Toxicology in Vitro 36 (2016) 197–209

Contents lists available at ScienceDirect

Toxicology in Vitro

journal homepage: www.elsevier.com/locate/toxinvit

Consensus of classification trees for skin sensitisation hazard prediction

D. Asturiol *, S. Casati, A. Worth

Joint Research Centre, Via Enrico Fermi 2749, Ispra 21027, VA, Italy

ARTICLE INFO

Article history:
Received 18 December 2015
Received in revised form 8 July 2016
Accepted 21 July 2016
Available online 22 July 2016

Keywords:
QSAR
Skin sensitisation
In vitro
In silico
Prediction
Decision tree

ABSTRACT

Since March 2013, it is no longer possible to market in the European Union (EU) cosmetics containing new ingredients tested on animals. Although several *in silico* alternatives are available and achievements have been made in the development and regulatory adoption of skin sensitisation non-animal tests, there is not yet a generally accepted approach for skin sensitisation assessment that would fully substitute the need for animal testing.

The aim of this work was to build a *defined approach* (*i.e.* a predictive model based on readouts from various information sources that uses a fixed procedure for generating a prediction) for skin sensitisation hazard prediction (sensitiser/non-sensitiser) using Local Lymph Node Assay (LLNA) results as reference classifications. To derive the model, we built a dataset with high quality data from *in chemico* (DPRA) and *in vitro* (KeratinoSens™ and h-CLAT) methods, and it was complemented with predictions from several software packages.

The modelling exercise showed that skin sensitisation hazard was better predicted by classification trees based on *in silico* predictions.

The defined approach consists of a consensus of two classification trees that are based on descriptors that account for protein reactivity and structural features. The model showed an accuracy of 0.93, sensitivity of 0.98, and specificity of 0.85 for 269 chemicals. In addition, the defined approach provides a measure of confidence associated to the prediction.

1. Introduction

The assessment of skin sensitisation potential represents a key requirement within several pieces of chemicals' regulations in the EU. For example, the REACH regulation (EC, 2006) foresees that chemicals produced or marketed in quantities of one tonne or more *per annum* be assessed for their potential to cause allergic contact dermatitis in humans, and within the Cosmetics Regulation (EC, 2009) skin sensitisation is one of the toxicological endpoints that require particular focus. The REACH regulation demands that testing on vertebrate animals should be considered only as last resort. The Cosmetics Regulation banned the animal testing of cosmetic ingredients in 2009 and the marketing of cosmetics containing new ingredients tested on animals in 2013 (EC, 2009).

The main chemical and biological mechanisms underpinning skin sensitisation are established (Karlberg et al., 2008; Martin, 2015; Martin et al., 2011) and have been described in the form of an adverse outcome pathway (AOP) (OECD, 2012a,b). Within this AOP, four key events (KE) are considered necessary for the acquisition of skin sensitisation: the covalent binding to skin proteins (KE-1) – also considered to be the molecular initiating event (MIE) –, the activation of keratinocytes (KE-2), the maturation of dendritic cells (KE-3), and the activation and proliferation of memory T-cells.

Progress has been made over the past ten years in the development of non-testing and testing methods addressing the key events of the skin sensitisation AOP. Three animal-free methods that account for KEs 1, 2, and 3 have been formally assessed by the European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM). These methods are: the direct peptide reactivity assay (DPRA) (EURL ECVAM, 2013; Gerberick et al., 2004), KeratinoSens™ (Emter et al., 2010; EURL ECVAM, 2014; Natsch and Emter, 2008), and the human cell-line activation test (h-CLAT) (Ashikaga et al., 2006; EURL ECVAM, 2015; Sakaguchi et al., 2006).

The DPRA, KeratinoSens™, and h-CLAT have been adopted by the Organisation for Economic Cooperation and Development (OECD) as Test Guidelines 442C (OECD, 2015a) and 442D (OECD, 2015b) and 442E (not yet published), respectively. Despite these methods predict LLNA responses with an accuracy of about 80% they are not proposed to be used as standalone alternatives. One of the reasons put forward for this is that they model specific KEs of the AOP and not the final adverse effect.

Progress has been made in the integration of results from *in silico*, *in chemico* and *in vitro* methods in defined approaches (OECD, 2016a, 2016b) to improve skin sensitisation hazard/potency prediction with respect to the individual methods. The first approach of this kind was developed by Natsch et al. (Natsch et al., 2009). The authors made a

* Corresponding author at: European Commission; Directorate General Joint Research Centre; Directorate F – Health, Consumers and Reference Materials; Chemicals Safety and Alternative Methods, VA, Italy.
E-mail address: david.asturiol-bofill@ec.europa.eu (D. Asturiol).

http://dx.doi.org/10.1016/j.tiv.2016.07.014

proof of concept of the prediction model based on scores proposed by Jowsey et al. (Jowsey et al., 2006), which was intended for predicting skin sensitisation potency. The model did not predict LLNA potency successfully, but a good performance was achieved in predicting skin sensitisation hazard for 116 chemicals (sensitivity = 0.86, specificity = 0.94, and accuracy = 0.88). Since then, a number of other approaches integrating non-animal data which use the AOP as a framework and are proposed for skin sensitisation hazard and/or potency assessments have been published. These range from simple weight-of-evidence (WoE) approaches (*e.g.* Bauch et al., 2012; Guyard-Nicodème et al., 2015; Macmillan et al., 2016; Urbisch et al., 2015), tiered approaches involving interim decision steps at the end of each tier (*e.g.* Takenouchi et al., 2015; van der Veen et al., 2014), and multiple regression models (Natsch et al., 2015) to more complex mathematical models (MacKay et al., 2013), artificial neural networks (*e.g.* Hirota et al., 2013, 2015; Tsujita-Inoue et al., 2014), and support vector machine-based approaches (Strickland et al., 2016). Another model integrating data from various sources and developed for LLNA potency prediction is the one based on a Bayesian Network (Jaworska, 2011; Jaworska et al., 2013, 2015). Bayesian networks are probabilistic models that can work with data gaps and can guide additional testing by quantifying the additional test information value before performing the testing.

Some authors have analysed in detail the performance of various *in silico* methods and expert systems when predicting skin sensitisation potential (Teubner et al., 2013; van der Veen et al., 2014). They showed that in general this kind of skin sensitisation methods had sensitivities above 0.70 and specificities below 0.65, even when some of them were combined. They concluded that the methods evaluated were not sufficiently accurate to be broadly used for skin sensitisation prediction. Alves et al. (Alves et al., 2015) recently showed that random forest models built from *in silico* descriptors obtained from the 2D structure of chemicals can have higher accuracy and larger applicability domains than the *in silico* methods reviewed by Teubner et al. and van der Veen et al. Alves et al. developed a series of consensus random forest models that predict skin sensitisation hazard (sensitiser *vs.* non sensitiser) using LLNA results as reference data. Their models used descriptors calculated with Dragon (Talete Srl, 2010) and SiRMS (Muratov et al., 2010) and were applied to a total of 406 chemicals, the largest skin sensitisation dataset published to date. The authors finally used a model based on a consensus of random forests that showed an accuracy of 0.82, sensitivity of 0.79, and specificity of 0.85 for the training set. These predictive performance values were obtained for 82% of the chemicals of the training set (chemical space coverage = 82%) as the predictions of the remaining 18% of chemicals were discarded because the two forests had contradictory outputs and the overall prediction was considered equivocal. The corresponding statistics for the validation set are not reported here as they are not representative because the validation set was highly unbalanced, *i.e.* contained 152 sensitisers and only 5 non-sensitisers. It is worth mentioning that the coverages of the validation sets of the different models developed by Alves et al. were significantly lower than those of the training sets, being of 50% the highest amount of chemicals of the validation tests that could be predicted.

The aim of our work was to build a model for predicting skin sensitisation hazard (sensitiser/non-sensitiser) that was simple, accurate, highly sensitive, and if possible integrating data from different sources, *i.e.* a defined approach. In order to develop the best model possible we have built a high quality database of 269 chemicals with LLNA data and skin sensitisation results obtained from DPRA, KeratinoSens™, and h-CLAT. The dataset has been quality checked by EURL ECVAM in collaboration with the test developers, and has been completed with a number of descriptors predicted with several free and licensed software packages yielding about 4500 descriptors for each of the 269 chemicals. This database has been used to build different classification trees to predict skin sensitisation hazard using LLNA results as reference. The two trees with the highest specificities and accuracies against LLNA classifications were used in a conservative consensus approach as final prediction model. In addition, a qualitative confidence measure on the prediction was added to the model by taking into account the leaves that were used in each individual tree to obtain the final consensus prediction.

2. Materials and methods

2.1. Dataset compilation

A dataset of 269 organic chemicals (170 sensitisers and 99 non sensitisers) identified by their chemical name and SMILES codes with *in chemico, in vitro*, and *in vivo* skin sensitisation data (LLNA and human) was built to develop a model to predict skin sensitisation hazard.

The initial collected dataset contained a total of 315 substances with human and/or LLNA data. Of these, 22 substances with only human data available were not considered. 16 inorganic chemicals and two mixtures (Pepperwood and Kathon CG) were discarded since they could not be calculated with most *in silico* software packages. Ammonium peroxodisulphate was also discarded because it was considered an inorganic chemical by TIMES (Dimitrov et al., 2005b), and 1,6-diisocyanatohexane, methylisoeugenol, 4-methylcatechol, diphenylmethane-4,4′-diisocyanate, and 4-nitrobenzyl chloride were discarded because they were considered as sensitisers or respiratory sensitisers in the sources but had no associated LLNA EC3 values, which was interpreted as an indication of lower quality data. The remaining 269 chemicals were used for modelling.

The dataset can be found in the Supporting Information (SI_Dataset.xls) and contains: name, SMILES, human skin sensitisation classification (1 to 6 categories), NOEL values (μg/cm^2) (Basketter et al., 2014), human GHS derived classifications (1A, 1B, NS), the LLNA EC3 values obtained from the different sources with a corresponding final call made by the authors for those cases in which multiple LLNA studies were available for the same chemical, and the *in chemico* and *in vitro* readouts that are explained in the next section. Binary descriptors indicating positive or negative predictions for each of the methods and the LLNA skin sensitisation hazard are also included in the dataset. In addition, the values of DRAGON and TIMES-SS descriptors used in the consensus model, a column indicating the use given to each chemical for each tree (*i.e.* training set, test set, or external test set), and the final consensus model predictions with the corresponding qualitative confidence measures are reported.

2.2. In chemico and in vitro data

The non-animal data included in the dataset were those generated with the three validated and OECD adopted methods, *i.e.* DPRA, KeratinoSens™, and h-CLAT, and were obtained from the validation study reports (EURL-ECVAM, 2012, 2015; EURL-ECVAM, 2014) and the scientific literature (Bauch et al., 2012; Emter et al., 2010; Gerberick et al., 2004, 2007; Natsch and Emter, 2008; Natsch et al., 2013; Nukada et al., 2013; Takenouchi et al., 2013).

The DPRA (OECD, 2015a) is an *in chemico* method which addresses peptide reactivity, considered to be the Molecular Initiating Event (MIE) or Key Event (KE)-1 in the skin sensitisation AOP (OECD, 2012a), by measuring the depletion of synthetic heptapeptides containing either cysteine or lysine following 24 hour incubation with a single concentration of the test substance. Depletion of the peptide in the reaction mixture is measured by HPLC using UV detection. Average peptide depletion data for cysteine and lysine are interpreted using a classification model in which chemicals classified as having minimal reactivity are considered to lack skin sensitisation potential whereas chemicals classified as having low, moderate, or high reactivity are considered to be skin sensitisers. DPRA data included in the dataset were: a) the % cysteine and b) the % lysine depletion values, c) average of cysteine and lysine depletion values, d) the DPRA positive or negative prediction, and

e) the reactivity class (minimal, low, moderate and high) assigned to the substance.

KeratinoSens™ (OECD, 2015b) is a luciferase reporter gene assay in which quantification of the luciferase gene is used as an indicator of the activity of the Nrf2 transcription factor in keratinocytes following 48 hour exposure to twelve serial concentrations of the test substance. By measuring activation of a relevant pathway in keratinocytes, the KeratinoSens™ is addressing KE-2 of the skin sensitisation AOP. KeratinoSens™ readouts included in the dataset were: a) the IC50 value, *i.e.* the concentration of test chemical yielding 50% reduction in cell viability; b) the EC 1.5, EC 2.0 and the EC 3.0 values, *i.e.* the extrapolated concentration of test chemical inducing the luciferase activity above the 1.5-fold, 2-fold, and 3-fold thresholds, respectively; c) the I_{max} value, *i.e.* the maximal fold induction of the luciferase activity over the negative control (solvent); and d) the KeratinoSens™ positive or negative prediction.

The h-CLAT test method addresses KE-3 of the skin sensitisation AOP by quantifying changes in the expression of cell surface markers associated with the process of activation of monocytes and dendritic cells (DC), *i.e.* CD86 and CD54, in the human monocytic leukaemia cell line THP-1 following 24 hour exposure to eight serial concentrations of the test chemical. The h-CLAT readouts included in the dataset were: a) the CV75, *i.e.* the test chemical concentration resulting in 75% cell viability compared to the solvent/vehicle control; b) two binary descriptors indicating whether the cell surface markers (CD86 and CD54) were expressed; c) the EC150 and EC200 values corresponding to the concentrations at which the test chemicals induces a relative fluorescence intensity (RFI) equal or above 150% for the CD86 and equal or above 200% for the CD54; and d) the h-CLAT positive or negative prediction.

Data from *in chemico* and *in vitro* tests were double checked and updated with the latest results provided by the test developers. When discrepancies between different sources of data for the same chemical were found, the test developers were contacted for clarification and a final call was made on the basis of the quality of data and the precautionary principle. When data from validation studies was available for a chemical, it was preferred over any other source since it is considered of higher quality because it was generated under blind testing conditions.

A number of chemicals were tested in the *in chemico* and *in vitro* methods with slightly different formulations, *e.g.* sulphate *vs.* disulphate, aldehyde *vs.* hydro aldehyde, sulphate heptahydrate *vs.* sulphate, racemic mixtures *vs.* pure enantiomers, *etc.* These variations in the formulations are not expected to affect chemical reactivity or hazard classification because they do not represent modifications of the reactive part of the molecule. These formulations were merged into single entries in order to have a larger number of chemicals with complete data, *i.e.* data from all *in chemico* and *in vitro* methods. In some cases, chemicals with the same name and different CAS numbers were found in different sources. These data were also merged into single entries. Table 1 shows the chemicals that were found in different sources with different formulation, name, and/or CAS number; and that were merged into single entries.

Table 1
List of chemicals with data from different sources that were merged into single entries.

CAS from sources	Names from sources	Data merged into
1405-10-3	Neomycin	Neomycin
1404-04-2	Neomycin sulphate	
59-01-8	Kanamycin (*Streptomyces*)	Kanamycin
64013-70-3	Kanamycin monosulphate	
8063-07-8	Kanamycin disulphate	
19317-11-4	Farnesal (mixture)	Farnesal
502-67-0	Farnesal (specific isomer)	
69-57-8	Penicillin G sodium salt	Penicillin G
61-33-6	Penicillin G (free acid)	
61-33-6	Penicillin Potassium salt	
104-55-2	Cinnamic aldehyde	Cinnamic aldehyde
14371-10-9	Cinnamaldehyde-hydro	
127-65-1	Chloramine T	Chloramine T
149358-73-6	Chloramine T	
10191-41-0	(±)-α-Tocopherol	Tocopherol
59-02-9	Tocopherol	
18031-40-8	Perillaldehyde	Perillaldehyde
2111-75-3	Perillaldehyde	
5989-27-5	D-Limonene (pure isomer)	D-Limonene
138-86-3	Limonene (racemic mixture)	
99-49-0	S-Carvone	Carvone
6485-40-1	R-Carvone	
69-09-0	Chlorpromazine hydrochloride	Chlorpromazine
50-53-3	Chlorpromazine	

2.3. In silico descriptors

The dataset of 269 chemicals was complemented with *in silico* descriptors generated with various software packages: the OECD QSAR ToolBox (OECD, 2013), Derek Nexus (LHASA, 2015), Toxtree (Ideaconsult Ltd. on behalf of the JRC, 2005), Dragon (Talete Srl, 2010), Vega (Istituto di Ricerche Farmacologiche Mario Negri, 2013), TIMES (Dimitrov et al., 2005b), and ADMET Predictor (Simulations Plus, Inc., Lancaster, 2014).

- *OECD Toolbox* (*v.* 3.2): the results of the chemical profilers were transformed into binary descriptors indicating the presence or absence of each alert of each profile. All profilers were used. No metabolism was considered when processing the chemicals.
- *Vega* (*v.* 1.0.8): all Vega modules (*e.g.* mutagenicity, skin sensitisation, *Daphnia magna* LC50, BCF Read-Across, and LogP) were run and the results were used as descriptors.
- *Toxtree* (*v.* 2.6.6): all modules (*e.g.* Ames mutagenicity, carcinogenicity, Cyp450, DNA binding, skin sensitisation) were run and the results were used as descriptors.
- *Derek Nexus* (*v.* 1.7.6 (4.0.6)): the predictions obtained for skin and respiratory sensitisation, skin and eye irritation, and photoallergenicity for human, mammal, mouse, and rat were used as descriptors.
- *ADMET* (*v.* 7.1): the results obtained from the physicochemical, biopharmaceutical, and toxicity modules were used as descriptors.
- *TIMES-SS* (*v. 2.27.13*): the results of models of different endpoints like *Daphnia magna* 24 h EC50, *Vibrio fischeri* 5 min, phototoxicity, estrogen receptor, Ames mutagenicity, and skin sensitisation, were used as descriptors. The last three endpoints also included the metabolism simulator with autoxidation. All the predictions were carried out using the default setting that includes a conformer optimisation step.
- *Dragon* (*v.* 6.0.7): All descriptors were calculated and the highly correlated (>0.95) and invariant ones were filtered out.

The descriptors listed above were calculated from the SMILES codes of the chemicals, which were preferably obtained from the data sources, and when not provided, from Chemspider (www.chemspider.com), Sigma-Aldrich, or PubChem by using as queries the names and CAS numbers found in the sources. The 3D structures needed by Dragon to calculate 3D descriptors were generated from the SMILES codes of the chemicals using the Open Babel (O'Boyle et al., 2011) node implemented in KNIME (Berthold et al., 2007). The SMILES codes that were used for the calculations did not include salts as they were stripped. A final dataset of 269 chemicals with about 4500 descriptors for each chemical was obtained by combining the *in silico* descriptors with the *in chemico* and *in vitro* data.

2.4. Subsets of data

Due to the different number of chemicals with available *in chemico* and/or *in vitro* data, five different subsets of the entire dataset were used for modelling. The first subset, "complete subset", consisted of 127 chemicals with DPRA, KeratinoSens™, h-CLAT data, and *in silico*

predictions. The second subset, "DPRA subset", consisted of 167 chemicals with DPRA data and *in silico* predictions. The third subset, "KeratinoSens™ subset", consisted of 223 chemicals with KeratinoSens ™ data and *in silico* predictions. The fourth subset, "h-CLAT subset", consisted of 168 chemicals with h-CLAT data and *in silico* predictions. The fifth subset, "*in silico* subset" consisted of 269 chemicals with *in silico* predictions only. The training and test sets of each subset were generated using the procedure explained next.

2.5. Data split

80% of the skin sensitisation subsets were used as training sets, or modelling sets, and 20% as external test sets, or validation sets. The chemicals were structurally diversely picked in order to maximise the overlap between the two sets and increase the chances that the models obtained with the training set could be successfully applied to the validation set (Tong et al., 2003). The procedure was carried out separately for sensitisers and non-sensitisers to have a better representation of both groups. The proportion of sensitisers and non-sensitisers was kept as given in the subsets to maximise the number of chemicals used for modelling and to improve the prediction of sensitisers, *i.e.* reduce the number of false negatives (FN) as sensitisers were represented in larger proportions in the subsets. In practice, the training and test sets were defined as depicted in Fig. 1.

The chemicals were first divided between sensitisers (S) and non-sensitisers (NS) according to their LLNA results. Fingerprints accounting for structural descriptors were calculated using the RDKit fingerprints node (Landrum, 2015) in a KNIME workflow. A principal component analysis (PCA) on the Tanimoto similarity matrix was carried out and the components accounting for 90% of the variability were kept. This step was carried separately for S and NS. The remaining components were used to cluster the chemicals using the k-Means algorithm implemented in KNIME in as many clusters as the number of chemicals divided by 10 and 5 for sensitisers and non-sensitisers, respectively. In the final step, 80% of the chemicals of each cluster were assigned to the training set using a stratified random selection algorithm, and the rest of chemicals were assigned to the test set (Fig. 1).

2.6. Generation of classification trees (CT)

Machine learning methods generate better performing models when datasets with smaller number of descriptors are used (Witten and Frank, 2005). In order to reduce the number of descriptors, *i.e.* reduce the dataset dimensionality, feature selection methods like classifier subset evaluation, correlation based subset evaluation (Hall, 1998), or information gain attribute evaluation, combined with different search algorithms like genetic algorithm (Goldberg, 1989), ranking and best first, were applied to the training sets. Various lower-dimensional sets of the subsets of data mentioned above were obtained after applying different feature selection methods, and these lower-dimensional sets were used to build CTs. All CTs were generated with the Weka (Hall et al., 2009) node implemented in KNIME using the *C*4.5 algorithm (Salzberg, 1994). The trees were generated using the default parameters except the minimum number of components per leaf, which was increased from 2 to 8 for the *in silico* dataset and to 5 for the other datasets containing *in vitro*/*in chemico* descriptors.

3. Results

3.1. Most discriminating descriptor

All CTs obtained with the various subsets of data selected the TIMES-SS protein adduct formation descriptor (hereafter TIMES-ProtBind) as first and most discriminating descriptor. This descriptor represents a prediction of the amount (in moles) of test chemical that would covalently bind to a mole of protein and is named "AmountAduct/mol/" in TIMES-SS outputs. TIMES-ProtBind not only considers the parent chemical reactions but also the reactions of its possible metabolites and autoxidation products, which are generated by the skin metabolism simulator implemented in TIMES-SS. Thus, TIMES-ProtBind addresses the MIE of the AOP and accounts for both biotic and abiotic transformations. Surprisingly, other descriptors that also account for the MIE and that were present in the datasets, *i.e.* DPRA cysteine and/or lysine depletion values or even KeratinoSens™ readouts, were never selected as first nodes if TIMES-ProtBind was present in the training set. This is due to the fact that the algorithm that we used to build the classification trees (*C*4.5) selects the descriptor with the highest information gain ratio as first node. This descriptor corresponds to TIMES-ProtBind for all datasets as is shown in Figs. SI4–SI8.

In order to investigate further the reason why TIMES-ProtBind was always preferred to *in chemico* or *in vitro* readouts as first decision node in the CTs, the predictions of TIMES-ProtBind, DPRA, and KeratinoSens™ were compared and the results are presented in Fig. 2.

Pie chart B informs about the discrepancies in predictions between DPRA and TIMES-ProtBind by representing chemicals (n = 26) predicted as negative by the DPRA and positive by TIMES-ProtBind. The largest proportion of these chemicals are nevertheless positive in the LLNA (18 in the 1B and 3 in the 1A skin sensitisation GHS subcategories, pink and red sectors of the pie, respectively) and are therefore false negative (FNs) in the DPRA but are correctly predicted as true positives (TPs) by TIMES-ProtBind. The green sector of the pie represents LLNA negative chemicals (n = 5) which are correctly predicted by DPRA, and are false positives (FPs) in TIMES-ProtBind. Pie chart D provides the same kind of information when comparing KeratinoSens™ with TIMES-ProtBind results. Pie charts E and G represent chemicals the vast majority of which are correctly predicted as TNs by TIMES-ProtBind but are FPs in both DPRA (pie chart E) and KeratinoSens™ (pie chart G). The last row of Fig. 2 shows the results of TIMES-ProtBind predictions for the chemicals for which DPRA (pie charts I and J) and KeratinoSens™ (pie charts K and L) data was not available.

In general it is observed that TIMES-ProtBind correctly predicts around 80% of the chemicals that are mispredicted by the *in chemico*/

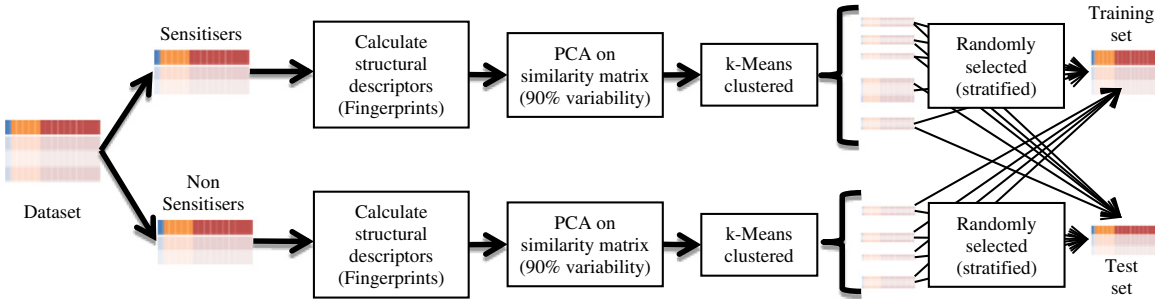

Fig. 1. Data split procedure applied to the skin sensitisation subsets of data. The similarity matrix was calculated using Tanimoto's distance. All algorithms correspond to the ones implemented in KNIME.

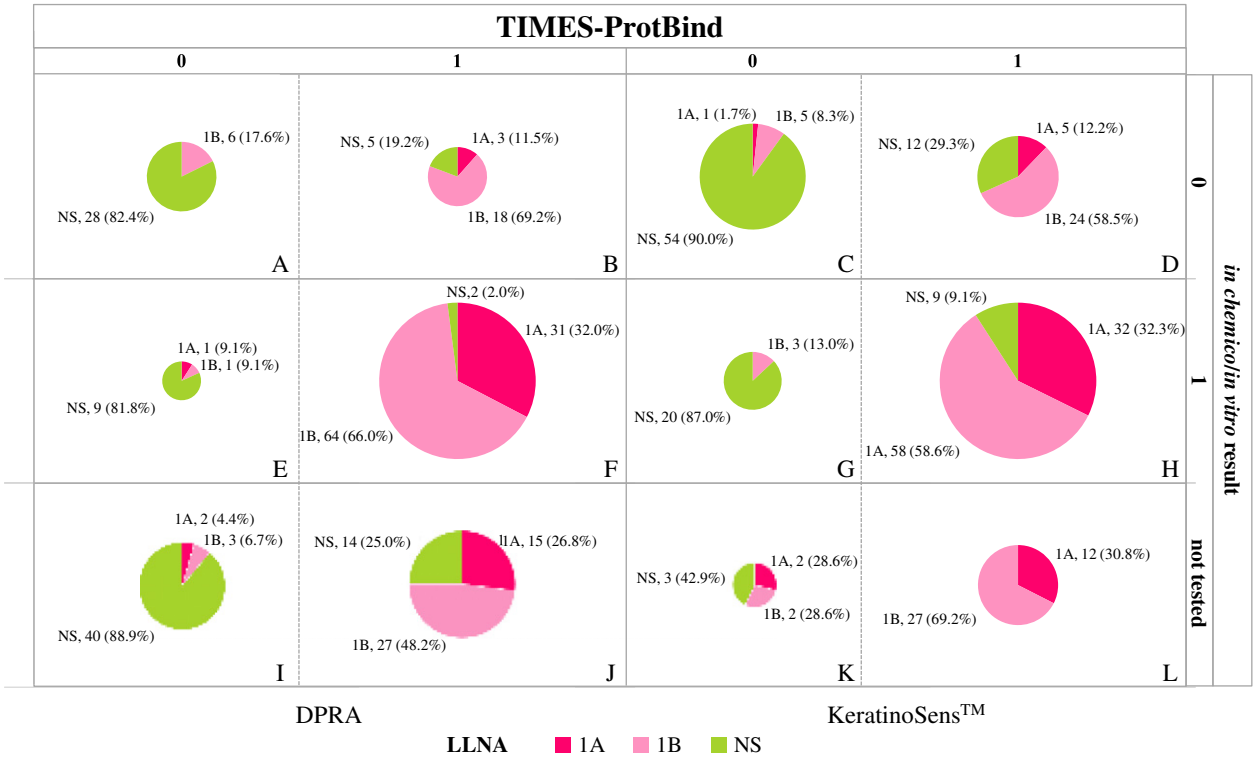

Fig. 2. Pie charts comparing the predictions of TIMES-ProtBind *vs.* DPRA and KeratinoSens™ when predicting LLNA skin sensitisation results. The colours of the pie charts group chemicals by their LLNA results: green for non-sensitisers, pink for Cat 1B sensitisers, and red for Cat 1A sensitisers. The sizes of the pie charts are proportional to the number of chemicals they contain. Each pie chart corresponds to chemicals binned by the DPRA and KeratinoSens™ skin sensitisation predictions (rows) vs. TIMES-ProtBind prediction (columns). Negative predictions are indicated with 0 whereas positive predictions are indicated with 1. As an example, the top left pie chart (A) corresponds to the chemicals that are predicted to be non-sensitisers by DPRA and that are predicted not to bind to proteins by TIMES-ProtBind. The larger sector of the pie shows that 82% (n = 28) of the chemicals are negative in the LLNA (represented in green in the pies) and, therefore, are correctly predicted as true negatives (TNs) by DPRA and TIMES-ProtBind. The smaller sector of the pie corresponds to chemicals (n = 6) with positive LLNA classification (Cat 1B, represented in pink in the pies) which are FNs in both DPRA and TIMES-ProtBind. The same kind of information for KeratinoSens™ is given by pie chart C.

in vitro methods. Pie charts comparing the predictive performance of TIMES-ProtBind and h-CLAT are not shown since the analysis and values are very similar to the DPRA ones. However, they can be found in Fig. SI1 of the Supporting Information.

3.2. Consensus model

In order to obtain a model for skin sensitisation hazard that provides the safest possible results for the consumers, a conservative consensus model constituted by two classification trees, *i.e.* a decision forest of two trees (Tong et al., 2003), was built. This kind of model is characterised by having a lower number of FNs than the individual trees that constitute them since a chemical is classified as non-sensitiser only when both individual trees provide a negative prediction for that chemical. For any other combination of results from the individual trees, *i.e.* discordant predictions or concordant positive predictions, the chemical is predicted as sensitiser.

The C4.5 algorithm was used on the lower-dimensional sets of data that were obtained from the complete, DPRA, KeratinoSens™, h-CLAT, and *in silico* subsets to generate a number of trees from which to build the consensus model. Only trees with sensitivities and specificities above 0.80 were considered as potential candidates for the consensus model. The two classification trees that were selected to form the consensus model were those with the highest specificity and accuracy, and are shown in Fig. 3.

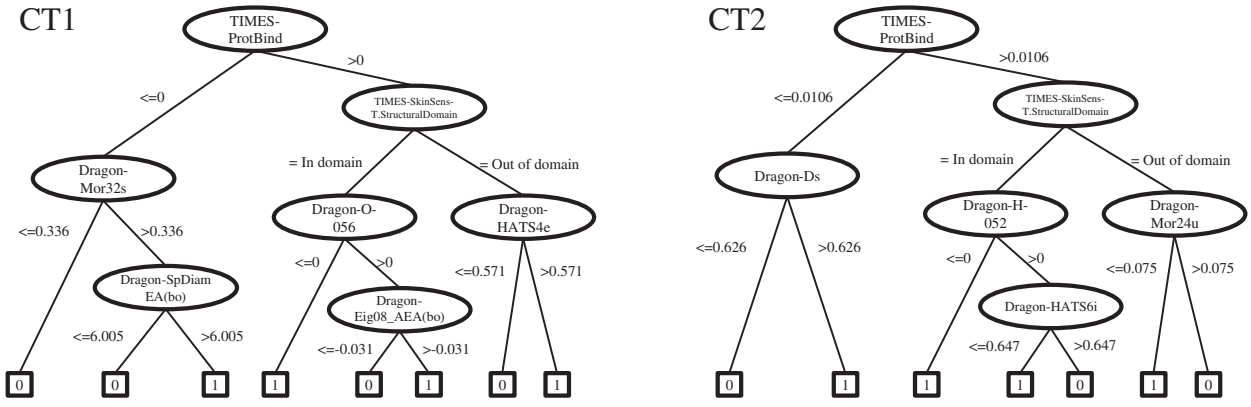

Fig. 3. Classification trees with the highest specificity and accuracy that were selected to build the conservative consensus model.

The selected classification trees, named CT1 and CT2, are based solely on TIMES-SS and Dragon descriptors and have similar shapes. Other trees with different types of descriptors including Vega and KeratinoSens™ IC50 were generated but their predictive performances were lower. CT-1 and CT-2 share the first node, and also share the second node that separates protein binders that fall inside and outside the total skin sensitisation structural applicability domain of TIMES-SS (hereafter TIMES-SkinSens-T.StructDomain). The descriptors used by the consensus model are shown in Table 2 together with an interpretation of their function in the classification trees.

TIMES-ProtBind is the predicted amount of test chemical that would bind to skin protein. This quantity (mol of test chemical/mol of protein) is obtained by the software in two simultaneous steps. The first step is to determine the metabolites and autoxidation products that the parent chemical can be transformed into. This is done by the metabolism simulator which submits the test chemical and each of the predicted metabolites to a set of predefined possible chemical transformations that include spontaneous reactions and enzyme catalysed biotransformation reactions (phase I and II). In the second step, each of the generated metabolites/products is given a probability to covalently bind to proteins. The covalent reactions to proteins are described by 47 alerting groups and are in accordance with the existing knowledge on electrophilic mechanisms underlying skin sensitisation (Aptula et al., 2005; Enoch et al., 2011). The ultimate probabilities of each reaction to take place are determined by a combination of sub-models and 3D QSARs that take into account the nature of the reactive group, sub-structural features, steric effects, molecular size, shape, and lipophilicity among other parameters (Dimitrov et al., 2005b). Thus, TIMES-SS ProtBind ("TIMES AmountAduct/mol/" in the software output) accounts for both the test chemical and its metabolites/products that covalently bind to proteins.

TIMES Skin Sensitisation Total Structural Domain (TIMES-SkinSens-T.StructDomain) is a binary descriptor provided by TIMES-SS determining whether the test chemicals fall in the structural applicability domain of TIMES-SS. This descriptor is the result of the combination of three measures:

a) percentage of atom-centred fragments of test chemical corresponding to fragments extracted only from training set chemicals correctly predicted by the model
b) percentage of atom-centred fragments of test chemical corresponding to fragments extracted only from training set chemicals incorrectly predicted by the model
c) percentage of atom-centred fragments of test chemical not found in training set chemicals of the model

TIMES-SkinSens-T.StructDomain only considers test chemicals "In domain" if 100% of the atom-centred fragments of the test chemical correspond to fragments found in training set chemicals correctly predicted by the model, *i.e.* if all the fragments fall in measure a) above. Test chemicals with any other combination are considered "Out of domain".

The other descriptors used in the consensus model correspond to Dragon descriptors (see Table 2). Of these, O-056 and H-052 are easily interpretable. They are used as binary descriptors indicating the presence of OH– groups and the presence of a hydrogen atom attached to a sp3 carbon with a heteroatom attached to the next carbon atom, respectively. The rest of the descriptors account for structural features and are derived from molecular influence matrixes, spectra, edge adjacent matrix, or the Cartesian coordinates of the atoms. Some of them are weighted by electronic properties of the molecules like electronegativity, ionization potential, or levels of electronic states. Of the other

Table 2
List of descriptors used by the consensus model with their definition and observed function in the individual trees. The tree in which each of the descriptors is used is indicated in brackets after the name of the descriptor.

Descriptor	Description	Chemical/functional interpretation
TIMES-ProtBind (CT-1 & CT-2)	Prediction of the amount of test chemical – either parent chemical or any of the predicted metabolites or autoxidation products – that will bind covalently to a mole of skin protein	Distinguishes protein binders from non-binders
TIMES-SkinSens-T.StructDomain (CT-1 & CT-2)	It determines whether the test chemical falls in the structural applicability domain of TIMES-SS	Distinguishes chemicals whose all atom-centred fragments are found within the correctly predicted chemicals of the training set of TIMES, from others
Mor32s (CT-1) and Mor24u (CT-2)	3D MoRSE descriptors (3D Molecule Representation of Structures based on Electron diffraction) are derived from infrared spectra simulation using a generalised scattering function. Mor32s corresponds to signal 32 weighted by I-state, and Mor24u to the un-weighted signal 24.	Distinguish large and long molecules like Kanamycin or sodium lauryl sulphate from phenol and benzoate like molecules
SpDiam_EA(bo) (CT-1) and Eig08_AEA(bo) (CT-1)	Spectral diameter from edge adjacency matrix weighted by bond order Eighth eigenvalue of the augmented edge adjacency matrix weighted by bond order	Distinguish molecules with multiple carboxylic bonds and esters with long aliphatic chains from other molecules with no double bonds or with shorter side chains
O-056 (CT-1)	Presence of alcohol (—OH) groups	Distinguishes molecules containing alcohol groups
HATS4e (CT-1) and HATS6i (CT-2)	The GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors are molecular descriptors derived from the Molecular Influence Matrix (MIM). HATS4e is a leverage-weighted autocorrelation of lag 4 weighted by Sanderson electronegativity, and HATS6i is a leverage-weighted autocorrelation of lag 6 weighted by ionization potential	Distinguish molecules that contain highly electrophilic groups like cyano-, nitro(so)-, or halo-substituents, from others with less electrophilic groups like alcohol and amines. HATS6i distinguishes acrylates and sulphates from other molecules
Ds (CT-2)	D total accessibility index weighted by I-state. It is built in such a way as to capture relevant molecular 3D information regarding the molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. It increases when the variability of the distribution of the electrotopological charges is relevant. Indicates the presence of electronegative groups at one end of the molecule (Todeschini and Gramatica, 1997; Todeschini et al., 1994).	Distinguishes nitro benzenes from other molecules (predicted non-reactive to proteins)
H-052 (CT-2)	H attached to C(sp3) with 1 heteroatom attached to the next C	Distinguishes saturated molecules from unsaturated

descriptors, Ds is a directional weighted holistic invariant molecular descriptor (Todeschini and Gramatica, 1997) and is larger for molecules that have regions of high density of mass or electronegativity. Thus, it can indicate the presence of nitro, tri-fluoro, carbonyl, or halogen groups at one end of a molecule. In our model, Ds separates nitro benzenes (that were predicted non-reactive to proteins) from other molecules. Mor32s and Mor24u are 3D molecule representations of structures based on electron diffraction. Their interpretation is not straightforward and in general indicate the presence of pairs of atoms beyond a certain distance (Devinyak et al., 2014). In our model, they seem to separate large and long molecules like Kanamycin or sodium lauryl sulphate from phenol and benzoate like molecules, all of them predicted non-reactive to proteins. HATS4e and HATS6i are geometry, topology, and atom-weights assembly (GETAWAY) descriptors (Consonni et al., 2002a). These descriptors try to match 3D molecular geometry with chemical information by using Sanderson electronegativity and ionization potential as atomic weightings, respectively. At the practical level, they encode local information related to molecular fragments and substituent groups (Consonni et al., 2002b). In our model they separate molecules that contain highly electrophilic groups like cyano-, nitro(so)-, or halo-substituents, from less electrophilic groups like alcohol and amines. HATS6i determines the skin sensitisation prediction for many acrylates and sulphates too. SpDiam_EA(bo) and Eig08_AEA(bo) are descriptors derived from graph-theory matrices (Janežič et al., 2007), which are molecular graphs accounting for the 2D structure of molecules by indicating the atoms or chemical bonds that are adjacent to each other (adjacency matrices). SpDiam and Eig08 are properties of such matrices. SpDiam stands for spectral diameter and is calculated as the difference between the largest and smallest matrix eigenvalue of the matrix, and Eig08 is the eighth eigenvalue of the matrix. SpDiam_EA(bo) is, thus, the spectral diameter of the edge adjacency matrix, *i.e.* the diameter of the matrix that indicates the adjacency of edges (chemical bonds) of a molecule, weighted by the bond order. Similarly, Eig08_AEA(bo) is the eighth eigenvalue of the augmented edge adjacency matrix weighted by bond order. The augmented edge adjacency matrix is a variation of the adjacency matrix that is used to differentiate between bond types by giving them different weights. In our model, these descriptors separate molecules with multiple carboxylic bonds and esters with long aliphatic chains from other molecules with no double bonds or shorter chains.

3.3. Applicability domain

The applicability domain (AD) of the consensus model is determined by the applicability domain of the individual trees and is shown in Figs. SI2 and SI3 of the Supporting Information. The AD of each tree was determined from the correctly predicted chemicals of the corresponding training sets (Tong et al., 2005) as the range of values expanded by the descriptors used in CT-1 and CT-2 ± 15%. Following this strategy, the AD of TIMES-ProtBind was set to values between 0 and 2.30. This descriptor is a quantitative value indicating the moles of test chemical capable of binding to proteins and, in our model, is used as a binary descriptor indicating whether a molecule will bind to proteins. Thus, strictly speaking this descriptor should have no upper limit as the model will also correctly predict chemicals with more reactive groups. For obvious reasons, such a descriptor cannot have negative values. A similar situation is found for Dragon-O-056 and Dragon-H-052 as they account for the presence of structural features.

TIMES predictions are also subject to their own applicability domain (Dimitrov et al., 2005a), which is derived from 3 sub-applicability domains, *i.e.* general parametric domain, structural domain, and interpolation space domain. In the development of the consensus model all the predictions provided by TIMES were considered valid as suggested by the developers (Jaworska et al., 2015), but each AD was included in the pool of descriptors. TIMES-SkinSens-T.StructDomain was finally selected by the algorithm to form CT-1 and CT-2, and obviously its values are limited to "In domain" and "Out of domain".

Three chemicals, Dextran, dimethyl formamide, and fluorescein-5-isothiocyanate seem to fall outside the AD of CT-1 in Fig. SI2 (Supporting Information). This is the case for the global domain but not for the focused domain, *i.e.* the domain defined by only those descriptors used to derive the prediction of these chemicals (Tong et al., 2005). Dextran falls outside the AD of Dragon-O-056, but its prediction is derived from Dragon-HATS4e. Dimethyl formamide falls outside the AD of Dragon-HATS4e but is predicted through Dragon-Mor32s and Dragon-SpDiam_EA(bo). Similarly, fluorescein-5-isothiocyanate falls outside the AD of Dragon-Mor32s but its prediction is derived from Dragon-O-056. Thus, the predictions of chemicals falling outside the global domain should not be affected if they fall in the focused domain.

3.4. Predictive performance of the consensus model

The predictive performances of the consensus model and of the individual trees are shown in Table 3. Common Cooper statistic measures (*i.e.* accuracy, sensitivity, and specificity) and the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for the training and test sets are provided (Cooper et al., 1979). Balanced performance values (prevalence independent) are not reported as they are very similar to the Cooper ones.

The accuracy and sensitivity of the consensus model for predicting LLNA classifications (positive/negative) for a total of 269 chemicals are 0.93 and 0.98, respectively. The high sensitivity achieved is the result of using the conservative consensus approach, which boosts sensitivity at expense of specificity, which is reduced to 0.85 with respect to those of the individual trees (Table 3). Even though the objective of this work was not to predict human skin sensitisation, the performance of the consensus model that results from using the human sensitisation potential present in our dataset (Basketter et al., 2014) as reference was added for comparison. The performance of the LLNA in predicting human skin sensitisation was also added for comparison. Both the consensus model and LLNA predict human skin sensitisation potential almost identically, with an accuracy of 0.80–0.81, sensitivity of 0.90–0.92, and specificity 0.58–0.64 for a total of 99 chemicals. Regarding the performance of the individual trees that constitute the consensus

Table 3
Predictive performance (Cooper statistics) of CT-1 and CT-2 *vs.* LLNA, and of the consensus model *vs.* LLNA and human data (Basketter et al., 2014). Performances for the training and test sets are shown separately. CT-1 and CT-2 have different number of chemicals as they were derived from different subsets of data. CT-1 was derived from the KeratinoSens™ subset (223 chemicals) and CT-2 from the *in silico* subset (269 chemicals). External Test Set consists of the chemicals that were not tested with KeratinoSens™.

Model	Subset	TP	TN	FP	FN	Accuracy	Sensitivity (TP/TP + FN)	Specificity (TN / TN + FP)
Consensus model	*vs.* LLNA	166	84	15	4	0.93	0.98	0.85
	vs. Human	57	23	13	6	0.81	0.90	0.64
CT1	Training (80%)	99	71	4	2	0.97	0.98	0.95
	Test (20%)	23	16	3	4	0.85	0.85	0.84
	External test	34	5	0	8	0.83	0.81	1.00
CT2	Training (80%)	125	72	6	9	0.93	0.93	0.92
	Test (20%)	29	18	3	7	0.82	0.81	0.86
LLNA	*vs.* Human	58	21	15	5	0.80	0.92	0.58

Table 4
List of FN results of the consensus model with respect to the LLNA.

Chemical Name	TIMES-ProtBind	Rationale and notes
Hexyl salicylate	0	No alerts for reactivity. It is an irritant. It has been suggested (Urbisch et al., 2015) as possible FP in the LLNA. Additionally, h-CLAT data shows it is positive for CD54 activation, only.
Pyridine	0	No alerts for reactivity. It has been suggested (Urbisch et al., 2015) as possible FP in the LLNA. Additionally, h-CLAT data shows it is positive for CD86 activation only, with very high CV75 and EC150(CD86).
Xylene	0	No alerts for reactivity. It has been suggested (Urbisch et al., 2015) as possible FP in the LLNA. It is also negative in all *in vitro* methods.
Tocopherol	0.16	Predicted by TIMES to have metabolites that react with proteins. It falls outside Total Structural Domain of TIMES. It is a NS in humans and, therefore, a FP in the LLNA (Urbisch et al., 2015).

Table 5
List of FP predictions of the consensus model with respect to LLNA classifications. The values obtained for TIMES-ProtBind and the individual CT-1 and CT-2 predictions are also included.

Chemical Name	TIMES-ProtBind	CT-1 Prediction	CT-2 Prediction
1-Chloro-2-methyl-3-nitrobenzene	0	0	1
1-Iodohexane	0.93	1	0
1-Methoxy-4-methyl-2-nitrobenzene	0.29	0	1
1-Octen-3-yl acetate	1.44	1	0
2-Mercaptobenzoxazole	0.75	0	1
2-Nitro-3-pyridinol	0	0	1
3-Hydroxy-4-nitrobenzoic acid	0	0	1
Dihydromyrcenol	0.31	1	0
Ethyl benzoylacetate	0.46	1	0
Geranyl nitrile	0.74	1	1
Hydrocortisone	0.24	0	1
N-p-Benzonitrile menthanecarboxamide	0.93	0	1
p-Nitro-benzaldehyde	0.07	1	0
Saccharin	0.58	1	0
Sodium 1-nonanesulfonate	0	0	1

model, *i.e.* CT-1 and CT-2, they show accuracies, sensitivities and specificities above 0.90 for the training sets and above 0.82, 0.81 and 0.84 for the test sets. The performance for the external test set, *i.e.* the chemicals not included in the KeratinoSens™ subset, is also similar but with no FPs. These values show that the performances of the consensus model for the training and test sets of CT-1 and CT-2 are similar and above 0.82, which indicates that the models are robust and not overfitted.

3.5. Mispredictions

Table 3 shows that the consensus model predicts 88 chemicals as negative (non-sensitisers) and that 84 of these correspond to chemicals classified as non-sensitisers by the LLNA. The other 4 predictions are FNs and correspond to hexyl salicylate, pyridine, xylene, and tocopherol. They are listed in Table 4 together with the possible reasons for the misprediction and additional considerations. TIMES-ProtBind values are also listed.

Hexyl salicylate, pyridine, and xylene do not have chemical features that indicate that they are reactive, and they are predicted to be non-reactive to proteins by TIMES. This is supported by the fact that the three chemicals are negative in the DPRA and KeratinoSens™, and xylene is also negative in the h-CLAT assay. Therefore, most certainly neither these compounds nor their metabolites/products are reactive to proteins. Xylene is a well-known FP in the LLNA, hexyl salicylate is a known irritant (which is a confounding factor in the LLNA and may give FPs results), and pyridine is positive for only one of the markers measured in h-CLAT (*i.e.* CD54) but at very high cytotoxic concentration. Thus, evidence suggests that the three chemicals might be FPs in the LLNA. This possibility is in agreement with other works (Basketter et al., 2014; Urbisch et al., 2015).

The other compound, tocopherol, is also a FP in the LLNA and is negative in h-CLAT. However, it is predicted to be reactive to proteins by TIMES although it falls outside the total structural domain of TIMES. Neither Dragon-HATS4e nor Dragon-Mor24u is able to fix the TIMES-ProtBind prediction and render tocopherol as sensitiser. The reason for this is probably the fact that it falls in the "weak part" of the model, which will be explained next.

Table 3 shows that of the 181 positive predictions generated by the consensus model, 166 of them correspond to chemicals classified as sensitisers by the LLNA. Thus, the 15 remaining chemicals are FPs of the consensus model and are listed in Table 5 with TIMES-ProtBind values and with the CT-1 and CT-2 individual predictions. The reason for including the individual predictions of CT-1 and CT-2 is to show that neither is systematically over-predicting sensitisers, but that the misclassifications are attributable to both trees in very similar proportion with 6 and 8 misclassifications, respectively. Of the FPs, only geranyl nitrile is predicted as sensitiser by both trees.

Table 5 shows that most FPs of the consensus model correspond to discordant predictions of CT-1 and CT-2. In fact, 14 out of 15 FPs of the consensus model correspond to discrepancies between CT-1 and CT-2 (see Table 6). However, when the predictions of CT-1 and CT-2 are concordant the probability of being a TP or TN is very high. 145 chemicals in our dataset correspond to CT-1 and CT-2 positive concordant predictions and only 1 of these is a FP. If we would only consider concordant predictions from CT-1 and CT-2 as valid outputs of our model (*i.e.* pure consensus), the predictive performance would be: accuracy = 0.98, sensitivity = 0.97, and specificity = 0.99. However, this would also mean that only 87% of the chemicals of our dataset could be predicted (coverage = 87%) as the rest (13%) would correspond to contradictive outputs of CT-1 and CT-2, which are considered equivocal in pure consensus models.

Besides the discordance between CT-1 and CT-2 predictions, TIMES-SS predictions also play an important role in the mispredictions. Considering the 15 FPs generated by the consensus model, 11 are predicted by TIMES-ProtBind to be reactive to proteins and 8 of these are considered

Table 6
Summary of the results of the consensus model with respect to the LLNA classifications binned by the individual predictions of CT-1 and CT-2, values of TIMES-ProtBind, and TIMES-SkinSens-T.StructDomain.

Binning Result *vs.* LLNA	Concordant CT-1 & CT-2	Discordant CT-1 & CT-2	Positive TIMES-ProtBind & outside TIMES-SkinSens-T.StructDomain	Negative Consensus model & Positive TIMES-ProtBind
TP	144	22	15	–
FP	1	14	8	–
TN	84	–	7	11
FN	4	–	1	1

outside the applicability domain of TIMES. There are 31 chemicals in our dataset that are predicted positive by TIMES-ProtBind and that fall outside the structural skin sensitisation applicability domain of TIMES (see 3rd column of Table 6). A lower predictive performance for these chemicals is observed, *i.e.* only 71% of these 31 chemicals are properly predicted by the consensus model since the predictions turn out to be 7 TN, 15 TP, 8 FP, and 1 FN with respect to the LLNA classifications. The FN prediction corresponds to tocopherol (see Table 4). The fact that the chemicals are predicted to bind to proteins by TIMES and are predicted as non-sensitisers by the consensus model is not a source of concern because this combination takes place for 12 substances and there is only one chemical misclassified, tocopherol (see 4th column of Table 6).

Predictions with concordant results from CT-1 and CT-2 have much higher confidence than those that are not concordant. In order to provide more accurate prediction confidences, we have analysed the consensus model predictions with respect to the combinations of CT-1 and CT-2 in more detail and have assigned to each prediction a qualitative confidence value, *i.e.* very high, high, low, or very low, depending on the combination of leaves used in CT-1 and CT-2. The qualitative confidence measures are based on the percentage of correct predictions obtained by each combination of leaves. The qualitative confidence measures for each combination of CT-1 and CT-2 with the corresponding PPV and NPV obtained for our dataset are provided in Table 7.

Concordant positive predictions of CT-1 and CT-2 have a high rate of TPs, 144 out of 145. Therefore concordant positive predictions of CT-1 and CT-2 are assigned a very high confidence. Table 6 also shows that only 61% of the discordant positive predictions correspond to TPs and that, therefore, discordant positive predictions should correspond to low confidence predictions. Table 7 shows that this is the case for the majority of discordant predictions between CT-1 and CT-2, but there are some combinations for which the rate of TPs is very high. For instance, 4 out of 4 predictions based on positive CT-1 obtained from Dragon-SpDiam_EA(bo), and negative CT-2 obtained from Dragon-Ds, correspond to TPs. Thus, predictions obtained from this combination are considered highly reliable even though they correspond to discordant outputs of CT-1 and CT-2. The combination of positive CT-1 *via* Dragon-HATS4e and negative CT-2 *via* Dragon-Mor24u correctly predicts as sensitiser only 1 out of 4 chemicals. Consequently, predictions obtained with this combination are considered of very low confidence.

The conservative consensus model was built to improve sensitivity and, as shown in Table 6, negative predictions are highly reliable. However, Table 7 shows that the combination of (negative) Dragon-HATS4e and Dragon-Mor24u is not as good as the combination of other descriptors since only 5 out of 6 (~83%) of the chemicals are TNs. Thus, a slightly lower degree of confidence (high confidence instead of very high) was assigned to this combination, which also corresponds to a negative prediction by the consensus model but positive TIMES-ProtBind and outside the structural domain of TIMES (4th column of Table 6).

4. Discussion

A dataset of 269 chemicals with high quality DPRA, KeratinoSens™, and h-CLAT data was built. The number of chemicals in the dataset with *in vitro* data lies in between those reported in Urbisch et al. (Urbisch et al., 2015; n = 202) and Natsch et al. (Natsch et al., 2015; n = 312), which are the largest datasets published to date although the latter does not contain h-CLAT data. Our dataset was complemented with *in silico* descriptors from different software packages (*e.g.* TIMES-SS, ADMET Predictor, Derek Nexus, The OECD Toolbox, Vega, and Dragon). The dataset was used to derive a model for skin sensitisation hazard prediction.

The results obtained with our dataset show that *in silico* descriptors are better predictors of skin sensitisation hazard than *in chemico* and *in vitro* methods/readouts. Out of about 4500 descriptors present in our initial dataset, the most discriminating one was TIMES-ProtBind (see Figs. SI4–SI8 in the Supporting Information), an *in silico* descriptor that accounts for the protein binding of the test chemical and their predicted metabolites and autoxidation products, which are generated by an autoxidation and skin metabolism simulator (Dimitrov et al., 2005b). Other works (Jaworska et al., 2013; Natsch et al., 2009; Patlewicz et al., 2014) have also used TIMES-SS but instead of using its skin sensitisation predictions or alerts (Jaworska et al., 2015; Urbisch et al., 2015), our model uses the amount of hapten-protein formation, named "AmountAduct/mol/" in the software output, to which we have referred throughout the present manuscript as TIMES-ProtBind. The adequacy of such a descriptor is two-fold: it is consistent with the skin sensitisation AOP by addressing the MIE and it provides the consensus model with a highly mechanistic relevance as TIMES-ProtBind is mainly derived from structural alerts that relate to chemical reactions relevant for skin sensitisation that were defined by experts and that are encoded in TIMES-SS. We compared the predictions of TIMES-ProtBind in predicting LLNA classifications with the ones of the *in vitro* and *in chemico* methods (see Fig. 2 and Fig. SI1 in the Supporting Information). It is shown that TIMES-ProtBind correctly predicts about 80% of the chemicals mispredicted by the *in vitro* and *in chemico* methods.

Apart from TIMES-ProtBind, our model also uses TIMES-SkinSens-T.StructDomain as a descriptor. This measure of structural domain of TIMES is very restrictive. In order to be considered in the structural domain of TIMES, a test chemical needs to have 100% of the atom-centred fragments in the set of chemicals of the training set of TIMES that were correctly predicted for skin sensitisation. In fact, the developer recommended other researchers to use TIMES-SS predictions independently of whether the chemical of interest was considered in the domain of TIMES(Jaworska et al., 2015). Thus, given that TIMES-ProtBind has a highly mechanistic character and that the applicability domain measure of TIMES is very restrictive, we considered all TIMES predictions valid and included the different applicability domain measures as descriptors in our dataset.

Table 7
List of qualitative confidence factors assigned to the predictions of the conservative consensus model depending on the leaves used in the individual trees. Only the combinations that occurred for 4 or more chemicals are shown. The corresponding PPV and NPV are also provided.

Consensus Prediction	CT-1 *vs.* CT-2	CT-1		CT-2		Conf.	PPV or NPV (%)
		Leaf	Pred.	Leaf	Pred.		
S	Concordant					Very high	144/145 = 99%
S	Discordant					Low	22/36 = 61%
S	Discordant	Dragon-SpDiam_EA(bo)	1	Dragon-Ds	0	Very high	4/4 = 100%
S	Discordant	Dragon-O-056	1	Dragon-HATS6i	0	Low	5/8 = 63%
S	Discordant	Dragon-HATS4e	0	Dragon-Mor24u	1	Very low	4/8 = 50%
S	Discordant	Dragon-Mor32s	0	Dragon-Ds	1	Very low	3/6 = 50%
S	Discordant	Dragon-HATS4e	1	Dragon-Mor24u	0	Very low	1/4 = 25%
NS	Concordant					Very high	84/88 = 96%
NS	Concordant	Dragon-Mor32s	0	Dragon-Ds	0	Very high	67/70 = 96%
NS	Concordant	Dragon-SpDiam_EA(bo)	0	Dragon-Ds	0	Very high	6/6 = 100%
NS	Concordant	Dragon-Eig08_AEA(bo)	0	Dragon-HATS6i	0	Very high	4/4 = 100%
NS	Concordant	Dragon-HATS4e	0	Dragon-Mor24u	0	High	5/6 = 83%

A concern that could be raised on the use of TIMES-SS as a descriptor in a skin sensitisation prediction model is that the model could be overfitted due to a possible overlap between the training set of TIMES and the dataset. This is not the case for our consensus model. Although about 66% of the chemicals in our database are found in the training set of TIMES, the accuracy of the individual trees of the consensus model in predicting skin sensitisation hazard is similar (~0.90) for the chemicals that belong to the training set of TIMES and those that do not (see Table SI1), indicating that the consensus model is not overfitted by the use of TIMES-SS results as descriptors.

The division of the dataset into smaller subsets of data (*i.e.* complete, DPRA, KeratinoSens™, h-CLAT, and *in silico* subsets) was intended to obtain classification trees composed of *in vitro*/*in chemico* and *in silico* descriptors. Surprisingly, in the very few cases in which such a combination was obtained, the models turned out to have lower specificities and accuracies. This is mainly due to the higher discriminating power of TIMES-ProtBind, and also to the fact that TIMES-ProtBind predictions largely overlap with the predictions of the different *in vitro* methods (see Fig. 2), what prevents them from being included into the tree once TIMES-ProtBind is selected. Classification trees with DPRA, KeratinoSens™, and h-CLAT descriptors were obtained when TIMES-ProtBind was manually removed from the pool of descriptors, but in all cases the Cooper statistics of the resulting classification trees were significantly lower. This shows that if any experimental method is capable of predicting protein binding similarly to TIMES (perhaps including an efficient metabolic system) TIMES-ProtBind could probably be substituted in our model.

We aimed at providing a model with the lowest reasonable number of FNs (highest sensitivity) since, besides the obvious safety and ethical reasons, such a model would have the potential to be used as a screening tool by different stakeholders like regulators and industry. A model that predicts non-sensitisers with high confidence can be used to gain assurance before releasing products to the market, to detect substances of high concern, or even in a 2-tiered approach to predict skin sensitisation potency. We achieved a predictive model with very high sensitivity and accuracy by combining the two classification trees with the highest specificity in a conservative way. This means that the consensus model only predicts a chemical as non-sensitiser if CT-1 and CT-2 have concordant negative predictions. With any other combination of CT-1 and CT-2 the consensus model predicts a chemical to be a sensitiser. This combination is less accurate than a "pure consensus" model (*i.e.* concordant positive and concordant negative predictions of CT-1 and CT-2), but allows us to predict a larger amount of chemicals.

Since neither CT-1 nor CT-2 contain experimental descriptors, our consensus model is not subject to physicochemical limitations like water solubility or evaporation of the test chemical as it is the case for *in vitro*/*in chemico* methods and testing strategies based on their combination. The consensus model allows the prediction of a number of chemicals within a few minutes, it has no inter- or intra-lab variability, and only needs the chemical structure of the test chemical to obtain a prediction. The main limitations of the consensus model are the fact that it uses predictions from licensed software packages, that it can only be used to predict organic chemicals with defined structures (no mixtures), and that some of the descriptors correspond to 3D descriptors whose values depend on the 3D structure of the molecule, which means that some predictions can be affected by the quality of the geometrical optimisation process.

The higher accuracy of the consensus model (accuracy 0.93; n = 269) with respect to that of the validated and regulatory adopted individual methods (accuracies of DPRA (0.77; n = 167), KeratinoSens™ (0.71; n = 223), and h-CLAT (0.77; n = 168)) is in line with the results provided by other defined approaches that use multiple readouts from *in silico* and/or *in chemico* and/or *in vitro* methods (see Table 8).

In order to properly compare the performance of different approaches, the same set of chemicals should ideally be used since the predictive performance values are dependent on the set of chemicals considered. In general, larger datasets will tend to have lower performance statistics as the chemical structural diversity, reaction mechanism and applicability domain will be larger and, therefore, the models will be more general. It is beyond the scope of this analysis to do a comprehensive comparison of available approaches for skin sensitisation, which among other issues should take into account the difference in datasets used to develop and test each of the models. Nevertheless, we made an attempt to benchmark the performance of our model. Table 8 shows the performance statistics for some of the defined approaches for skin sensitisation that have been recently published.

Our model shows an accuracy of 0.93 and has been tested with 269 chemicals. Other approaches available at the moment, *i.e.* Hirota et al. (2015), Natsch et al. (2015), Urbisch et al. (2015), Alves et al. (2015), or Takenouchi et al. (2015), show accuracies ranging from 0.81 to 0.84 except that of Jaworska et al. (2015) which shows an accuracy of 0.96.

Table 8
Summary of results of defined approaches and models published in 2015 for skin sensitisation hazard prediction against LLNA. The number of true negatives (TN), true positives (TP), false positives (FP), false negatives (FN), sensitivity (Sens), specificity (Spec), accuracy (Acc), and total number of chemicals (n) are reported. The values were obtained from the original publications, and the values in parenthesis correspond to other models present in the same publication or to different sets of chemicals used in the publication.

Measure	This work (pure consensus)[a]	Hirota et al. 2015 h-CLAT & DPRA (h-CLAT/SH test/ARE assay)[b]	Takenouchi et al. (2015) ITS (STS)[c]	Urbisch et al. (2015)[d]	Jaworska et al., (2015) Train. Set (Test Set w/AD)[e]	Natsch et al. (2015)[f]	Alves et al. (2015) Set A (Set A & B)[g]
TN	84 (84)	21 (10)	26 (20)	43	36 (14)	84	73 (74)
TP	166 (144)	92 (52)	91 (92)	117	105 (46)	118	73 (123)
FP	15 (1)	16 (11)	11 (17)	15	3 (0)	15	13 (13)
FN	4 (4)	10 (0)	11 (10)	27	3 (0)	27	19 (43)
Sens	0.98 (0.97)	0.90 (1.0)	0.89 (0.91)	0.81	0.97 (1.0)	0.81	0.79 (0.74)
Spec	0.85 (0.99)	0.57 (0.48)	0.70 (0.54)	0.74	0.92 (1.0)	0.85	0.85 (0.85)
Acc	0.93 (0.98)	0.81 (0.85)	0.84 (0.81)	0.79	0.96 (1.0)	0.83	0.82(0.78)
n	269 (233)	139 (73)	139 (139)	202	147 (60)	244	178/254 (253/405)

[a] Data in brackets correspond to the values obtained if discordant results between CT-1 and CT-2 are considered as ambiguous.
[b] Data corresponds to the artificial neural network model (ANN) for potency prediction translated into hazard by joining Cat 1A and Cat 1B into sensitisers. Data for two models are presented, ANN with h-CLAT and DPRA data, and ANN with h-CLAT, SH test, and ARE assay data.
[c] Data corresponds to the integrated testing strategy (ITS) and to the sequential testing strategy (STS) in parenthesis.
[d] Data corresponds to the model that uses LuSens/MUSST values when no h-CLAT data is available.
[e] Data corresponds to the performance of the training set and the test set. The latter set takes into account the applicability domain of the individual data inputs.
[f] Data corresponds to the performance of the model that combines global and local domains.
[g] Data of set A corresponds to chemicals used in the training and test set. Set A is a balanced dataset, and set B corresponds to chemicals that are mostly sensitisers and that were used as external validation set. The model is a pure consensus model and when the two sources are not concordant, the prediction is considered ambiguous and no prediction is given (Alves et al., 2015). This fact is reflected in the number of chemicals, which indicates the amount of chemicals that were predicted from the total of chemicals present in each dataset.

These approaches have been tested with a range of 139 to 244 chemicals. Only the models proposed by Alves et al. have been tested with more than 400 chemicals. However, many of their predictions were considered inconclusive, and thus the approach was in reality tested on 253 chemicals. The coverage of our consensus model would be reduced to 233 chemicals (87%) if a pure consensus was applied. The accuracy, sensitivity, and specificity would, however, be increased to >0.97. Transforming our conservative consensus model to a pure consensus model would affect one of its strong features, which is that it can be applied to a large number of chemicals. In addition, we showed in Table 7 that the fact that CT-1 and CT-2 are discordant is not necessarily an indication of low confidence in the prediction. In any case, when considering the performance of our consensus model and of the other defined approaches, one should bear in mind that these may be biased by the datasets used and the uncertainty associated to the reference data (Dumont et al., 2016).

Even though our model was optimised to predict LLNA classifications, its performance in predicting human responses (Basketter et al., 2014) appears satisfactory with an accuracy of 0.81, sensitivity of 0.90, and specificity of 0.64. These predictive performance values are very similar to those of the LLNA in predicting human data, which in our dataset correspond to an accuracy of 0.80, sensitivity of 0.92, and specificity of 0.58. These values show that our model is capable of predicting human skin sensitisation as accurately as the LLNA does.

The consensus model complies with the 5 OECD principles for the validation of QSARs(OECD, 2004):

1) It predicts a well-defined endpoint that is skin sensitisation hazard (LLNA)
2) It has an unambiguous algorithm as it uses a conservative consensus of classification trees
3) It has a defined applicability domain that consists of any organic chemical with a defined structure whose descriptors fall within the limits shown in Fig. SI2 and SI3 of the Supporting Information
4) It has proper measures of goodness of fit (see Table 3 and Table 8)
5) It has a mechanistic interpretation as it is based on the prediction of protein binding considered to be the MIE of the skin sensitisation AOP. This prediction is performed by TIMES-SS and the result is modified or confirmed by a series of descriptors that mainly account for chemical reactivity features as shown in Table 2

Thanks to the use of classification trees we were able to give qualitative confidence measures to each prediction depending on the combination of descriptors used to generate the prediction (see Table 6 and Table 7). This measure of confidence, although not being a quantitative measure like the one provided by Bayesian models (Jaworska et al., 2013, 2015) or decision forests (Tong et al., 2003), still gives an added value to our model by providing a measure of uncertainty. In fact, predictions of very low or low confidence are an indication that additional evidence needs to be generated to come to a sound conclusion on the lack or presence of skin sensitisation potential.

5. Conclusions

We have built a high quality dataset of 269 chemicals with *in vivo* (Basketter et al., 2014), *in chemico* (Gerberick et al., 2007), and *in vitro* (Ashikaga et al., 2006; Natsch and Emter, 2008) skin sensitisation data. The dataset has been obtained from the literature, test submissions to EURL-ECVAM, and validation studies. The dataset has been completed with *in silico* predictions from several licensed and free software packages (*e.g.* TIMES, Dragon, Vega, Derek Nexus, Dragon, or Toxtree) and has been used to develop a predictive model for skin sensitisation hazard (sensitiser/non-sensitiser). The collected human, LLNA, *in chemico*, *in vitro*, and the *in silico* descriptors used in the model can be found in the Supporting Information (SI_Dataset.xls).

The modelling exercise showed that skin sensitisation hazard, as measured in the LLNA, was better predicted by classification trees based on *in silico* descriptors accounting for reactivity and structural features, being TIMES-ProtBind the most discriminating one. TIMES-ProtBind predicts the amount of test chemical that would bind to proteins and accounts for skin metabolism and autoxidation processes. It addresses the MIE of the skin sensitisation AOP (OECD, 2012a) and is shown to correctly predict about 80% of the chemicals that are mispredicted by the validated methods, DPRA, KeratinoSens™, and h-CLAT.

A conservative consensus model of two classification trees purely based on *in silico* descriptors that predicts skin sensitisation hazard (using LLNA classifications as a reference) with an accuracy of 0.93, sensitivity of 0.98, and specificity of 0.85 for 269 chemicals is provided and analysed in this manuscript. The consensus model can be used to predict organic substances with defined chemical structures (mixtures, inorganic substances and natural products cannot be predicted), and provides a qualitative measure of confidence associated to the hazard prediction. The model is very simple, can be implemented easily in different platforms, and complies with the OECD principles for the validation of QSARs (OECD, 2004).

In summary, we propose a defined approach for predicting skin sensitisation hazard that is highly accurate and sensitive, 100% reproducible, and fast. Depending on the user's acceptance criteria, the predictions generated by our model may be adequate for the intended purpose (*e.g.* hazard classification), or may need to be combined with other information within a WoE approach that yields higher confidence. In applying a WoE approach, it is important to consider that our model is optimised for the reliable identification of negatives, so the "additional weight" should focus on checking the positive predictions (to correct for false positives). In this respect, little added value in predictive performance is likely to be gained by applying an *in chemico* peptide binding model, since this key (molecular initiating) event is well captured by the TIMES- ProtBind descriptors in our model. A better information gain is expected by follow up testing with one of the methods capable of identifying downstream key events (*i.e.* keratinocyte activation, dendritic cell activation).

Our model could also be used in a two-tiered strategy to predict skin sensitisation potency where the first step would be the identification of skin sensitisation potential. In case of a positive (sensitiser) prediction in the first step, the second step would determine whether the sensitising effect is likely to be strong/extreme.

The next steps of our work will focus on the dissemination of the model to make it publicly accessible and on the testing of additional sets of chemicals to further challenge the model.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.tiv.2016.07.014.

Transparency document

The Transparency document associated with this article can be found, in online version.

Acknowledgements

The authors acknowledge Dr. Julien Burton for the fruitful discussions and guidance on the use of Spotfire and thank the Laboratory of Mathematical Chemistry (LMC) and Talete Srl. for giving their permission to publish the data generated with TIMES-SS and Dragon.

References

Alves, V.M., Muratov, E., Fourches, D., Strickland, J., Andrade, C.H., Tropsha, A., Hill, C., 2015. Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. Toxicol. Appl. Pharmacol. 284, 262–272. http://dx.doi.org/10.1016/j.taap.2014.12.014.

Aptula, A.O., Patlewicz, G., Roberts, D.W., 2005. Skin sensitization: reaction mechanistic applicability domains for structure–activity relationships. Chem. Res. Toxicol. 18, 1420–1426. http://dx.doi.org/10.1021/tx050075m.

Ashikaga, T., Yoshida, Y., Hirota, M., Yoneyama, K., Itagaki, H., Sakaguchi, H., Miyazawa, M., Ito, Y., Suzuki, H., Toyoda, H., 2006. Development of an in vitro skin sensitization test using human cell lines: the human Cell Line Activation Test (h-CLAT): I. Optimization of the h-CLAT protocol. Toxicol. in Vitro 20, 767–773. http://dx.doi.org/10.1016/j.tiv.2005.10.012.

Basketter, D.A., Alépée, N., Ashikaga, T., Barroso, J., Gilmour, N., Goebel, C., Hibatallah, J., Hoffmann, S., Kern, P., Martinozzi-Teissier, S., Maxwell, G., Reisinger, K., Sakaguchi, H., Schepky, A., Tailhardat, M., Templier, M., 2014. Categorization of chemicals according to their relative human skin sensitizing potency. Dermatitis 25, 11–21. http://dx.doi.org/10.1097/DER.0000000000000003.

Bauch, C., Kolle, S.N., Ramirez, T., Eltze, T., Fabian, E., Mehling, A., Teubner, W., van Ravenzwaay, B., Landsiedel, R., 2012. Putting the parts together: combining in vitro methods to test for skin sensitizing potentials. Regul. Toxicol. Pharmacol. 63, 489–504. http://dx.doi.org/10.1016/j.yrtph.2012.05.013.

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., 2007. {KNIME}: The {k}onstanz {i}nformation {m}iner. Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007). Springer.

Consonni, V., Todeschini, R., Pavan, M., 2002a. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J. Chem. Inf. Comput. Sci. 42, 682–692. http://dx.doi.org/10.1021/ci015504a.

Consonni, V., Todeschini, R., Pavan, M., Gramatica, P., 2002b. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR S tudies. J. Chem. Inf. Comput. Sci. 42, 693–705. http://dx.doi.org/10.1021/ci0155053.

Cooper, J.A., Saracci, R., Cole, P., 1979. Describing the validity of carcinogen screening tests. Br. J. Cancer 39, 87–89. http://dx.doi.org/10.1038/bjc.1979.10.

Devinyak, O., Havrylyuk, D., Lesyk, R., 2014. 3D-MoRSE descriptors explained. J. Mol. Graph. Model. 54, 194–203. http://dx.doi.org/10.1016/j.jmgm.2014.10.006.

Dimitrov, S., Dimitrova, G., Pavlov, T., Dimitrova, N., Patlewicz, G., Niemela, J., Mekenyan, O., 2005a. A stepwise approach for defining the applicability domain of SAR and QSAR models. J. Chem. Inf. Model. 45, 839–849. http://dx.doi.org/10.1021/ci0500381.

Dimitrov, S., Low, L., Patlewicz, G., Kern, P., Dimitrova, G., Comber, M., Phillips, R., Niemela, J., Bailey, P., Mekenyan, O., 2005b. Skin sensitization: modeling based on skin metabolism simulation and formation of protein conjugates. Int. J. Toxicol. 24, 189–204.

Dumont, C., Barroso, J., Matys, I., Worth, A., Casati, S., 2016. Analysis of the local lymph node assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches. Toxicol. in Vitro 34, 220–228. http://dx.doi.org/10.1016/j.tiv.2016.04.008.

EC, 2006. REACH Regulation (1907/2006/EC): Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, Amending Directive 1999.

EC, 2009. EU Cosmetic Products Regulation (1223/2009/EC).

Emter, R., Ellis, G., Natsch, A., 2010. Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. Toxicol. Appl. Pharmacol. 245, 281–290. http://dx.doi.org/10.1016/j.taap.2010.03.009.

Enoch, S.J., Ellison, C.M., Schultz, T.W., Cronin, M.T.D., 2011. A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity. Crit. Rev. Toxicol. 41, 783–802. http://dx.doi.org/10.3109/10408444.2011.598141.

EURL ECVAM, 2013. Recommendation on the Direct Peptide Reactivity Assay (DPRA) for Skin Sensitisation Testing. http://dx.doi.org/10.2788/48229.

EURL ECVAM, 2014. Recommendation on the KeratinoSensTM Assay for Skin Sensitisation Testing. http://dx.doi.org/10.1017/CBO9781107415324.004.

EURL ECVAM, 2015. Recommendation on the Human Cell Line Activation Test (h-CLAT) for Skin Sensitisation Testing. http://dx.doi.org/10.2788/29986.

EURL-ECVAM, 2012. Direct Peptide Reactivity Assay (DPRA) ECVAM Validation Study Report.

EURL-ECVAM, 2014. KeratinoSens™ Validation Study Reports.

EURL-ECVAM, 2015. h-CLAT Validation Study Reports.

Gerberick, G.F., Vassallo, J.D., Bailey, R.E., Chaney, J.G., Morrall, S.W., Lepoittevin, J.P., 2004. Development of a peptide reactivity assay for screening contact allergens. Toxicol. Sci. 81, 332–343. http://dx.doi.org/10.1093/toxsci/kfh213.

Gerberick, G.F., Vassallo, J.D., Foertsch, L.M., Price, B.B., Chaney, J.G., Lepoittevin, J.P., 2007. Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. Toxicol. Sci. 97, 417–427. http://dx.doi.org/10.1093/toxsci/kfm064.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. first ed. Boston, MA, USA.

Guyard-Nicodème, M., Gerault, E., Platteel, M., Peschard, O., Veron, W., Mondon, P., Pascal, S., Feuilloley, M.G.J., 2015. Development of a multiparametric in vitro model of skin sensitization. J. Appl. Toxicol. 48–58. http://dx.doi.org/10.1002/jat.2986.

Hall, M.A., 1998. Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software. SIGKDD Explor. 11.

Hirota, M., Kouzuki, H., Ashikaga, T., Sono, S., Tsujita, K., Sasa, H., Aiba, S., 2013. Artificial neural network analysis of data from multiple in vitro assays for prediction of skin sensitization potency of chemicals. Toxicol. in Vitro 27, 1233–1246. http://dx.doi.org/10.1016/j.tiv.2013.02.013.

Hirota, M., Fukui, S., Okamoto, K., Kurotani, S., Imai, N., Fujishiro, M., Kyotani, D., Kato, Y., Kasahara, T., Fujita, M., Toyoda, A., Sekiya, D., Watanabe, S., Seto, H., Takenouchi, O., Ashikaga, T., Miyazawa, M., 2015. Evaluation of combinations of in vitro sensitization test descriptors for the artificial neural network-based risk assessment model of skin sensitization. J. Appl. Toxicol. 1333–1347 http://dx.doi.org/10.1002/jat.3105.

Ideaconsult Ltd. on behalf of the JRC, 2005. Toxtree.

Istituto di Ricerche Farmacologiche Mario Negri, 2013. Vega.

Janežič, D., Miličević, A., Nikolić, S., Trinajstić, N., 2007. Graph Theoretical Matrices in Chemistry, Mathemat. third ed. CRC Press.

Jaworska, J., 2011. Integrating non-animal test information into an adaptive testing strategy – skin sensitization proof of concept case. ALTEX 28, 211–225. http://dx.doi.org/10.14573/altex.2011.3.211.

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. Appl. Toxicol. 33, 1353–1364. http://dx.doi.org/10.1002/jat.2869.

Jaworska, J., Andreas, N., Ryan, C., Strickland, J., Takao, A., Masaaki, M., 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. http://dx.doi.org/10.1007/s00204-015-1634-2.

Jowsey, I.R., Basketter, D.A., Westmoreland, C., Kimber, I., 2006. A future approach to measuring relative skin sensitising potency: a proposal. J. Appl. Toxicol. 26, 341–350. http://dx.doi.org/10.1002/jat.1146.

Karlberg, A.T., Bergstrom, M.A., Borje, A., Luthman, K., Nilsson, J.L.G., 2008. Allergic contact dermatitis-formation, structural requirements, and reactivity of skin sensitizers. Chem. Res. Toxicol. 21, 53–69. http://dx.doi.org/10.1021/tx7002239.

Landrum, G., 2015. RDKit: Open-source Informatics.

LHASA, 2015. Nexus.

MacKay, C., Davies, M., Summerfield, V., Maxwell, G., 2013. From pathways to people: applying the adverse outcome pathway (AOP) for skin sensitization to risk assessment. ALTEX 30, 473–486. http://dx.doi.org/10.14573/altex.2013.4.473.

Macmillan, D.S., Canipa, S.J., Chilton, M.L., Williams, R.V., Barber, C.G., 2016. Predicting skin sensitisation using a decision tree integrated testing strategy with an in silico model and in chemico/in vitro assays. Regul. Toxicol. Pharmacol. 76, 30–38. http://dx.doi.org/10.1016/j.yrtph.2016.01.009.

Martin, S.F., 2015. New concepts in cutaneous allergy. Contact Dermatitis 72, 2–10. http://dx.doi.org/10.1111/cod.12311.

Martin, S.F., Esser, P.R., Weber, F.C., Jakob, T., Freudenberg, M.a., Schmidt, M., Goebeler, M., 2011. Mechanisms of chemical-induced innate immunity in allergic contact dermatitis. Allergy 66, 1152–1163. http://dx.doi.org/10.1111/j.1398-9995.2011.02652.x.

Muratov, E.N., Artemenko, A.G., Varlamova, E.V., Polischuk, P.G., Lozitsky, V.P., Fedchuk, A.S., Lozitska, R.L., Gridina, T.L., Koroleva, L.S., Sil'nikov, V.N., Galabov, A.S., Makarov, V.A., Riabova, O.B., Wutzler, P., Schmidtke, M., Kuz'min, V.E., 2010. Per aspera ad astra: application of Simplex QSAR approach in antiviral research. Future Med. Chem. 2, 1205–1226. http://dx.doi.org/10.4155/fmc.10.194.

Natsch, A., Emter, R., 2008. Skin sensitizers induce antioxidant response element dependent genes: application to the in vitro testing of the sensitization potential of chemicals. Toxicol. Sci. 102, 110–119. http://dx.doi.org/10.1093/toxsci/kfm259.

Natsch, A., Emter, R., Ellis, G., 2009. Filling the concept with data: integrating data from different in vitro and in silico assays on skin sensitizers to explore the battery approach for animal-free skin sensitization testing. Toxicol. Sci. 107, 106–121. http://dx.doi.org/10.1093/toxsci/kfn204.

Natsch, A., Ryan, C.A., Foertsch, L., Emter, R., Jaworska, J., Gerberick, F., Kern, P., 2013. A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. J. Appl. Toxicol. 33, 1337–1352. http://dx.doi.org/10.1002/jat.2868.

Natsch, A., Emter, R., Gfeller, H., Haupt, T., Ellis, G., 2015. Predicting skin sensitizer potency based on in vitro data from KeratinoSens and kinetic peptide binding: global versus domain-based assessment. Toxicol. Sci. 143, 319–332. http://dx.doi.org/10.1093/toxsci/kfu229.

Nukada, Y., Miyazawa, M., Kazutoshi, S., Sakaguchi, H., Nishiyama, N., 2013. Data integration of non-animal tests for the development of a test battery to predict the skin sensitizing potential and potency of chemicals. Toxicol. in Vitro 27, 609–618. http://dx.doi.org/10.1016/j.tiv.2012.11.006.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: an open chemical toolbox. J. Chem. Inf. 3, 33. http://dx.doi.org/10.1186/1758-2946-3-33.

OECD, 2004. OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationships Models., OECD.

OECD, 2012a. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Part 1: Scientific Evidence. Ser. Test. Assessment, No. 168, ENV/JM/MONO(2012) 10 PART 1.

OECD, 2012b. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Part 2: Use of the AOP to Develop Chemical Categories and Integrated Assessment and Testing Approaches. Ser. Test. Assessment, No. 168, ENV/JM/MONO(2012) 10 PART 2.

OECD, 2013. OECD QSAR Application Toolbox.

OECD, 2015a. OECD Guideline for Testing of Chemicals - Guideline 422C: In Chemico Skin Sensitisation Direct Peptide Reactivity Assay (DPRA).

OECD, 2015b. OECD Guideline for Testing of Chemicals - Guideline 442D: In Vitro Skin Sensitisation : ARE-Nrf2 Luciferase Test.

OECD, 2016a. OECD Guidance Document on the Reporting of Defined Approaches to be Used Within Integrated Approaches to Testing and Assessment (No. ENV/JM/MONO(2016)28).

OECD, 2016b. OECD Guidance Document on the Reporting of Defined Approaches and Individual Information Sources to be Used Within Integrated Approaches to Testing and Assessment (IATA) for Skin Sensitization (No. ENV/JM/MONO(2016)29).

Patlewicz, G., Kuseva, C., Mehmed, A., Popova, Y., Dimitrova, G., Ellis, G., Hunziker, R., Kern, P., Low, L., Ringeissen, S., Roberts, D., Mekenyan, O., 2014. TIMES-SS – recent refinements resulting from an industrial skin sensitisation consortium. SAR QSAR Environ. Res. 25, 367–391. http://dx.doi.org/10.1080/1062936X.2014.900520.

Sakaguchi, H., Miyazawa, M., Yoshida, Y., Ito, Y., Suzuki, H., 2006. Prediction of preservative sensitization potential using surface marker CD86 and/or CD54 expression on

human cell line, THP-1. Arch. Dermatol. Res. 298, 427–437. http://dx.doi.org/10.1007/s00403-006-0714-9.
Salzberg, S.L., 1994. C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach. Learn. 16, 235–240. http://dx.doi.org/10.1007/BF00993309.
Simulations Plus, Inc., Lancaster, C., 2014. ADMET Predictor.
Strickland, J., Zang, Q., Kleinstreuer, N., Paris, M., Lehmann, D.M., Choksi, N., Matheson, J., Jacobs, A., Lowit, A., Allen, D., Casey, W., 2016. Integrated decision strategies for skin sensitization hazard. J. Appl. Toxicol. n/a–n/a http://dx.doi.org/10.1002/jat.3281.
Takenouchi, O., Miyazawa, M., Saito, K., Ashikaga, T., Sakaguchi, H., 2013. Predictive performance of the human Cell Line Activation Test (h-CLAT) for lipophilic chemicals with high octanol-water partition coefficients. J. Toxicol. Sci. 38, 599–609.
Takenouchi, O., Fukui, S., Okamoto, K., Kurotani, S., Imai, N., Fujishiro, M., Kyotani, D., Kato, Y., Kasahara, T., Fujita, M., Toyoda, A., Sekiya, D., Watanabe, S., Seto, H., Hirota, M., Ashikaga, T., Miyazawa, M., 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. 1318–1332 http://dx.doi.org/10.1002/jat.3127.
Talete Srl, 2010. DRAGON (Software for Molecular Descriptor Calculation).
Teubner, W., Mehling, A., Schuster, P.X., Guth, K., Worth, A., Burton, J., van Ravenzwaay, B., Landsiedel, R., 2013. Computer models versus reality: how well do in silico models currently predict the sensitization potential of a substance. Regul. Toxicol. Pharmacol. 67, 468–485. http://dx.doi.org/10.1016/j.yrtph.2013.09.007.
Todeschini, R., Gramatica, P., 1997. 3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. Quant. Struct. Relationships. 16, pp. 113–119.
Todeschini, R., Lasagni, M., Marengo, E., 1994. New molecular descriptors for 2D and 3D structures. Theory. J. Chemom. 8, 263–272. http://dx.doi.org/10.1002/cem.1180080405.
Tong, W., Hong, H., Fang, H., Xie, Q., Perkins, R., 2003. Decision forest: combining the predictions of multiple independent decision tree models. J. Chem. Inf. Comput. Sci. 43, 525–531. http://dx.doi.org/10.1021/ci020058s.
Tong, W., Hong, H., Xie, Q., Shi, L., Fang, H., Perkins, R., 2005. Assessing QSAR limitations - a regulatory perspective. Curr. Comput. Aided Drug Des. 1, 195–205. http://dx.doi.org/10.2174/1573409053585663.
Tsujita-Inoue, K., Hirota, M., Ashikaga, T., Atobe, T., Kouzuki, H., Aiba, S., 2014. Skin sensitization risk assessment model using artificial neural network analysis of data from multiple in vitro assays. Toxicol. in Vitro 28, 626–639. http://dx.doi.org/10.1016/j.tiv.2014.01.003.
Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., Landsiedel, R., Jaworska, J., Kern, P.S., Gerberick, F., Natsch, A., Emter, R., Ashikaga, T., Miyazawa, M., Sakaguchi, H., 2015. Assessing skin sensitization hazard in mice and men using non-animal test methods. Regul. Toxicol. Pharmacol. 71, 337–351. http://dx.doi.org/10.1016/j.yrtph.2014.12.008.
van der Veen, J.W., Rorije, E., Emter, R., Natsch, A., van Loveren, H., Ezendam, J., 2014. Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals. Regul. Toxicol. Pharmacol. 69, 371–379. http://dx.doi.org/10.1016/j.yrtph.2014.04.018.
Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. second ed.