme whereby EC3 values are used to categorize sensitizers into one of four potency categories: extreme ($< 0.1\%$), strong ($\geq 0.1 - < 1\%$), moderate ($\geq 1 - < 10\%$) or weak ($\geq 10\%$) (Kimber, 2001; ECETOC, 2003; Kimber et al., 2003). Similarly, a Globally Harmonized System of Classification and Labelling (GHS) was suggested by United Nations Economic Commission for Europe (UNECE) to standardize the classification of chemicals, labels and safety data sheets at a national, regional and worldwide level. Skin sensitizers are split into sub-category 1A and 1B, corresponding to EC3 values ≤ 2 and $> 2\%$, respectively (UNECE, 2015).

Despite the success of both the LLNA and GPMT to correctly identify sensitising chemicals, a significant amount of recent research has sought to develop non-animal alternatives. This has been motivated by a desire to improve animal welfare and to reduce the cost of chemical safety assessments by replacing animal tests. The cosmetics industry has a particular need for these alternative tests, driven largely by the implementation of European Union (EU) Regulation 1223/2009 (replacing EU Directive 76/768/EEC) which forbids the use of animal tests for cosmetic ingredients sold in the European Union (2013). To date, several *in chemico* and *in vitro* tests have been developed including the Direct Peptide Reactivity Assay [DPRA, (OECD 2015a)], KeratinoSens™ (OECD, 2015b) and the human Cell Line Activation Test [h-CLAT, (OECD 2016)], all of which have been validated and approved by the OECD for use as alternative methods to assess skin sensitization.

In silico models for skin sensitization potency

In addition to research into *in chemico* and *in vitro* tests, the last two decades have seen the development of several *in silico* models for predicting skin sensitization potential. Binary models for skin sensitization activity, where chemicals are classified as a sensitizer or non-sensitizer, have been available for many years; predicting skin sensitization potency is, however, a more difficult challenge. Recently, several commercial and non-commercial *in silico* systems have been developed that are capable of providing quantitative and semi-quantitative predictions of sensitization e.g. CADRE-SS (Kostal and Voutchkova-Kostal, 2016) assigns an ECETOC class to

a query chemical, TopKat (Enslein et al., 1997) can predict potency as weak or strong, CASE Ultra (MultiCASE) has several models for skin sensitization including separate models for weak, moderate and extreme sensitization and TIMES SS (Roberts et al., 2007) is a semi-quantitative hybrid expert system.

To fully replace existing animal assays, alternative tests and *in silico* models should provide not only reliable predictions of hazard but also match the performance of assays such as the LLNA in the prediction of potency. With this goal in mind, an EC3 prediction model has been developed based on publicly available LLNA studies and mechanistic (alert) domains as provided by Derek Nexus, an expert knowledge base system for toxicity predictions (Lhasa).

Materials and method

Derek Nexus

Derek Nexus 5.0.2 (using Derek Knowledge Base 2015 2.0 and Derek EC3 Model - 1.0.6), an expert knowledge base system for toxicity predictions (Lhasa), was used to make an *in silico* assessment of the skin sensitizing potential of the chemicals in the internal and external validation datasets. The datasets were processed against the skin sensitization endpoint, selecting mammal as the species of interest. Skin sensitization alerts are activated for, and describe haptens, pre-haptens and pro-haptens, meaning metabolic requirements have been considered in the structure of each alert. If the predicted skin sensitization outcome was defined as certain/probable/plausible/equivocal, the test chemical was considered as a sensitizer and the predicted EC3 value used, if available. If more than one skin sensitization alert was activated, the most potent EC3 value and corresponding alert were used. If no skin sensitization alert was activated with a likelihood of equivocal or above, the test chemical was considered as a non-sensitizer.

KNIME

Analysis of data and workflow construction was carried out using KNIME 3.1.2, a data analytics, reporting and integration platform freely available from <https://www.knime.org/> (Berthold et al., 2008).

Model construction

The model utilizes a k-Nearest Neighbours (kNN) approach. Initially, both compounds (query and dataset) must activate the same skin sensitization structural alert in Derek to be considered as a Nearest Neighbour. These Nearest Neighbours are then arranged by Tanimoto similarity generated by an in-house radial fingerprinting method. The minimum number of neighbours required is 3, otherwise no prediction is given, and the most similar neighbours, up to 10th place (Fig. 1), are considered (using standard competition ranking, e.g. 12 Nearest Neighbours may be used for the prediction if neighbours 10, 11 and 12 all have identical Tanimoto scores). The predicted EC3 value is the weighted average of all the valid neighbours as shown below in Eqn (1).

$$\frac{MW_Q}{EC3_Q} = \frac{\sum (Tanimoto * \frac{MW_{NN}}{EC3_{NN}})}{\sum Tanimoto} \quad (1)$$

where Q is query compound and NN is the Nearest Neighbour.

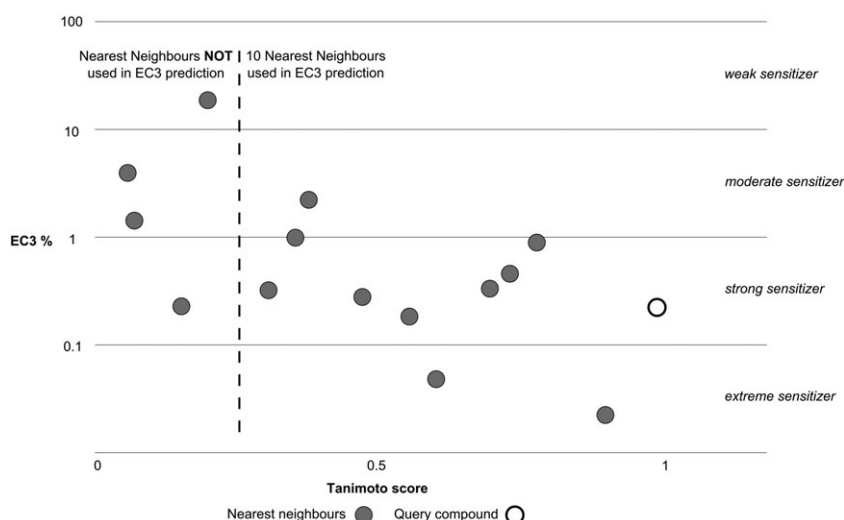


Figure 1. Example of 10 Nearest Neighbours used in a typical EC3 prediction.

Dataset and validation of model

The dataset comprised 1096 separate LLNA studies corresponding to 704 unique chemicals. Six hundred and fifty-nine of these compounds were used to develop the prediction model (training set, Derek EC3 Model - 1.0.6) and consisted of 84 non-sensitizers, 208 weak, 231 moderate, 90 strong and 46 extreme sensitizers. Compounds activating more than one alert were removed to prevent interference from potentially differing and/or competing mechanisms. Finally, after removal of compounds activating alerts with fewer than three Nearest Neighbours, the dataset consisted of 465 compounds. A separate dataset consisting of 45 compounds, containing 8 weak, 26 moderate, 8 strong and 3 extreme sensitizers, was used as a test set as shown in Fig. 2. The LLNA data were compiled from multiple sources of published literature using the EC3 value calculated by the author cited unless otherwise stated. For LLNA studies where only the stimulation index (SI) values were reported, an EC3 value was calculated using interpolation or

extrapolation as described by Ryan *et al.* (2007). The chemicals were chosen based on the quality of the LLNA studies, e.g. studies reporting LLNA negatives only tested to low concentrations (< 50%) were not used, leading to few non-sensitizers in the dataset.

Leave-one-out cross-validation (LOO-CV). A leave-one-out cross-validation (LOO-CV) of the prediction model was carried out using KNIME. One compound was removed from the dataset, the model built on the remaining compounds and the performance of the model against the individual compound was assessed. This was repeated for all compounds in the dataset with at least four compounds in the alert space ($n = 4$, one compound for the query and a minimum of three compounds per alert for an EC3 prediction) which reduced the dataset from $n = 465$ to $n = 407$.

Internal validation (test set). A test set was used for internal validation and consisted of 45 chemicals previously unseen by the prediction model. All chemicals activated at least one Derek Nexus

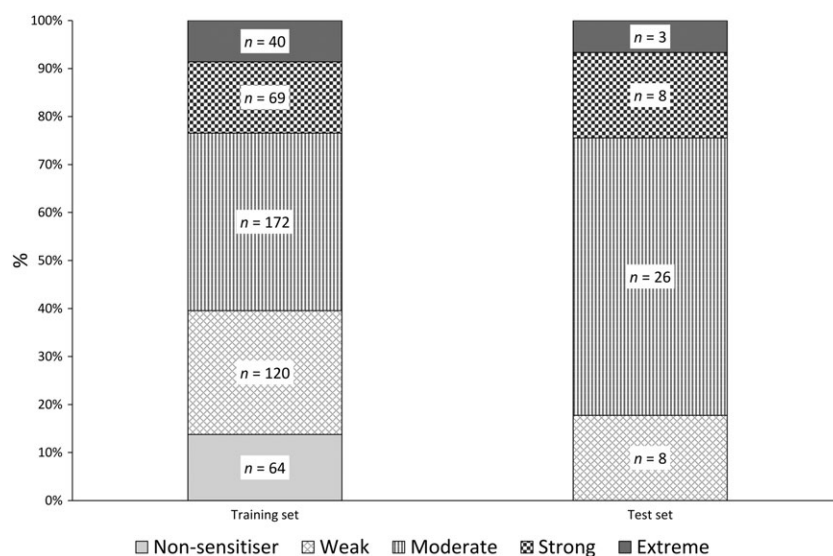


Figure 2. Distribution of compounds between the ECETOC classifications of skin sensitization potency in the training and test sets (number of compounds in each category is shown in bar segments. Corresponding % of dataset shown on Y-axis).

skin sensitization alert and all alerts activated had sufficient data (≥ 3 Nearest Neighbours) to predict an EC3 value. Where more than one skin sensitization alert was activated, the most potent EC3 value was taken as the predicted EC3.

External validation dataset. LLNA data were donated from seven Lhasa Limited members to validate the prediction model. Nine hundred and sixty-two unique compounds were donated, of which 328 had EC3 data. Compounds present in the EC3 training set, non-alerting compounds and compounds activating a Derek Nexus skin sensitization alert but with insufficient data to predict an EC3 value (< 3 Nearest Neighbours) were removed from the dataset. The overall external validation dataset contained 103 compounds previously unseen by the prediction model and contained 38 weak, 29 moderate, 24 strong and 12 extreme sensitizers.

Data transparency

The expertly reviewed and curated LLNA dataset ($n=1096$) described above used to develop the EC3 prediction model and the internal validation test set ($n=45$) will not be published within this article as we believe it to have commercial value, although all the data used are available in the public domain. Additionally, as the external validation dataset ($n=962$) consists solely of proprietary data from our members, publication of these data in this article is not permitted. This limitation means that others will not be able to verify or replicate the results of this study independently. However, the non-proprietary data are accessible to our members who sponsor our toxicity database Vitic Nexus, and we do intend to publish it in the future.

Results

Initial model development in KNIME

'Standard EC3' method. The initial development of the model arose from an investigation into the variance in the EC3 values for all compounds which activate each individual alert. The rationale was that if each alert described a unique mechanism, then it was possible that each mechanism would lead to a similar skin sensitization potency. To test this hypothesis, LLNA positive

compounds in an in-house EC3 dataset were gathered that activated only one skin sensitization alert in Derek. The EC3 values were collected on a per alert basis and the mean calculated. Each mean value was used as the 'standard EC3' value for that alert, provided that there were at least three compounds that activated the alert. In this manner, a 'standard EC3' value was calculated for 34 alerts. Figure 3 illustrates the range of EC3 values for three alerts and the area under the curve represents the relative frequency of each EC3 value. Some alerts only contained compounds with similar EC3 values e.g. within only one ECETOC category (Alert A, Fig. 3) whereas some alerts contained mixed values, with a small (Alert B, Fig. 3) or large range of EC3 values/ECETOC categories (Alert C, Fig. 3).

Nevertheless, when this model was applied to a selected test set containing 145 chemicals (Jaworska *et al.* 2013), it could assign the correct ECETOC category 48% of the time. This suggested that it would be possible to use the mechanistic separation of the alerts to aid us in the prediction of an EC3 value. The in-house dataset was then augmented with publicly available data and this expanded dataset used to develop the model further.

Nearest Neighbour method. The first attempt to improve the model was with an extension to a Nearest Neighbour method. With this method, the dataset was taken and again sub-selected to compounds activating only one alert. This was to ensure the chemicals in the training set were unlikely to undergo multiple different mechanisms at different sites. Unlike before negative LLNA compounds were included and assigned an EC3 value of 101 for calculation purposes. The EC3 values of the compounds in this dataset were predicted using a leave-one-out cross-validation to estimate the predictive efficiency of the method. For many of the alert spaces, there was only a limited number of compounds, so no other cross-validation was attempted. To calculate the EC3 value predicted for the query compound, firstly, those compounds activating the same alert were considered. Then only those examples which had a Tanimoto similarity to the query compound of greater than 0.5 were considered. The similarity threshold represented a balance between getting a sufficient number of compounds to be Nearest Neighbours and getting compounds similar enough that their activity might be informative in predicting for the query compound. Up to 10 most similar compounds in the

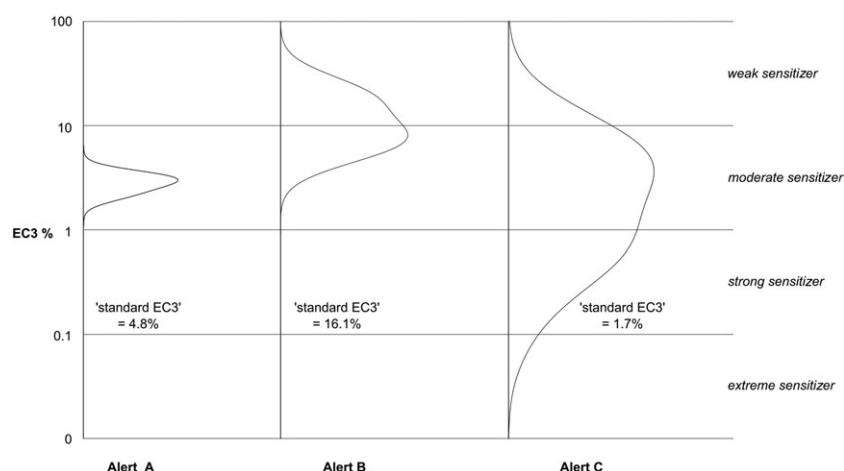


Figure 3. The range and 'standard EC3' for 3 skin sensitization alerts. The area under the curve represents the number of compounds smoothed assuming a normal distribution around the experimental EC3 value. Alert A contains only moderate sensitizers, Alert B has a range of weak, moderate and strong sensitizers and Alert C has compounds spread across the ECETOC categories.

dataset were used for the calculation; where there were fewer than three examples, the query compound was classified as out of domain. The EC3 was then calculated as the average of the MW/EC3 values weighted by the Tanimoto distance. This model improved the performance for the compounds which could be predicted by the model but resulted in a loss of coverage.

Two-tiered approach. To expand the coverage, a two-tiered approach was used. This consisted of taking an EC3 value from the Near Neighbour method where possible, supplemented by the 'standard EC3' method when the Near Neighbour method was unable to provide an answer. This two-tiered approach resulted in the same coverage as the 'standard EC3' method, but with increased accuracy for the compounds where it was possible to make a Nearest Neighbour prediction.

Optimization of a two-tiered approach. Optimization of the various parameters was undertaken, namely varying the fingerprinting engine used to derive a Tanimoto score, the minimum number of Nearest Neighbours and the minimum Tanimoto similarity. By modifying the fingerprinting engine between the in-house fingerprinting engine, Chemical Development Kit (CDK), Extended-Connectivity Fingerprints (ECFP) or MolPrint, it was possible to observe the impact of the fingerprinting method on the results. The in-house fingerprinting method performed similarly to CDK, and all other fingerprint methods performed significantly worse. The next parameter investigated was the minimum number of Nearest Neighbours required for generating a prediction; here the minimum was increased from 3 to 5 and then finally to 9. As expected, increasing the minimum number of neighbours needed for a prediction improved the accuracy – presumably because the chemical space was better defined – but the number of compounds out of domain also increased significantly. A minimum of three compounds in the example set required for a prediction was found to be a fair balance between accuracy and coverage. In the final optimization, the minimum Tanimoto threshold varied between 0, 0.5, 0.7 and 0.9. As with increasing the minimum number of neighbours, increasing the minimum Tanimoto similarity gave rise to an increase in prediction accuracy – the model would only consider increasingly similar compounds to be Nearest Neighbours – with a very steep drop in the number of compounds that were considered in domain.

It was realized that this method lacking the minimum Tanimoto threshold was effectively applying the two-tiered approach in one method. It was still able to leverage the mechanistic information encoded in each alert, as the requirement for the compounds to activate the same alert still stands, but without the threshold on the minimum Tanimoto value nearly every compound was in domain and the accuracy of the prediction was as high as using the two systems. This elegantly eliminated the two-system problem and made the model that was used henceforth.

The Lhasa EC3 prediction model. The optimized EC3 prediction model is shown in the schematic below (Fig. 4). First, a query compound is submitted to Derek Nexus and processed against skin sensitization. Upon activating a skin sensitization alert, the query is fingerprinted then compared against the fingerprints of compounds activating the same alert in the EC3 dataset (currently Derek EC3 model 1.0.6) to assign Nearest Neighbours. When three or more Nearest Neighbours are available an EC3 value is predicted and subsequently the predicted ECETOC category is shown. For queries activating alerts with fewer than three Nearest Neighbours in the EC3 dataset, no prediction is made.

Variability of the LLNA and choosing performance criteria for model

While assessing the performance of the EC3 model, quantifying the inherent variability of the LLNA was of significant interest. In a recent analysis of data taken from the publicly available NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) LLNA database (NICEATM 2013), Hoffmann (Hoffmann, 2015) calculated the median standard deviation of log(EC3) values as 0.252 ($n = 27$) and 0.312 ($n = 11$) and the variability of the LLNA to be ± 3.2 -fold and ± 4.2 -fold for compounds tested in the same vehicle and different vehicles, respectively. Roberts *et al.* (2016) reported the variability of the LLNA as twofold based on the analysis of 12 chemicals with multiple studies in the same vehicle. Both Hoffmann and Roberts' analyses are approximately in agreement with previously reported studies on the variability/reproducibility of the LLNA (ICCVAM 1999, 2009; Basketter and Cadby, 2004; Jowsey *et al.* 2008).

To assess the variability for the compounds in the dataset with multiple LLNA studies, compounds with three or more EC3 values were analysed using the different methods described above – due to data availability, separation by vehicle was not possible.

Hoffmann's methodology

The median standard deviation of the log(EC3) values was calculated as 0.245 ($n = 50$), which yielded an EC3 value variability of ± 3.1 -fold for the dataset which compared well with Hoffmann's previously reported figures. Furthermore, 90% of the 354 EC3 values that were analysed fell within ± 3.1 fold of the predicted EC3 value for each compound (Fig. 5).

Roberts' methodology

The overall standard deviation of the log(EC3) values was calculated as 0.314 ($n = 50$), which yielded an EC3 value variability of ± 4.3 -fold for the dataset and 94% of the 354 EC3 values analysed fell within ± 4.3 fold of the predicted EC3 value for each compound (Fig. 5).

The variability of the LLNA assay will directly affect the maximum possible performance of any model built upon this

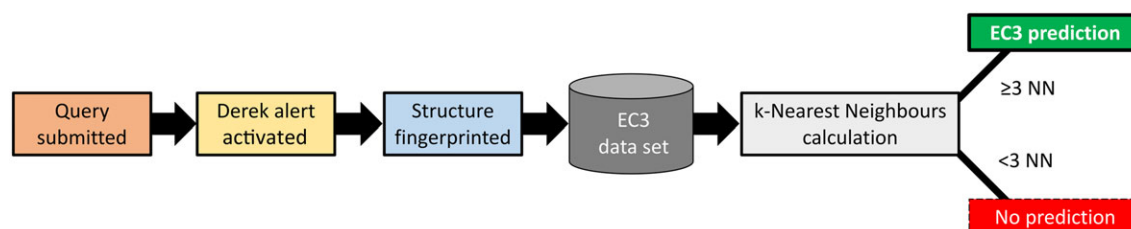


Figure 4. Schematic illustrating the EC3 prediction model after optimization. [Colour figure can be viewed at wileyonlinelibrary.com]

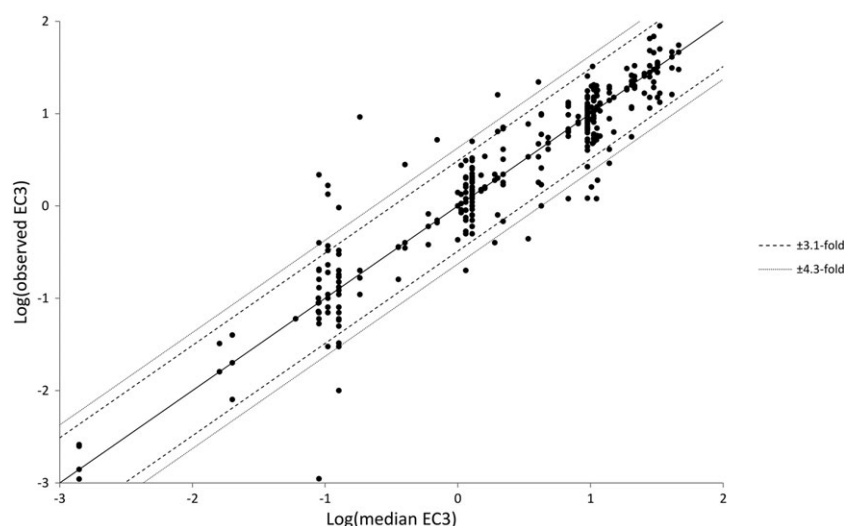


Figure 5. Scatter plot of the dataset showing variability of predicted log(EC3) values against observed log(EC3) values for compounds with multiple LLNA studies using Hoffmann's and Roberts' methodologies, respectively. The dashed line represents the calculated variability of ± 3.1 -fold using Hoffmann's methodology and grey line represents the calculated variability of ± 4.3 -fold using Roberts' methodology.

training data. Given the values of ± 3.1 -fold and ± 4.3 -fold variability calculated in the EC3 data using the methodologies above, assessing the prediction model's performance within three- and fivefold of the experimental EC3 seemed appropriate.

Model performance and validation

The EC3 prediction model was assessed by performing three different validation techniques:

- *Cross-validation* – this refers to the leave-one-out cross-validation (LOO-CV) technique.
- *Internal validation* (test set) – this refers to the validation using an in-house dataset collected from the publicly available literature by scientists at Lhasa Limited ($n = 45$). As these data were excluded from the training set used to build the EC3 model it forms a suitable test set to validate the model.
- *External validation* – this refers to the validation using LLNA data donated to Lhasa Limited by a number of members with the specific purpose of providing a test set of unseen compounds ($n = 103$).

Then four different criteria were chosen to assign a prediction as correct:

Fold accuracy:

- Predicted EC3 value falls within $\pm$ threefold of the experimental value (henceforth known as *3-fold accuracy*).
- Predicted EC3 value falls within $\pm$ fivefold of the experimental value (*5-fold accuracy*).

Category accuracy (Table 1):

- ECETOC potency category – predicted potency category matches the experimentally determined potency category (*ECETOC accuracy*).
- GHS classification category – predicted potency category matches the experimentally determined potency category (*GHS accuracy*).

Table 1. Classification categories of ECETOC and GHS and corresponding EC3 values

ECETOC Classification	
EC3 (%)	Category
≥ 10	Weak sensitizer
$\geq 1 - < 10$	Moderate sensitizer
$\geq 0.1 - < 1$	Strong sensitizer
< 0.1	Extreme sensitizer
GHS Classification	
EC3 (%)	Category
≤ 2	1A - strong sensitizer
> 2	1B - other sensitizer

Cross-validation

Fold accuracy. 151/407 (37%) compounds were correctly assessed within a threefold range of the experimentally derived EC3 value (Fig. 6). 207/407 (51%) were over-predicted (where the prediction value was more potent than the experimental value), and 49/407 (12%) were under-predicted (where the prediction value was less potent than the experimental value).

198/407 (49%) were correctly assessed within a fivefold range of the experimental EC3, 171/407 (42%) were over-predicted, and 38/407 (9%) were under-predicted.

Category accuracy. 157/407 (39%) compounds were correctly assigned to their ECETOC category (Fig. 6). 198/407 (49%) were over-predicted (assigned a more potent category than observed experimentally) and 52/407 (13%) were under-predicted (assigned a less potent category than observed experimentally).

224/407 (55%) compounds were correctly assigned to their GHS category. 146/407 (36%) were over-predicted, and 37/407 (9%) were under-predicted (Fig. 6).

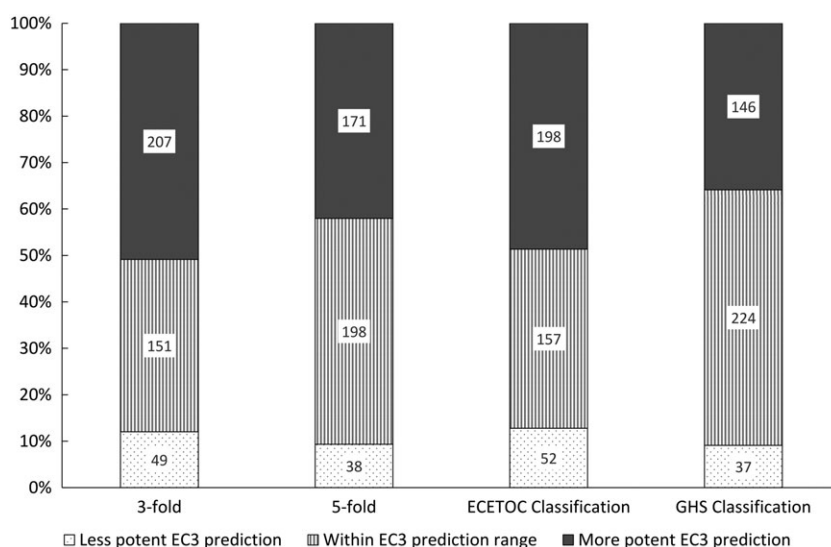


Figure 6. Leave one out cross-validation results (number of compounds in each category is shown in bar segments. Corresponding % of dataset shown on Y-axis).

Internal validation (test set)

Fold accuracy. 20/45 (44%) compounds were correctly assessed within a threefold range of the experimentally derived EC3 value (Fig. 7). 16/45 (36%) were over-predicted, and 9/45 (20%) were under-predicted.

27/45 (60%) were correctly assessed within a fivefold range of the experimental EC3, 12/45 (27%) were over-predicted, and 6/45 (13%) were under-predicted.

Category accuracy. 23/45 (51%) compounds were correctly assigned to their ECETOC category. 15/45 (33%) were over-predicted, and 7/45 (16%) were under-predicted.

29/45 (64%) compounds were correctly assigned to their GHS category. 11/45 (24%) were over-predicted, and 5/45 (11%) were under-predicted (Fig. 7).

External validation

Fold accuracy. 30/103 (29%) compounds were correctly assessed within a threefold range of the experimentally derived EC3 value. 49/103 (48%) were over-predicted, and 24/103 (23%) were under-predicted.

51/103 (50%) were correctly assessed within a fivefold range of the experimental EC3, 38/103 (37%) were over-predicted, and 14/103 (14%) were under-predicted.

Category accuracy. 37/103 (36%) compounds were correctly assigned to their ECETOC category. 46/103 (45%) were over-predicted, and 19/103 (18%) were under-predicted.

66/103 (64%) compounds were correctly assigned to their GHS category. 26/103 (25%) were over-predicted, and 11/103 (11%) were under-predicted (Fig. 8).

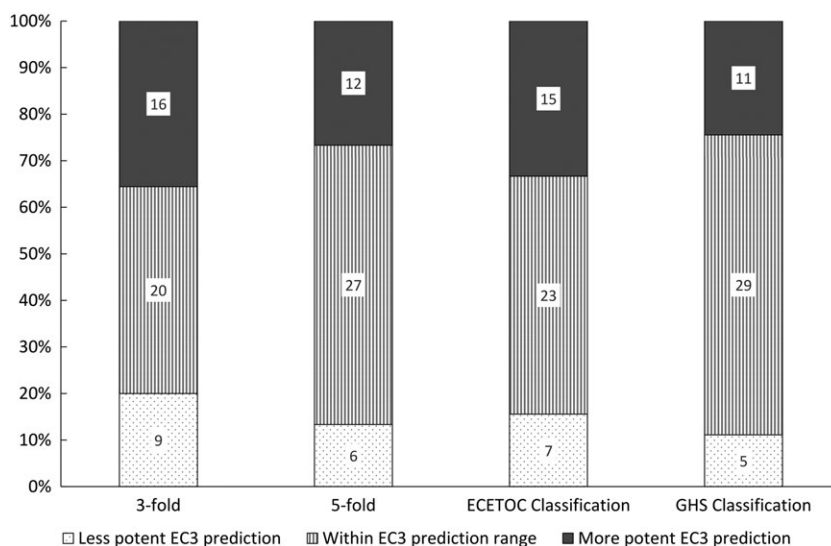


Figure 7. Internal validation (test set) results (number of compounds in each category is shown in bar segments. Corresponding % of dataset shown on Y-axis).

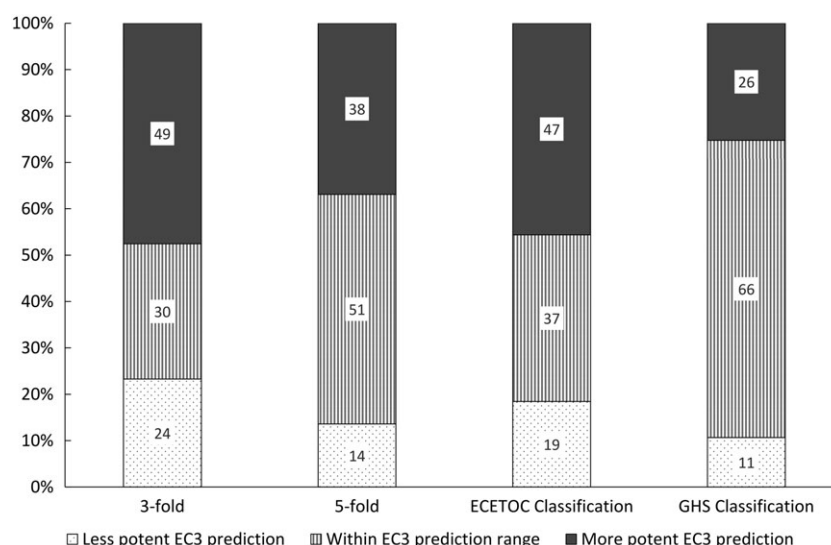


Figure 8. External validation results (number of compounds in each category is shown in bar segments. Corresponding % of dataset shown on Y-axis).

Discussion

Our results confirm that publicly available, curated, LLNA data can be used in a k-Nearest Neighbours model to predict the EC3 value of skin sensitizers. This novel EC3 prediction model has good alert coverage and predicts well for both internal and external validation datasets although some chemicals and/or alert domains are predicted incorrectly. The case studies below tackle the most common reasons for mispredictions observed in our internal dataset (Table 2).

Limitations of prediction model

Fixed category classifications. There are a number of compounds which, although numerically very close in EC3 value, fall into different ECETOC or GHS categories e.g. CS1, CS2 and CS3 are mispredicted by ECETOC categories and CS4, CS5, CS6, CS7, CS8 are mispredicted by GHS categories. CS1 is predicted as a moderate sensitizer (EC3 = 1.4) when experimentally it is strong (EC3 = 0.49). CS2 is predicted as weak (EC3 = 14) when experimentally it is moderate (EC3 = 9.41) and CS3 is predicted as an extreme

Table 2. List of chemicals from the internal validation test set mispredicted by EC3 prediction model. ✓ = correctly predicted by accuracy criteria, ✗ = mispredicted by accuracy criteria

Name	Accuracy Criterion				Alert Number	Experimental EC3%*	Predicted EC3%	Fold Error
	3-fold	5-fold	ECETOC	GHS				
CS1	✓	✓	✗	✓	414	0.49	1.4	2.86
CS2	✓	✓	✗	✓	481	9.41	14	1.49
CS3	✓	✓	✗	✓	447	0.10	0.056	0.56
CS4	✓	✓	✓	✗	401	1.94	3.6	1.86
CS5	✓	✓	✓	✗	414	2.00	2.3	1.15
CS6	✓	✓	✓	✗	435	1.37	2.5	1.82
CS7	✗	✓	✓	✗	432	4.05	1.3	0.32
CS8	✓	✓	✓	✗	443	2.15	1.8	0.84
CS9	✗	✗	✗	✓	443	0.04	0.4	10.0
CS10	✗	✓	✓	✓	443	0.16	0.74	4.63
CS11	✗	✗	✓	✓	443	0.12	0.73	6.08
CS12	✗	✗	✗	✓	439	21.46	3	0.14
CS13	✗	✗	✗	✗	439	9.33	0.017	0.0018
CS14	✗	✗	✗	✗	439	22.07	0.16	0.01
CS15	✗	✗	✗	✓	439	45.47	3.3	0.07
CS16	✗	✗	✗	✓	462	4.47	100	22.4
CS17	✗	✓	✓	✓	462	24.00	100	4.17

* EC3 values calculated from reported SI values.

sensitizer (EC₃ = 0.056) but is observed as a strong sensitizer experimentally (EC₃ = 0.1). CS₄, CS₅ and CS₆ are predicted as category 1B when they are observed experimentally as 1A. Conversely, CS₇ and CS₈ are predicted as 1A sensitizers but are 1B sensitizers experimentally. However, it is clear to see that a number of these mispredicted compounds are very close to the category thresholds (e.g. only one ECETOC category out/very close to the GHS threshold) and cases like this will always be seen when using arbitrarily defined thresholds. Furthermore, given the inherent variability in the LLNA, one might expect the results to cross the boundaries in many of those cases upon duplicate testing.

Challenging alert domains. As previously observed when beginning to build the prediction model, some alert domains were significantly easier to predict than others. As Derek alerts are mechanism-based one might expect similar EC₃ values from compounds activating the same alert. However, within this test set, some alert domains such as 443 (metal or metal salt) and 439 (substituted phenol or precursor) are more challenging to predict than others. For example, CS₈, CS₉, CS₁₀ and CS₁₁ all activate alert 443 (metal or metal salt) and are all mispredicted by at least one accuracy criterion with a fold error ranging from 0.84- to 10.0-fold of the experimental EC₃ value. Some sensitizing metal salts are known to perform poorly in the LLNA e.g. nickel sulphate producing conflicting results in the LLNA between studies (Gerberick *et al.*, 2000; ICCVAM, 2012). This then can have an impact on the LLNA data for the alert domain subsequently influencing the accuracy of the alert predictivity. Another challenging alert domain is phenols and precursors, CS₁₂, CS₁₃, CS₁₄, and CS₁₅ are all mispredicted by more than one accuracy criteria with a fold error ranging from 0.0018- to 0.14-fold. Phenols and precursors are thought to elicit skin sensitization through several possible mechanisms (Barratt and Basketter, 1992), and these different mechanisms of action may have a bearing on the sensitization potential for specific chemicals within the class. As a result, the mechanistic basis for the phenol alert has been refined and a subset of compounds given a separate alert to improve predictivity.

Need for coverage in dataset. CS₁₆ and CS₁₇ activate alert 462 (alkyl ester of phosphoric or phosphonic acid) and are predicted as weak sensitizers (EC₃ = 100). CS₁₆ is only predicted correctly by GHS accuracy whereas CS₁₇ is predicted correctly by all methods except a threefold accuracy. These mispredictions are caused by poor coverage in the dataset. Alert 462 in Derek Nexus is built using mainly guinea pig data and LLNA data lacking EC₃ values. As only 3 compounds in the dataset currently activate alert 462, all of which are non-sensitizers, the prediction model currently makes a prediction of EC₃ = 100 for all compounds activating alert 462. However, the compounds in the internal dataset will be added to the dataset in the future which will improve the predictions for this alert (and others activated in the internal dataset).

Alert coverage

In particular, the alert coverage of the model is impressive considering that the training set is relatively small. Alert coverage of the EC₃ prediction model was assessed by calculating the number of skin sensitization alerts activated by compounds in the previously unseen external validation dataset and comparing to the number of alerts the prediction model can make a prediction for Eqn (2).

$$\text{Alert coverage of external dataset} = \frac{\text{no. of alerts with EC}_3 \text{ predictions}}{\text{no. of alerts activated overall}} * 100 \quad (2)$$

The model predicts an EC₃ value for every skin sensitization alert in Derek Nexus unless there are fewer than three compounds activating the same alert in the dataset. Forty-five skin sensitization alerts were activated in the external validation dataset (including exact matches in the dataset) of which 40 could predict an EC₃ value. However, these 40 predicting alerts corresponded to almost 99% of the external dataset due to multiple compounds activating identical alerts. Four of the 5 non-predicting alerts were solely derived from guinea pig data, and the remaining alert had fewer than three EC₃ values from which to calculate an EC₃ prediction. Furthermore, as the 5 non-predicting alerts were single compounds, this equated to only 1.4% of the external dataset.

Impact of expert review on mispredicted chemicals

The Lhasa EC₃ prediction model, when it mispredicts a chemical, tends to predict a more potent EC₃ value. There are a number of reasons that this may be preferred to under-prediction. The most cited reason is that regulators and risk assessors consider it prudent to be vigilant and use a more conservative prediction for chemicals which may have an impact on human health. However, over-prediction is not without issue as it can lead to an over cautious approach to chemicals. To this end, the user may apply an expert review of the data and add and/or remove compounds within the alert domain in the software version of the prediction model (currently Derek Nexus v5.0.2 and Derek EC₃ Model - 1.0.6). A case study using 1-methoxy-4-(prop-1-en-1-yl)benzene is shown in Fig. 9. Based on the Nearest Neighbours which activate the same alert, substituted phenol or precursor, the predicted EC₃ is 0.014%. However, the majority of the compounds within this alert domain are significantly less potent. Upon further investigation of the structure of the lone extreme sensitizer, 1-(4-hydroxyphenyl)-3-phenylprop-2-en-1-one (EC₃ = 0.0016%), it is evident that this compound may undergo an alternative Michael acceptor mechanism or a quinone formation/phenolic radical mechanism which may contribute to its potency. Removing this compound from the EC₃ prediction led to a post-expert review predicted EC₃ of 4.1%, much more in line with the other compounds within this alert domain and the experimentally derived EC₃ value of 2.3%.

In summary, a novel quantitative *in silico* EC₃ prediction model used within Derek Nexus is presented, based on a k-Nearest Neighbours approach. When a query compound matches a skin sensitization alert in the knowledge base and has 3 or more Nearest Neighbours activating the same alert in the EC₃ dataset, an EC₃ prediction is made. The model has a good predictive ability, predicting 64% of chemicals in an external test set within the correct GHS category. Of the remaining chemicals in the previously unseen dataset, 25% were over-predicted (GHS 1A predicted; GHS 1B experimentally) and 11% were under-predicted (GHS 1B predicted; GHS 1A experimentally). As the classification categories used to assess the skin sensitization risk of a particular substance have strict limits, if a chemical is mispredicted the model is designed to predict a more conservative EC₃ value to help protect human health and the environment. Finally, expert review of the 10 Nearest Neighbours can also help to fine-tune the EC₃ value supplied by the model. The accuracy of the model has room for improvement, however, the accuracy will always be limited by

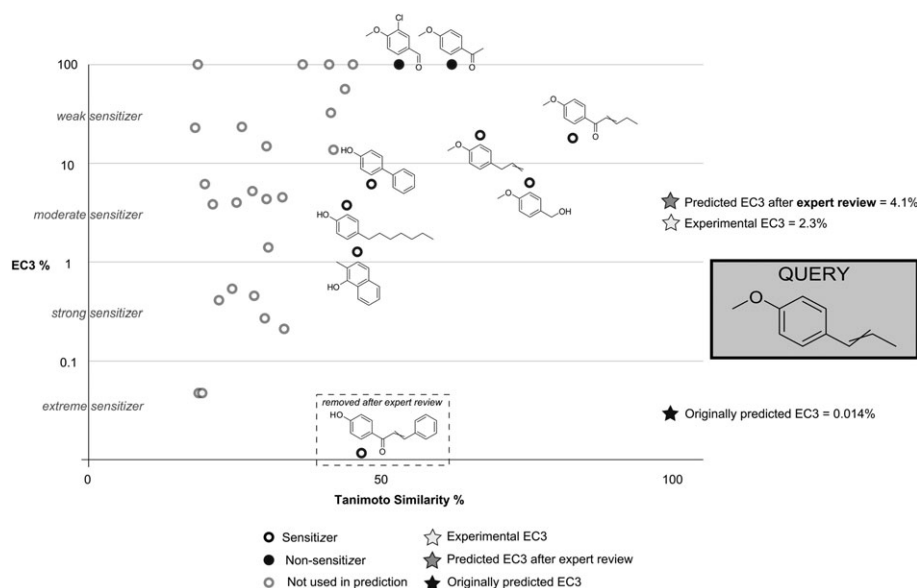


Figure 9. The impact of expert review on the predicted EC3 value of 1-methoxy-4-(prop-1-en-1-yl)benzene.

the inherent variability of the LLNA assay (three- to fourfold). Approaches to improve the predictive capacity of the model are ongoing. Further work will focus on encouraging the donation of data not yet available in the public domain and more specifically, data for those alerts which have insufficient data to predict EC3 values at present. Finally, we are eager to compare our EC3 prediction model to other statistical skin sensitization QSARs and an in-depth analysis is currently under investigation.

Conflict of interest

The authors declare that there are no conflict of interest in this paper.

References

- Barratt MD, Basketter DA. 1992. Possible origin of the skin sensitization potential of isoeugenol and related compounds. *Contact Dermatitis* **27**: 98–104. DOI:10.1111/j.1600-0536.1992.tb05217.x.
- Basketter DA, Blaikie L, Dearman RJ, Kimber I, Ryan CA, Gerberick GF, Harvey P, Evans P, White IR, Rycroft RJG. 2000. Use of the local lymph node assay for the estimation of relative contact allergenic potency. *Contact Dermatitis* **42**: 344–348. DOI:10.1034/j.1600-0536.2000.042006344.x.
- Basketter DA, Cadby P. 2004. Reproducible prediction of contact allergenic potency using the local lymph node assay. *Contact Dermatitis* **50**: 15–17. DOI:10.1111/j.0105-1873.2004.00278.x.
- Basketter DA, Clapp C, Jefferies D, Safford B, Ryan CA, Gerberick F, Dearman RJ, Kimber I. 2005. Predictive identification of human skin sensitization thresholds. *Contact Dermatitis* **53**: 260–267. DOI:10.1111/j.0105-1873.2005.00707.x.
- Basketter DA, Wright ZM, Warbrick EV, Dearman RJ, Kimber I, Ryan CA, Gerberick GF, White IR. 2001. Human potency predictions for aldehydes using the local lymph node assay. *Contact Dermatitis* **45**: 89–94. DOI:10.1034/j.1600-0536.2001.045002089.x.
- Berthold MR, Cebon N, Dill F, Gabriel TR, Köster T, Meinel T, Ohl P, Sieb C, Thiel K, Wiswedel B. 2008. In *KNIME: The Konstanz Information Miner*, Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds). Springer: Berlin; 319–326.
- Diepgen TL, Kanerva L. 2006. Skin diseases in Europe. *Eur. J. Dermatol.* **16**: 324–330.
- ECETOC. 2003. In *Contact Sensitization: Classification According to Potency*, Carpanini FMB (ed). European Centre for Ecotoxicology and Toxicology of Chemicals: Brussels. Technical Report No. 87.
- Enslein K, Gombar VK, Blake BW, Maibach HI, Hostynek JJ, Sigman CC, Bagheri D. 1997. A quantitative structure-toxicity relationships model for the dermal sensitization guinea pig maximization assay. *Food Chem. Toxicol.* **35**: 1091–1098. DOI:10.1016/S0278-6915(97)87277-8.
- Gerberick FG, Ryan CA, Kimber I, Dearman RJ, Lea LJ, Basketter DA. 2000. Local lymph node assay: Validation assessment for regulatory purposes. *Am. J. Contact Dermatit.* **11**: 3–18. DOI:10.1016/S1046-199X(00)90027-7.
- Gerberick GF, Robinson MK, Ryan CA, Dearman RJ, Kimber I, Basketter DA, Wright Z, Marks JG. 2001. Contact allergenic potency: correlation of human and local lymph node assay data. *Am. J. Contact Dermat.* **12**: 156–161. DOI:10.1053/ajcd.2001.23926.
- Hoffmann S. 2015. LLNA variability: An essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. *ALTEX* **32**: 379–383. DOI:10.14573/altex.1505051.
- ICCVAM. 1999. *The Murine Local Lymph Node Assay: A Test Method for Assessing the Allergic Contact Dermatitis Potential of Chemicals/Compounds*. National Institute of Environmental Health Sciences: Research Triangle Park, NC.
- ICCVAM. 2009. *Recommended Performance Standards: Murine Local Lymph Node Assay*. National Institute of Environmental Health Sciences: Research Triangle Park, NC.
- ICCVAM. 2012. *Test Method Evaluation Report on Using the Murine Local Lymph Node Assay for Testing Pesticide Formulations, Metals, Substances in Aqueous Solutions, and Other Products*. National Institute of Environmental Health Sciences: Research Triangle Park, NC.
- Jaworska J, Dancik Y, Kern P, Gerberick F, Natsch A. 2013. Bayesian integrated testing strategy to assess skin sensitization potency: from theory to practice. *J. App. Toxicol.* **33**: 1353–1364. DOI:10.1002/jat.2869.
- Jowsey IR, Clapp CJ, Safford B, Gibbons BT, Basketter DA. 2008. The impact of vehicle on the relative potency of skin-sensitizing chemicals in the local lymph node assay. *Cutan. Ocul. Toxicol.* **27**: 67–75. DOI:10.1080/15569520801904655.
- Kimber I. 2001. Skin Sensitization Testing in Potency and Risk Assessment. *Toxicol. Sci.* **59**: 198–208. DOI:10.1093/toxsci/59.2.198.
- Kimber I, Basketter DA, Butler M, Gamer A, Garrigue JL, Gerberick GF, Newsome C, Steiling W, Vohr HW. 2003. Classification of contact allergens according to potency: proposals. *Food Chem. Toxicol.* **41**: 1799–1809. DOI:10.1016/S0278-6915(03)00223-0.
- Kostal J, Voutchkova-Kostal A. 2016. CADRE-SS, an in Silico Tool for Predicting Skin Sensitization Potential Based on Modeling of Molecular Interactions. *Chem. Res. Toxicol.* **29**: 58–64. DOI:10.1021/acs.chemrestox.5b00392.
- Lhasa Limited. Derek Nexus v5.0.2. <http://www.lhasalimited.org>
- MultiCASE Inc. CASE Ultra. <http://www.multicase.com>
- NICEATM. 2013. Murine local lymph node assay (LLNA) database. <http://ntp.niehs.nih.gov/iccvm/methods/immunotox/niceatm-llnadatabase-23dec2013.xls>

- OECD. 1992. Test No. 406: Skin Sensitization.
- OECD. 2010. Test No. 429: Skin Sensitization: Local Lymph Node Assay.
- OECD. 2015a. Test No. 442C: In Chemico Skin Sensitization: Direct Peptide Reactivity Assay (DPRA).
- OECD. 2015b. Test No. 442D: In Vitro Skin Sensitization: ARE-Nrf2 Luciferase Test Method.
- OECD. 2016. Test No. 442E: In Vitro Skin Sensitization.
- Roberts DW, Api AM, Aptula AO. 2016. Chemical applicability domain of the Local Lymph Node Assay (LLNA) for skin sensitization potency. Part 2. The biological variability of the murine Local Lymph Node Assay (LLNA) for skin sensitization. *Reg. Toxicol. Pharmacol.* DOI:10.1016/j.yrtph.2016.07.013.10.1016/j.yrtph.2016.07.013
- Roberts DW, Patlewicz G, Dimitrov SD, Low LK, Aptula AO, Kern PS, Dimitrova GD, Comber MI, Phillips RD, Niemela J, Madsen C, Wedebye EB, Bailey PT, Mekenyan OG. 2007. TIMES-SS—a mechanistic evaluation of an external validation study using reaction chemistry principles. *Chem. Res. Toxicol.* **20**: 1321–1330. DOI:10.1021/bx700169w.
- Ryan CA, Chaney JG, Gerberick GF, Kern PS, Dearman RJ, Kimber I, Basketter DA. 2007. Extrapolating local lymph node assay EC3 values to estimate relative sensitizing potency. *Cutan. Ocul. Toxicol.* **26**: 135–145. DOI:10.1080/15569520701212258.
- UNECE. 2015. *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*. United Nations: New York.
- European Union. 2013. Commission Regulation (EU) No 1197/2013 of 25 November 2013 amending Annex III to Regulation (EC) No 1223/2009 of the European Parliament and of the Council on cosmetic products. *Off. J. Eur. Union* **56**: 34–66. DOI:10.3000/19770677.L_2013.315.eng.
- van Och F. 2001. Comparison of dose–responses of contact allergens using the guinea pig maximization test and the local lymph node assay. *Toxicol. Sci.* **167**: 207–215. DOI:10.1016/S0300-483X(01)00425-5.