HHS Public Access

Author manuscript

J Appl Toxicol. Author manuscript; available in PMC 2018 March 01.

Published in final edited form as:

J Appl Toxicol. 2017 March ; 37(3): 347–360. doi:10.1002/jat.3366.

Multivariate Models for Prediction of Human Skin Sensitization Hazard

Judy Strickland[a,*], **Qingda Zang**[a], **Michael Paris**[a], **David M. Lehmann**[b], **David Allen**[a], **Neepa Choksi**[a], **Joanna Matheson**[d], **Abigail Jacobs**[e], **Warren Casey**[c], and **Nicole Kleinstreuer**[c]

[a]ILS, Research Triangle Park, NC, 27709, USA

[b]U.S. Environmental Protection Agency, Research Triangle Park, NC, 27709, USA

[c]National Institutes of Environmental Health Sciences, Research Triangle Park, NC, 27709, USA

[d]U.S. Consumer Product Safety Commission, Rockville, MD, 20850, USA

[e]U.S. Food and Drug Administration, Silver Spring, MD, 20993, USA

Abstract

One of ICCVAM's top priorities is the development and evaluation of non-animal approaches to identify potential skin sensitizers. The complexity of biological events necessary to produce skin sensitization suggests that no single alternative method will replace the currently accepted animal tests. ICCVAM is evaluating an integrated approach to testing and assessment based on the adverse outcome pathway for skin sensitization that uses machine learning approaches to predict human skin sensitization hazard. We combined data from three *in chemico* or *in vitro* assays—the direct peptide reactivity assay (DPRA), human cell line activation test (h-CLAT), and KeratinoSens™ assay—six physicochemical properties, and an *in silico* read-across prediction of skin sensitization hazard into 12 variable groups. The variable groups were evaluated using two machine learning approaches, logistic regression (LR) and support vector machine (SVM), to predict human skin sensitization hazard. Models were trained on 72 substances and tested on an external set of 24 substances. The six models (three LR and three SVM) with the highest accuracy (92%) used: (1) DPRA, h-CLAT, and read-across; (2) DPRA, h-CLAT, read-across, and KeratinoSens; or (3) DPRA, h-CLAT, read-across, KeratinoSens, and log P. The models performed better at predicting human skin sensitization hazard than the murine local lymph node assay (accuracy = 88%), any of the alternative methods alone (accuracy = 63–79%), or test batteries combining data from the individual methods (accuracy = 75%). These results suggest that

[*]Correspondence to: Judy Strickland, ILS, P.O. Box 13501, Research Triangle Park, NC, 27709, USA. strickl2@niehs.nih.gov.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Disclaimer

This article may be the work product of an employee or group of employees of the National Institute of Environmental Sciences (NIEHS), National Institutes of Health (NIH), Environmental Protection Agency (EPA), U.S. Food and Drug Administration, U.S. Consumer Product Safety Commission, or other organizations; however, the statements, opinions, or conclusions contained therein do not necessarily represent the statements, opinions, or conclusions of NIEHS, NIH, EPA, U.S. Food and Drug Administration, U.S. Consumer Product Safety Commission, the United States government, or other organizations. ILS staff provide technical support for NICEATM, but do not represent NIEHS, NTP, or the official positions of any federal agency. The use of commercial product names is for comparative purposes only and does not constitute endorsement by any of the authors, organizations, or agencies.

computational methods are promising tools to effectively identify potential human skin sensitizers without animal testing.

Keywords

Skin sensitization; allergic contact dermatitis (ACD); integrated decision strategy; machine learning; LLNA; DPRA; KeratinoSens; h-CLAT

Introduction

Allergic contact dermatitis (ACD) is an adverse health effect that frequently develops in workers and consumers exposed to skin-sensitizing substances and products. The development of ACD, which includes induction and elicitation phases, is well understood (Jowsey *et al.,* 2006). The induction phase of ACD occurs when a susceptible individual is exposed topically to a skin-sensitizing substance. The substance passes through the epidermis, where it generally forms a hapten complex with dermal proteins. The hapten complex is processed by the Langerhans cells, the resident antigen-presenting cells in the skin. The processed hapten complex is then transported by the Langerhans cells to the draining lymph nodes, where the hapten complex is presented as an antigen to T-lymphocytes, leading to T-lymphocyte proliferation. Studies have shown that the magnitude of T-lymphocyte proliferation correlates with the likelihood that skin sensitization will develop (Kimber and Dearman 1991; Kimber and Dearman 1996). The elicitation phase of ACD occurs when the individual is topically re-exposed to the same substance. As in the induction phase, the substance penetrates the epidermis, complexes with dermal proteins, is processed by the Langerhans cells, and is presented to circulating T-lymphocytes. The antigen-specific T-lymphocytes are then activated, which causes release of cytokines and other inflammatory mediators. This release produces a rapid dermal immune response that can lead to ACD (Basketter *et al.,* 2003; ICCVAM 1999; Jowsey *et al.,* 2006; Sailstad *et al.,* 2001).

To minimize the occurrence of ACD from exposure to chemical products, national and international regulatory authorities require that skin-sensitizing substances be labeled to identify the potential hazard posed by these items. Such hazards have historically been characterized based on results from animal tests that can use large numbers of animals and produce a painful allergic reaction during testing. For example, the guinea pig maximization test and the Buehler test use 20 to 40 animals per substance (OECD 1992). An alternative method, the murine local lymph node assay (LLNA) reduces and refines animal uses compared to guinea pig tests, but still uses animals (ICCVAM 1999).

Alternative methods replace, reduce, and refine (cause less pain and distress) animal use for chemical safety testing. Fostering the evaluation and promoting the use of alternative test methods for regulatory use in skin sensitization hazard assessment has long been a priority for the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) (Dean *et al.,* 2001; ICCVAM 1999; NIEHS 2013; Sailstad *et al.,* 2001). Numerous non-animal alternatives for skin sensitization hazard assessment have been developed and are at various stages of evaluation (Johansson and Lindstedt 2014; Mehling *et*

al., 2012; Wong *et al.,* 2015). Because skin sensitization is a complex process, it is unlikely that any individual alternative method will completely replace current animal tests. In fact, the *in vitro* and *in chemico* methods that have gained acceptance as international test guidelines are not recommended as stand-alone replacements for animal test methods (OECD 2015b; 2015c). Thus, a number of approaches to integrate the information from multiple alternative methods as a way to overcome the limitations of individual tests and more accurately assess the potential for skin sensitization have also been developed (Jaworska *et al.,* 2013; Jaworska *et al.,* 2015; Natsch *et al.,* 2009; Nukada *et al.,* 2013; Strickland *et al.,* 2016; Urbisch *et al.,* 2015). These approaches use combinations of non-animal tests that align with key events in the adverse outcome pathway for skin sensitization (OECD 2012).

This paper, whose authors include members of the ICCVAM Skin Sensitization Working Group, describes integrated decision strategies that use non-animal data to predict human skin sensitization hazard. We have previously described the application of a number of machine learning approaches to integrate existing non-animal skin sensitization data and physicochemical properties that may be associated with skin penetration to predict skin sensitization hazard based on LLNA outcomes (Strickland *et al.,* 2016). The optimal approach achieved greater accuracy (96%) than any of the individual non-animal test methods (≤85%) when compared to LLNA outcomes. Our next step, which is reported here, was to develop models to predict human skin sensitization hazard better than the LLNA, which has previously been demonstrated as 72% accurate in predicting human skin sensitization hazard (ICCVAM 1999).

Materials and Methods

Data Collection and Substance Database

We compiled a chemical database by collecting publicly available data for the direct peptide reactivity assay (DPRA), KeratinoSens™; the human cell line activation test (h-CLAT), and the LLNA (Table 1). DPRA, KeratinoSens, and h-CLAT were selected because international test guidelines are in the process of being adopted or were recently adopted by the Organisation for Economic Co-operation and Development (OECD) (OECD 2015a; 2015b; 2015c).

Data from the LLNA were used to compare its performance in predicting human skin sensitization hazard with that of the non-animal integrated decision strategy. The majority of the LLNA data were collected previously by the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (http://ntp.niehs.nih.gov/go/40500). These data include sensitizer/nonsensitizer determinations for each substance as well as stimulation indices at the concentrations tested. The LLNA data for five substances that were not in this database came from published literature (Table 1).

The majority of the human skin sensitization hazard data were adapted from an ICCVAM evaluation of the LLNA for human potency categorization (ICCVAM 2011) and from potency categorizations listed in Basketter *et al.* (2014). While ICCVAM (2011) compiled sensitization results from human predictive patch tests, the potency assessments listed in

Basketter *et al.* (2014) were developed by a panel of experts that evaluated prevalence from dermatologic clinic data as well as data from predictive patch tests. Note that the Basketter *et al.* (2014) assessment has limitations for assessing predictive alternative tests because it is not based solely on the intrinsic allergenicity of a substance, but also considers exposure. Priority was given to the categorizations in Basketter *et al.* (2014) to resolve any conflicts (n=6) between these references. Human hazard data for 10 substances not included in these sources came from Basketter and Kimber (2006), Basketter *et al.* (1999), Bjorkner (1984), and White *et al.* (2006). References for each substance are provided in Supplemental File 1.

In total, the database for the analysis reported here includes 96 substances with human hazard data that were tested in DPRA, KeratinoSens, h-CLAT, and the LLNA. For each substance, we also collected data on six physicochemical properties relevant to skin exposure and penetration: octanol:water partition coefficient, water solubility, vapor pressure, melting point, boiling point, and molecular weight. These properties have been important for other models or weight-of-evidence assessments for skin sensitization potential (Jaworska *et al.,* 2013; 2011; Patlewicz *et al.,* 2014). We also performed an *in silico* prediction of skin sensitization hazard using the read-across algorithm in QSAR Toolbox v3.2 (OECD 2014).

Characterization of the Substances

Of the 96 substances in the database, 69% (66/96) are human sensitizers and 31% (30/96) are nonsensitizers. Of the 66 sensitizers, three are prehaptens that require oxidation to induce a skin sensitization reaction, 14 are prohaptens that require metabolism, and three are pre/prohaptens that require both oxidation and metabolism. See Supplemental File 1 for the prehapten and prohapten information on each substance and the corresponding reference.

The 96 substances represent 14 product categories (Fig. 1). Product category information was obtained from the following sources:

- U.S. National Library of Medicine (NLM) Hazardous Substances Databank (HSDB; http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB)
- NLM Haz-Map database (http://hazmap.nlm.nih.gov/)
- NLM Household Products Database (http://hpd.nlm.nih.gov/index.htm)
- International Programme on Chemical Safety INCHEM database (http://www.inchem.org/)
- NLM Drug Information Portal (http://druginfo.nlm.nih.gov/drugportal/drugportal.jsp?APPLICATION_NAME=drugportal)
- U.S. National Toxicology Program (http://ntp.niehs.nih.gov/)

- List of pesticides registered by the U.S. Environmental Protection Agency (A. Lowit, personal communication)
- The United Nations Joint Expert Committee on Food Additives (http://apps.who.int/food-additives-contaminants-jecfa-database/search.aspx?)
- The Good Scents Company (http://www.thegoodscentscompany.com/)
- Scientific literature (i.e., papers that also presented test method data)
- Chemical Book (http://www.chemicalbook.com/)

Structural variety among substances in the database was assessed using ChemoTyper v1.0 (https://chemotyper.org), a free software developed under contract with the U.S. Food and Drug Administration. ChemoTyper defines 729 chemotypes, generic structural fragments that represent chemical features, including connected and unconnected chemical patterns as well as atom, bond, and molecular properties (Yang *et al.,* 2015). The 96 substances in the database included 183 chemotypes that appeared at a frequency of 1 to 59 over the entire dataset (Fig. 2). The most common chemotypes were bond:C=O_carbonyl_generic (59 substances), ring:aromatic_benzene (57 substances), chain:aromaticAlkane_Ph-C1_acyclic_generic (36 substances), bond:COH_alcohol_generic (31 substances), and chain:alkaneLinear_ethyl_C2(H_gt_1) (30 substances). Individual substances were characterized by 2–35 chemotypes each. See Supplemental File 1 for the chemotypes associated with each substance.

Data Inputs

The non-animal methods proposed for the integrated decision strategy are aligned to the adverse outcome pathway (AOP) for skin sensitization initiated by covalent binding to proteins (OECD 2012).

DPRA—DPRA is an *in chemico* test that assesses the ability of a substance to form a hapten–protein complex (Gerberick *et al.,* 2004; 2007; OECD 2015b), which is the molecular initiating event in the skin sensitization AOP as described in OECD (2012). The integrated decision strategy evaluated average cysteine peptide depletion (Cys), average lysine peptide depletion (Lys), average depletion of cysteine and lysine peptides (Avg.Lys.Cys), and sensitizer/nonsensitizer outcome.

KeratinoSens—The KeratinoSens test method assesses the ability of a substance to activate cytokines and induce cytoprotective genes in keratinocytes (Emter *et al.,* 2010; OECD 2015c), the second key event in the skin sensitization AOP (OECD 2012). We used a binary classification (sensitizer/nonsensitizer) because continuous KeratinoSens data (i.e., effective concentration at 1.5-fold luciferase induction) were not available for all substances at the time data were collected.

h-CLAT—h-CLAT assesses the ability of a substance to activate and mobilize dendritic cells in the skin (Ashikaga *et al.,* 2006; OECD 2015a), the third key event of the skin sensitization AOP (OECD 2012). We used a binary classification (sensitizer/nonsensitizer) because continuous h-CLAT data (i.e., effective concentration at 150% induction for the CD86 marker and the effective concentration at 200% induction for the CD54 marker) were not available for all substances when data were collected.

In Silico Read-across—QSAR Toolbox v3.2 (OECD 2009; 2014) was used to generate an *in silico* read-across prediction of whether each substance or its predicted auto-oxidation product or metabolite was a sensitizer or nonsensitizer based on *in vivo* data from structurally and mechanistically similar analogs. The *in silico* predictions cover the fourth key event of the AOP, T cell activation and proliferation (OECD 2012), and all preceding key events because *in vivo* data (LLNA, guinea pig, and human outcomes) are used to determine the read-across result. The read-across protocol for QSAR Toolbox is provided as Supplemental File 2. Briefly:

- The Chemical Abstracts Service Registry Number for a substance was provided as an input to QSAR Toolbox. All four protein binding profilers in QSAR Toolbox were used to search for protein binding alerts: OASISv1.2, OECD, potency, and alerts for skin sensitization by OASISv1.2.
- For substances with no protein binding alerts, auto-oxidation products and skin metabolites were generated and then those were profiled for protein binding alerts. If the oxidation products and metabolites had no alerts, the substance was then classified as a nonsensitizer.
- Test substances, products, or metabolites with protein binding alerts were grouped into categories with substances of similar structural and mechanistic characteristics. The read-across prediction of skin sensitization hazard was produced using the *in vivo* skin sensitization hazard data for the substances nearest the target substance, based on log K_{ow}.

Physicochemical Properties—We collected data for octanol:water partition coefficient, water solubility, vapor pressure, molecular weight, melting point, and boiling point from the following sources, with preference given to experimental values:

- SRC, Inc. – Epi Suite Data (http://esc.syrres.com/interkow/EPiSuiteData.htm)
- ChemID*plus* – NLM Toxicology Data Network (TOXNET) Database (http://chem.sis.nlm.nih.gov/chemidplus)
- ChemSpider – Royal Society of Chemistry database (http://www.chemspider.com/)

- HSDB – NLM Toxicology Data Network (TOXNET) Database (http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB)
- European Chemicals Agency – Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) database (http://echa.europa.eu/information-on-chemicals)

For seven substances, values for one or more physicochemical properties could not be located. In these cases, values were imputed via quantitative structure–property relationship models built using binary molecular fingerprints and machine learning approaches (Q. Zang et al., unpublished data). See Supplemental File 1 for the individual physicochemical properties and data sources for each substance.

Data Processing

If a substance had multiple continuous results for the DPRA, we calculated a geometric mean of those results after first setting negative peptide depletion values to zero. If a substance had multiple sensitizer and nonsensitizer results for a particular assay, we used the most prevalent result; if there were an equal number of sensitizer and nonsensitizer results for a substance, the substance was classified as a sensitizer. There were eight substances with an equal number of sensitizer and nonsensitizer results for DPRA, four substances for KeratinoSens, seven substances for h-CLAT, and four substances for the LLNA. The final results for each substance are provided in Supplemental File 1, along with the QSAR Toolbox read-across results.

Development of Models to Predict Human Skin Sensitization Hazard

Selection of Training and Test Sets—The database of 96 substances included 66 human sensitizers (69% [66/96]) and 30 human nonsensitizers (31% [30/96]). These substances were divided into training (n=72) and external test (n=24) sets with similar characteristics, such as product use categories, diversity of chemical structures, prehaptens/prohaptens, and mechanistic protein binding domains (Supplemental File 3). Training and test sets also included approximately the same ratio of sensitizers and nonsensitizers as the database of 96 substances. Of the 72 substances in the training set, there were 51 sensitizers (71% [51/72]) and 21 nonsensitizers (29% [21/72]), while the 24 substances in the test set were comprised of 15 sensitizers (63% [15/24]) and 9 nonsensitizers (37% [9/24]).

Model Variables—The 13 variables listed in Table 2 were considered for building and testing multivariate machine learning models. DPRA, h-CLAT, KeratinoSens and QSAR Toolbox are binary categorical variables that were assigned a value of 1 for sensitizers and 0 for nonsensitizers. The variable Avg.Lys.Cys represents the average lysine (Lys) and cysteine depletion (Cys) measurements from DPRA. All of these data were used as potential independent variables, in differing combinations, to predict human skin sensitization hazard.

Variable Importance Ranking—A random forest analysis was performed to assess the importance of each independent variable to the model based on how much the prediction error (i.e., mean squared error) increases when each variable in turn is replaced with random

noise while all others are left unchanged. Random forest is a non-linear consensus classification method based upon an ensemble of decision trees that are grown from separate bootstrap samples of the training data (Diaz-Uriarte 2007; Hao *et al.,* 2011). A subset of samples, called out-of-bag samples, are not employed for tree growth and are instead used to evaluate the prediction accuracy. The deterioration in model quality is evaluated by the relative change in the error for the out-of-bag validation over all of the trees. After all the variables are successively permuted for all the samples, the random forest algorithm provides a ranked list of the variables ranging from the most to the least important in descending order (Zang *et al.,* 2013).

Machine Learning Approaches—Two machine learning approaches, logistic regression (LR) and support vector machine (SVM), were used to develop a series of binary models to classify substances as human sensitizers or nonsensitizers.

LR is a probabilistic statistical classification method (Varmuza and Peter 2009). Binomial LR is used in situations in which the observed outcome for a dependent variable can have only two possible types, such as sensitizer and nonsensitizer. Probability scores are used to establish a relationship between the independent variables, i.e., the data inputs, and the categorical dependent variable, i.e., human skin sensitization hazard. The log-transformed posterior probabilities of the sensitizer and nonsensitizer classes are fitted to a linear function of the independent variables with the condition that each probability has a value of between 0 and 1 and the sum of the two probabilities is 1.

SVM performs classification by finding an optimal hyperplane as the decision boundary for separating the sensitizer and nonsensitizer classes. The hyperplane maximizes the margin between the closest data points of each class (Shen *et al.,* 2011). The complex class boundary is modeled by using the Gaussian radial basis function kernel, which maps linearly inseparable input data into a higher dimensional feature space where the non-linear relationship is expressed in linear form and the sensitizers and nonsensitizers can be linearly separated.

The following packages in the R statistical analysis software for Windows (v3.2.1) (R Core Team 2013) were used to build the models:

- Package *randomForest*: for random forest
- Package *MASS*: for logistic regression
- Package *e1071*: for support vector machine

Once each LR and SVM model was trained using the training set of substances, it was used to predict human skin sensitization outcomes for each substance in the test set. These outcomes were reported as probabilities; substances with a probability greater than 0.5 of being either a sensitizer or nonsensitizer were assigned to the respective class.

Evaluation of Model Performance—Model performance was evaluated by calculating sensitivity, specificity, and accuracy for the training and test sets. These metrics were calculated by the following formulae:

$$Sensitivity = \frac{True\ Postives}{True\ Positives + False\ Negatives}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

$$Accuracy = \frac{True\ Postives + True\ Negatives}{True\ Postives + False\ Negatives + True\ Negatives + False\ Positives}$$

To confirm their robustness and reliability, the predictive models with the best performance were also evaluated using a leave-one-out cross-validation (LOOCV) procedure. The LOOCV avoids any bias introduced during the selection of test and training sets. To implement this procedure, the training and test set substances were combined. Then, 95 substances were used as the training data for building the model. The single excluded substance then served as the test set. The procedure was performed 96 times with each of the substances used exactly once for external validation. The performance metrics were averaged over the 96 iterations.

Performance of the machine learning models for predicting human skin sensitization hazard was compared with the performance of the LLNA, the individual non-animal methods alone (DPRA, KeratinoSens, h-CLAT, and read-across), and with two test battery approaches using results from the non-animal methods. Test Battery 1 classified a substance as a sensitizer if any one non-animal method classified the substance as a sensitizer. Test Battery 2 classified a substance as a sensitizer if at least two non-animal methods classified the substance as a sensitizer.

Results

Analysis of Variable Importance

A random forest analysis was conducted to assess the relative importance of the seven non-animal test method variables and six physicochemical property variables (listed in Table 2) for predicting human skin sensitization hazard. Variable importance measures the degree of association between a given variable and the prediction results of a classification model, and hence variables with high importance have a strong association with the prediction performance. Fig. 3 presents the results, with the variables ranked in descending order of importance.

The most important variables were Cys and Avg.Lys.Cys from DPRA, followed by the read-across prediction from QSAR Toolbox and the outcome of the h-CLAT. The least important variables mainly represented the physicochemical properties, with the octanol:water partition coefficient, log P, exhibiting higher importance than the other properties. To eliminate redundancy among the DPRA variables, we decided to use only the Avg.Lys.Cys readout from the DPRA, since it incorporates both the Lys and the Cys measurements. Overall, the data from the non-animal methods captured important information and were more discriminative than the physicochemical properties.

Performance of the Variable Groups with the SVM and LR Models

Seven variable groups, Groups A–G in Table 3, were defined using different combinations of the non-animal methods plus either log P, the most important physicochemical property, or

all six physicochemical properties. One variable group, Group H, contained only the six physicochemical properties. Four variable groups, Groups I–L, used different combinations of two or three of Avg.Lys.Cys from DPRA, h-CLAT, or KeratinoSens along with the *in silico* read-across QSAR Toolbox method without any physicochemical properties. The performance of each model was examined in terms of sensitivity, specificity, and accuracy both against the training set used to develop the model and the test set used to evaluate the model.

Models with log P as the only physicochemical property variable performed better than similar models with all six physicochemical properties. For the LR models, all seven variable groups containing log P produced higher accuracy for the test set of 24 substances than those containing all six physicochemical properties (Fig. 4). For the SVM models, two of the seven variable groups containing log P, i.e., Group C (KeratinoSens + Toolbox + Log P) and Group D (Avg.Lys.Cys + Toolbox + Log P), produced higher accuracy for the test set than groups containing all physicochemical properties (Fig. 5). Because the variable groups containing log P had the same or higher accuracy than variable groups with all six physicochemical properties, variable groups used in subsequent analyses included only log P when physicochemical properties were included with the non-animal methods.

Table 4 summarizes the accuracy, sensitivity and specificity of the LR and SVM models for the training and test sets for the 12 variable groups listed in Table 3. These variable groups used only log P when both physicochemical properties and non-animal methods were included. The variable group with the worst performance was Group H, which included only the six physicochemical properties. Accuracy was 54–58%, sensitivity was 67%, and specificity was 33–44% for the test set.

The variable groups with the highest performance for the test set were the same using either the LR or SVM models. Three variable groups—Group A (Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox + Log P), Group I (Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox) and Group K (Avg.Lys.Cys + h-CLAT + Toolbox)—produced accuracy of at least 92%, sensitivity of at least 87% and specificity of at least 89% for both training and test sets. These six models (three variable groups times two machine learning approaches) correctly classified all prehaptens and prohaptens in the dataset. Variable Group A, with Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox + Log P, produced the best performing model using SVM, with accuracy of 94% (68/72), sensitivity of 94% (48/51), and specificity of 95% (20/21) for the training set; and accuracy of 92% (22/24), sensitivity of 93% (14/15), and specificity of 89% (8/9) for the test set.

In addition to testing the models with the external test set, we applied the LOOCV procedure to further assess the performance of the six models with classification accuracy of 92% (i.e., Variable Groups A, I, and K with the LR and SVM approaches). As shown in Table 5 for the whole dataset (training plus test sets), all models achieved an accuracy of at least 91%. These results are very close to those obtained from the external test set, confirming the robustness and reliability of the multivariate model.

Misclassified Substances

Nine substances were misclassified by one or more of the six SVM and LR models with the highest accuracy. The reasons for the misclassifications are not entirely clear. Looking for commonalities, we observed that five of the nine misclassified substances are pharmaceuticals: sulfanilamide, streptomycin sulfate, penicillin G, benzocaine, and coumarin. Sulfanilamide, streptomycin sulfate, and penicillin G are antibiotics. Penicillin G formerly contained impurities (de Weck *et al.,* 1968), which reminds us to mention that one of the limitations of the data is that the purity of the chemicals is not necessarily the same across all tests or across time. Sulfanilamide (Gao *et al.,* 2014) and benzocaine (Allen 1993) have been reported to produce photocontact allergy, which may be difficult to distinguish from simple allergic contact dermatitis. No other commonalities, including structural similarity, were noted among the misclassified substances.

Training Set—The six SVM and LR models with the highest accuracies misclassified a total of six substances, one false positive and five false negatives, in the training set (Table 6). None of the false negatives were prehaptens (n = 2 in the training set), prohaptens (n = 10 in the training set), or pre/prohaptens (n = 2 in the training set).

2-Methoxyl-4-methylphenol was the only false positive substance in the training set. It was misclassified as a sensitizer by all of the six best LR and SVM models; three of the four non-animal methods classified it as a sensitizer. 2-Methoxyl-4-methylphenol was tested in a human repeat insult patch test (HRIPT) at 118 μg/cm^2 with a negative result (ICCVAM 2011), however, other references classify it as a human sensitizer (Basketter *et al.,* 1999). Although no supporting test results were provided in Basketter *et al.* (1999), we assume that the highest dose tested in the HRIPT was inadequate to produce sensitization. HRIPT are typically performed to confirm a no adverse effect level from animal studies rather than assess hazard (Politano and Api 2008).

Of the five false negative substances, sulfanilamide and penicillin G were misclassified by all six LR and SVM models. All of the non-animal data yielded negative results for sulfanilamide. An ICCVAM report classified it as a Category 1B (i.e., weak) sensitizer (ICCVAM 2011) based on the Globally Harmonized System of Classification and Labelling of Chemicals (GHS) (UN 2013). For pencillin G, two of the four non-animal data inputs yielded negative results. Penicillin G is consistently positive in human predictive tests (ICCVAM 2011). Benzocaine was false negative in five of the best six SVM and LR models, streptomycin sulfate was false negative in four, and α-amyl cinnamaldehyde was false negative in one. Penicillin G, streptomycin sulfate, and benzocaine are consistently positive in human predictive tests (ICCVAM 2011). While streptomycin sulfate is a GHS 1A (i.e., strong) sensitizer in humans, penicillin G, and benzocaine are 1B sensitizers. Although α-amyl cinnamaldehyde produced negative results in human predictive tests (ICCVAM 2011), it is recognized as a weak sensitizer in humans by Basketter *et al.* (2014). Thus, four of the five false negative substances are weak human sensitizers.

Test Set—The six LR and SVM models with the highest accuracies (i.e., Variable Groups A, I, and K) misclassified three substances, one false positive and two false negatives, in the

test set (Table 6). Coumarin, was misclassified as a nonsensitizer by all six models. Three of the four non-animal methods misclassified coumarin as a nonsensitizer. Lilial was misclassified as a nonsensitizer by only one model and was misclassified by only one non-animal method. Neither coumarin nor lilial require oxidation or metabolism to produce skin sensitization. Thus, all six models correctly classified all six prehaptens, haptens, and pre/prohaptens in the test set. Both ICCVAM (2011) and Basketter *et al.* (2014) classified coumarin and lilial as weak human sensitizers. Pentachlorophenol was misclassified as a sensitizer by five of the six LR and SVM models. Three of the four non-animal methods misclassified it as a sensitizer. The human evidence for skin sensitization potential for pentacholorophenol is equivocal. Although ICCVAM (2011) classified it as a weak sensitizer, we chose to classify it as a nonsensitizer for this analysis based on its classification in Basketter *et al.* (2014), which acknowledges that although it is a sensitizer, sensitization in the general population is likely to be extremely rare, occurring only after prolonged exposure at high levels.

Performance of the LLNA, Individual Methods, and Test Battery Approaches

For comparison with the results from the machine learning approaches, Table 7 shows performance statistics for the individual non-animal methods, the LLNA, and two test battery approaches to predict human skin sensitization hazard for the training and test sets and the entire set of 96 substances. For the test set of 24 substances, the models performed better at predicting human skin sensitization hazard than the murine local lymph node assay (accuracy = 88%), any of the alternative methods alone (accuracy = 63–79%), or test batteries combining data from the individual methods (accuracy = 75%).

Because the test set was relatively small, we also include the performance of the individual non-animal tests for the entire set of 96 substances: accuracy = 74–81%, sensitivity = 77–88%, and specificity = 67–80%. Of the four non-animal methods, DPRA had the highest accuracy (83%) and specificity (80%), and h-CLAT had the highest sensitivity (88%). The LLNA, which is the recommended stand-alone animal test for skin sensitization (ICCVAM 1999), had slightly higher accuracy (84%) and sensitivity (92%) than the best non-animal tests, but had the same specificity (67%) as the lowest performing non-animal tests. Compared with the individual non-animal methods, Test Battery 1 similar accuracy (78%), higher sensitivity (99%), and much lower specificity (33%). Test Battery 2 had higher accuracy (85%) and higher sensitivity (94%) than the individual non-animal methods, with the same specificity as the lowest performing non-animal methods (67%). Thus, in comparison to individual non-animal methods, battery approaches, and the LLNA, our integrated strategies using machine learning provided superior predictions for human skin sensitization hazard and achieved a better balance between sensitivity and specificity.

Discussion

ICCVAM is committed to the evaluation and implementation of alternative test methods for regulatory use in skin sensitization hazard assessment (Dean *et al.,* 2001; ICCVAM 1999; NIEHS 2013; Sailstad *et al.,* 2001). Considering the inherent complexity of the AOP for substances that produce skin sensitization, it is likely that an integrated decision strategy

combining data from several non-animal methods is needed to accurately predict this adverse health outcome. Here, we used data from the DPRA, KeratinoSens, and h-CLAT assays along with six physicochemical properties and an *in silico* read-across prediction of skin sensitization hazard as inputs to two machine learning approaches to predict human skin sensitization potential.

This study affirms the widely-held belief that integrated approaches to skin sensitization testing outperform individual non-animal methods used in isolation (Rovida *et al.,* 2015). For the entire set of 96 substances used in this study, the highest accuracy for the prediction of human skin sensitization hazard outcomes for any non-animal method alone was only 83% for the DPRA (Table 7). Combining non-animal methods into simple test batteries slightly improved accuracy to 85%. However, the six best machine learning models markedly improved upon the individual methods and simple test batteries with accuracy of 93–94%. As a comparison, the LLNA, which is the recommended stand-alone animal test for skin sensitization (ICCVAM 1999), had an accuracy of only 84%.

High accuracy for both training and test sets (92–94%) was achieved for hazard classification predictions for six LR and SVM models using different variable combinations of non-animal data, read-across from QSAR Toolbox, and log P (i.e., Variable Groups A, I, and K). The LOOCV, which avoids bias introduced during the selection of test and training sets, yielded accuracies of 91–94%, which was nearly identical to the accuracies for the test set (92%). The similarity of these accuracies in the test set and LOOCV evaluations indicates that the test and training sets were well-chosen and that the models are stable. These results serve to demonstrate the potential utility of the integrated decision strategy developed here for identifying potential human skin sensitizers.

Models using log P in combination with non-animal methods often outperformed analogous models relying on six different physicochemical properties (Figs. 4 and 5). Interestingly, four of the six best performing models (i.e., Variable Group I, Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox with LR and SVM; and Variable Group K, Avg.Lys.Cys + h-CLAT + Toolbox with LR and SVM) required no physicochemical property information (Table 4). The two models using Variable Group K only required three of the four non-animal methods and may provide an avenue for some laboratories to conserve resources. However, there may be a very small preference for the SVM model using all of the non-animal methods and log P (i.e., Variable Group A, Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox + Log P) due to the slightly higher accuracy and sensitivity for the training set. The *in chemico* and *in vitro* non-animal methods used for the LR and SVM models described here come from internationally accepted (or, in the case of h-CLAT, nearly accepted) OECD test guidelines. Together, these methods assess three of the four key events in the AOP for skin sensitization initiated by covalent binding to proteins. Still, each method, when used alone, has limitations that hinder its identification of potential sensitizers. The DPRA has no metabolic capacity and thus would not be expected to correctly classify prohaptens (OECD 2015b). KeratinoSens (OECD 2015c) and h-CLAT (OECD 2015a) correctly classify some, but not all prohaptens. Inclusion of the *in silico* read-across input, which evaluated auto-oxidation products and skin metabolites if no protein-binding alerts were identified in the parent compound, may have facilitated correct prediction of prohaptens.

Although DPRA has consistently classified prehaptens correctly (OECD 2015b), KeratinoSens (OECD 2015c) and h-CLAT (OECD 2015a) have not. Collectively, DPRA, h-CLAT, and read-across from QSAR Toolbox (Variable Group K) correctly identified all prehaptens and prohaptens, with or without the inclusion of the variables KeratinoSens and log P (Variable Groups A and I). Given the known limitations of the individual assays and the relatively small dataset, additional substances requiring metabolic capacity should be evaluated to bolster confidence in this approach. Although DPRA was included in the models using continuous data (Avg.Lys.Cys), KeratinoSens and h-CLAT were included in the models as binary (sensitizer or nonsensitizer) inputs. Continuous data for KeratinoSens and h-CLAT, which were not available at the time of our data collection, may produce different model outcomes. Continuous data for all three methods would be preferable because it would provide more information for the modeling efforts to distinguish between the human sensitizer and nonsensitizer classes.

Previous efforts to integrate data to predict skin sensitization hazard without the use of animals have emphasized prediction of LLNA outcomes using uncomplicated test batteries (Bauch *et al.,* 2012; Natsch *et al.,* 2009; Natsch *et al.,* 2013; Nukada *et al.,* 2013; Urbisch *et al.,* 2015) and testing strategies (Nukada *et al.,* 2013; Takenouchi *et al.,* 2015) as well as various machine learning approaches (Hirota *et al.,* 2015; Jaworska *et al.,* 2013; Jaworska *et al.,* 2011; Luechtefeld *et al.,* 2015; Pirone *et al.,* 2014; Tsujita-Inoue *et al.,* 2014). While it is recognized as the gold standard for identification and characterization of skin-sensitizing chemicals (Anderson *et al.,* 2011), the LLNA only predicted human sensitization with 84% accuracy for the 96 substances used in this study (Table 7). This accuracy is somewhat higher than the 72% found in the ICCVAM evaluation of LLNA performance (ICCVAM 1999), which may be due to the difference in the substance sets evaluated.

While there is a clear and pressing need to predict skin sensitization outcomes in humans, only two studies prior to this one have evaluated the predictivity of non-animal methods based on what is known about human skin sensitization (Urbisch *et al.,* 2015; van der Veen *et al.,* 2014). Recently, Urbisch *et al.* (2015) published a study of 213 substances with LLNA and human data. In this study, the performance of the LLNA for predicting sensitization outcomes in humans was shown to be 82% whereas the accuracy of individual non-animal methods (i.e., DPRA, KeratinoSens and h-CLAT) ranged from 78–84% (Urbisch *et al.,* 2015). Accuracy was improved to 90% by using a “two-out-of-three” approach. Similarly, an analysis by van der Veen *et al.* (2014) found that the accuracy of the LLNA for predicting human skin sensitization was 78%, which was inferior to integrated testing strategies. Majority voting analysis (most prevalent result of DPRA, KeratinoSens or gene signature, and h-CLAT), yielded an accuracy of 96% whereas completion of a three-stage tiered approach, which included a QSAR analysis along with DPRA, KeratinoSens or gene signature, and h-CLAT, achieved 100% accuracy. These results are comparable to the highest performing LR and SVM models developed here (i.e., 92% accuracy), although the performance of our models cannot be directly compared with Urbisch’s and van der Veen’s because the three studies did not use the same substance set. While the tiered approach described by van der Veen *et al.* (2014) performed very well using a development set of substances, neither study tested assay performance using an external set. Consequently,

additional studies using more substances are warranted to more accurately gauge performance of the various models.

To date, all efforts to predict human skin sensitization potential using integrated testing strategies have been limited by an inability to support potency classification decisions. While potency data are necessary for risk assessors to identify the threshold level of exposure to a substance below which it is unlikely to produce skin sensitization, some regulatory classification and labeling applications only require hazard identification. For example, skin sensitization hazard information is used by the U.S. Environmental Protection Agency (EPA 2012a; 2012b) and the U.S. Occupational Safety and Health Administration (OSHA 2012) to caution consumers and workers about contact with potential skin sensitizers. In addition, consistent with the GHS (UN 2013), OSHA requires potency classification only if the skin sensitization data are adequate to characterize potency (OSHA 2012) (Appendix A). For hazard identification, the LR and SVM models developed here offer an advantage over the other published models designed to predict human outcomes (Urbisch *et al.,* 2015; van der Veen *et al.,* 2014) in that laboratories have more than one model to choose from and, depending on specific needs, models using only two non-animal laboratory methods can be selected. The tiered strategy discussed above (van der Veen *et al.,* 2014) requires the use of four different QSAR models in tier I alone. The results of tier I must then be integrated into a Bayesian prediction model. Furthermore, application of van der Veen *et al.*'s approach requires unique technical expertise and additional expense to assess changes in the expression of 10 genes. On the other hand, the approach we describe supports the generation of high-quality predictions using freely available software supported by the OECD (OECD 2014) and publicly available physicochemical property data.

For a number of years, the LLNA has been the gold standard for identifying and determining the relative potency of skin sensitizers. However, recent evaluation of one database revealed that one-third of strong human sensitizers are underclassified as weaker sensitizers by this method (ICCVAM 2011). Consequently, the LLNA is not recommended by ICCVAM as a stand-alone method to predict skin sensitization potency (ICCVAM 2011), leaving a void in the risk assessment toolbox. Given the limitations of the LLNA and the superior performance of screening strategies built using non-animal methods, a logical next step would be to continue the development of mechanistically rational integrated decision strategies for predicting skin sensitization hazard with the capacity to predict sensitizer potency in humans. To that end, our future work will explore the use of continuous variables for DRPA, h-CLAT, and KeratinoSens to support the development of models to predict human skin sensitization potency.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Drs. R. Luebke, M. Ward, D. Germolec, and B.A. Merrick for their thoughtful critical review of this manuscript. This project was funded in whole or in part with federal funds from the NIEHS, NIH under

contract HHSN273201500010C to ILS in support of the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods.

References

29 CFR 1910.1200. Title 29- Labor. OSHA; Washington DC: 2012. Occupational Safety and Health Standards.

40 CFR 158.500. Title 40- Protection of Environment. EPA; Washington DC: 2012a. Toxicology Data Requirement Table.

40 CFR 161.340. Title 40- Protection of Environment. EPA; Washington DC: 2012b. Toxicology Data Requirements.

Allen JE. Drug-induced photosensitivity. Clin Pharm. 1993; 12:580–587. [PubMed: 8222522]

Anderson SE, Siegel PD, Meade BJ. The LLNA: A brief review of recent advances and limitations. J Allergy. 2011; 2011:424203.doi: 10.1155/2011/424203

Ashikaga T, Sakaguchi H, Sono S, Kosaka N, Ishikawa M, Nukada Y, Miyazawa M, Ito Y, Nishiyama N, Itagaki H. A comparative evaluation of in vitro skin sensitisation tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). Altern Lab Anim. 2010; 38:275–284. [PubMed: 20822320]

Ashikaga T, Yoshida Y, Hirota M, Yoneyama K, Itagaki H, Sakaguchi H, Miyazawa M, Ito Y, Suzuki H, Toyoda H. 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. 2006; 20:767–773. DOI: 10.1016/j.tiv.2005.10.012 [PubMed: 16311011]

Ball N, Cagen S, Carrillo JC, Certa H, Eigler D, Emter R, Faulhammer F, Garcia C, Graham C, Haux C, Kolle SN, Kreiling R, Natsch A, Mehling A. Evaluating the sensitization potential of surfactants: Integrating data from the local lymph node assay, guinea pig maximization test, and in vitro methods in a weight-of-evidence approach. Regul Toxicol Pharmacol. 2011; 60:389–400. [PubMed: 21645576]

Basketter DA, Alepee 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. Categorization of chemicals according to their relative human skin sensitizing potency. Dermatitis. 2014; 25:11–21. DOI: 10.1097/DER.0000000000000003 [PubMed: 24407057]

Basketter DA, Gerberick GF, Kimber I, Loveless SE. The local lymph node assay: a viable alternative to currently accepted skin sensitization tests. Food Chem Toxicol. 1996; 34:985–997. [PubMed: 9012774]

Basketter, DA., Kimber, I. Predictive tests for irritants and allergens and their use in quantitative risk assessment. In: Frosch, P.Menné, T., Lepoittevin, J-P., editors. Contact Dermatitis. Springer Verlag; Heidelberg: 2006. p. 179-188.

Basketter DA, Lea LJ, Cooper K, Stocks J, Dickens A, Pate I, Dearman RJ, Kimber I. Threshold for classification as a skin sensitizer in the local lymph node assay: a statistical evaluation. Food Chem Toxicol. 1999; 37:1167–1174. [PubMed: 10654593]

Basketter DA, Smith Pease CK, Patlewicz GY. Contact allergy: the local lymph node assay for the prediction of hazard and risk. Clin Exp Dermatol. 2003; 28:218–221. [PubMed: 12653718]

Bauch C, Kolle SN, Fabian E, Pachel C, Ramirez T, Wiench B, Wruck CJ, Ravenzwaay BV, Landsiedel R. Intralaboratory validation of four in vitro assays for the prediction of the skin sensitizing potential of chemicals. Toxicol In Vitro. 2011; 25:1162–1168. [PubMed: 21669280]

Bauch C, Kolle SN, Ramirez T, Eltze T, Fabian E, Mehling A, Teubner W, van Ravenzwaay B, Landsiedel R. Putting the parts together: combining in vitro methods to test for skin sensitizing potentials. Regul Toxicol Pharmacol. 2012; 63:489–504. DOI: 10.1016/j.yrtph.2012.05.013 [PubMed: 22659254]

Bjorkner B. Contact allergy to 2-hydroxypropyl methacrylate (2-HPMA) in an ultraviolet curable ink. Acta Derm Venereol. 1984; 64:264–267. [PubMed: 6204493]

de Weck AL, Schneider CH, Gutersohn J. The role of penicilloylated protein impurities, penicillin polymers and dimers in penicillin allergy. Int Arch Allergy Appl Immunol. 1968; 33:535–567. [PubMed: 5666915]

Dean JH, Twerdok LE, Tice RR, Sailstad DM, Hattan DG, Stokes WS. ICCVAM evaluation of the murine local lymph node assay: II. Conclusions and recommendations of an independent scientific peer review panel. Regul Toxicol Pharmacol. 2001; 34:258–273. [PubMed: 11754530]

Diaz-Uriarte R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. Bmc Bioinformatics. 2007; 8:328. [PubMed: 17767709]

Emter R, Ellis G, Natsch A. Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. Toxicol Appl Pharmacol. 2010; 245:281–290. [PubMed: 20307559]

Estrada E, Patlewicz G, Chamberlain M, Basketter D, Larbey S. Computer-aided knowledge generation for understanding skin sensitization mechanisms: the TOPS-MODE approach. Chem Res Toxicol. 2003; 16:1226–1235. [PubMed: 14565764]

Gao L, Hu Y, Ni C, Xu Y, Ma L, Yan S, Dou X. Retrospective study of photopatch testing in a Chinese population during a 7-Year period. Dermatitis. 2014; 25:22–26. [PubMed: 24407059]

Gerberick GF, Vassallo JD, Bailey RE, Chaney JG, Morrall SW, Lepoittevin JP. Development of a peptide reactivity assay for screening contact allergens. Toxicol Sci. 2004; 81:332–343. [PubMed: 15254333]

Gerberick GF, Vassallo JD, Foertsch LM, Price BB, Chaney JG, Lepoittevin JP. Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. Toxicol Sci. 2007; 97:417–427. kfm064 [pii]; DOI: 10.1093/toxsci/kfm064 [PubMed: 17400584]

Hao M, Li Y, Wang Y, Zhang S. A classification study of respiratory Syncytial Virus (RSV) inhibitors by variable selection with random forest. Int J Mol Sci. 2011; 12:1259–1280. DOI: 10.3390/ijms12021259 [PubMed: 21541057]

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. Evaluation of combinations of in vitro sensitization test descriptors for the artificial neural network-based risk assessment model of skin sensitization. J Appl Toxicol. 2015; 35:1333–1347. DOI: 10.1002/jat.3105 [PubMed: 25824844]

ICCVAM. The Results of an Independent Peer Review Evaluation Coordinated by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and the National Toxicology Program Center for the Evaluation of Alternative Toxicological Methods (NICEATM). National Institute of Environmental Health Sciences; Research Triangle Park, NC: 1999. The Murine Local Lymph Node Assay: A Test Method for Assessing the Allergic Contact Dermatitis Potential of Chemicals/Compounds.

ICCVAM. ICCVAM Test Method Evaluation Report: Usefulness and Limitations of the Murine Local Lymph Node Assay for Potency Categorization of Chemicals Causing Allergic Contact Dermatitis in Humans. National Institute of Environmental Health Sciences; Research Triangle Park, NC: 2011.

Jaworska J, Dancik Y, Kern P, Gerberick F, Natsch A. Bayesian integrated testing strategy to assess skin sensitization potency: From theory to practice. J Appl Toxicol. 2013; 33:1353–1364. DOI: 10.1002/jat.2869 [PubMed: 23670904]

Jaworska J, Harol A, Kern PS, Frank Gerberick G. Integrating non-animal test information into an adaptive testing strategy - Skin sensitization proof of concept case. ALTEX. 2011; 28:211–225. [PubMed: 21993957]

Jaworska JS, Natsch A, Ryan C, Strickland J, Ashikaga T, Miyazawa M. 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. 2015; 89:2355–2383. DOI: 10.1007/s00204-015-1634-2 [PubMed: 26612363]

Johansson H, Lindstedt M. Prediction of skin sensitizers using alternative methods to animal experimentation. Basic & clinical pharmacology & toxicology. 2014; 115:110–117. DOI: 10.1111/bcpt.12199 [PubMed: 24548737]

Joint Research Centre of the European Union. EURL ECVAM Recommendation on the Direct Peptide Reactivity Assay (DPRA) for Skin Sensitisation Testing. Publications Office of the European Union; Luxembourg: 2013.

Joint Research Centre of the European Union. EURL ECVAM Recommendation on the KeratinoSens™ assay for skin sensitisation testing. Publications Office of the European Union; Luxembourg: 2014.

Jowsey IR, Basketter DA, Westmoreland C, Kimber I. A future approach to measuring relative skin sensitising potency: a proposal. J Appl Toxicol. 2006; 26:341–350. [PubMed: 16773645]

Kimber I, Dearman RJ. Investigation of lymph node cell proliferation as a possible immunological correlate of contact sensitizing potential. Food Chem Toxicol. 1991; 29:125–129. [PubMed: 2010142]

Kimber, I., Dearman, RJ. Contact hypersensitivity: immunological mechanisms. In: Kimber, I., Maurer, T., editors. Toxicology of Contact Hypersensitivity. Taylor and Francis; London: 1996. p. 4-25.

Luechtefeld T, Maertens A, McKim JM, Hartung T, Kleensang A, Sa-Rocha V. Probabilistic hazard assessment for skin sensitization potency by dose-response modeling using feature elimination instead of quantitative structure-activity relationships. J Appl Toxicol. 2015; 35:1361–1371. DOI: 10.1002/jat.3172 [PubMed: 26046447]

Mehling A, Eriksson T, Eltze T, Kolle S, Ramirez T, Teubner W, van Ravenzwaay B, Landsiedel R. Non-animal test methods for predicting skin sensitization potentials. Arch Toxicol. 2012; 86:1273–1295. DOI: 10.1007/s00204-012-0867-6 [PubMed: 22707154]

Natsch A, Emter R, Ellis G. 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. 2009; 107:106–121. [PubMed: 18832184]

Natsch A, Ryan CA, Foertsch L, Emter R, Jaworska J, Gerberick F, Kern P. A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. J Appl Toxicol. 2013; 33:1337–1352. DOI: 10.1002/jat.2868 [PubMed: 23576290]

NIEHS. Request for Information on Alternative Skin Sensitization Test Methods and Testing Strategies and for Comment on ICCVAM's Proposed Activities. Fed Regist. 2013; 78:68076–68077.

Nukada Y, Ashikaga T, Miyazawa M, Hirota M, Sakaguchi H, Sasa H, Nishiyama N. Prediction of skin sensitization potency of chemicals by human Cell Line Activation Test (h-CLAT) and an attempt at classifying skin sensitization potency. Toxicol In Vitro. 2012; 26:1150–1160. DOI: 10.1016/j.tiv.2012.07.001 [PubMed: 22796097]

Nukada Y, Ashikaga T, Sakaguchi H, Sono S, Mugita N, Hirota M, Miyazawa M, Ito Y, Sasa H, Nishiyama N. Predictive performance for human skin sensitizing potential of the human cell line activation test (h-CLAT). Contact Dermatitis. 2011; 65:343–353. [PubMed: 21767275]

Nukada Y, Miyazawa M, Kazutoshi S, Sakaguchi H, Nishiyama N. 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. 2013; 27:609–618. DOI: 10.1016/j.tiv.2012.11.006 [PubMed: 23149339]

OECD. OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. OECD Publishing; Paris: 1992. Test No. 406. Skin Sensitisation.

OECD. Guidance Document for Using the OECD (Q)SAR Application Toolbox to Develop Chemical Categories According to the OECD Guidance on Grouping of Chemicals. OECD Publishing; Paris: 2009. OECD Series on Testing and Assessment No. 102.

OECD. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Part 1: Scientific Assessment. OECD Publishing; Paris: 2012. OECD Series on Testing and Assessment No. 168.

OECD. [November 24, 2014] The OECD QSAR Toolbox. 2014. http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm

OECD. [August 12, 2015] Draft Proposal for a New Test Guideline. *In Vitro* Skin Sensitisation: human Cell Line Activation Test (h-CLAT). 2015a. http://www.oecd.org/env/ehs/testing/Draft-Proposal-for-a-new-Test-Guideline-on-in-vitro-skin-sensitisation-h-CLAT.pdf

OECD. OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. OECD Publsihing; Paris: 2015b. Test No. 442C. *In Chemico* Skin Sensitization: Direct Peptide Reactivity Assay (DPRA).

OECD. OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. OECD Publishing; Paris: 2015c. Test No. 442D. I*n Vitro* Skin Sensitisation: ARE-Nrf2 Luciferase Test Method.

Patlewicz G, Kuseva C, Kesova A, Popova I, Zhechev T, Pavlov T, Roberts DW, Mekenyan O. Towards AOP application--implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization. Regul Toxicol Pharmacol. 2014; 69:529–545. DOI: 10.1016/j.yrtph.2014.06.001 [PubMed: 24928565]

Pirone JR, Smith M, Kleinstreuer NC, Burns TA, Strickland J, Dancik Y, Morris R, Rinckel LA, Casey W, Jaworska JS. Open source software implementation of an integrated testing strategy for skin sensitization potency based on a Bayesian network. ALTEX. 2014; 31:336–340. http://dx.doi.org/10.14573/altex.1310151. [PubMed: 24687303]

Politano VT, Api AM. The Research Institute for Fragrance Materials' human repeated insult patch test protocol. Regul Toxicol Pharmacol. 2008; 52:35–38. [PubMed: 18177987]

R Core Team. R: A Language and Environment for Statistical Computing. Reference Index. R Foundation for Statistical Computing; Vienna, Austria: 2013.

Rovida C, Alepee N, Api AM, Basketter DA, Bois FY, Caloni F, Corsini E, Daneshian M, Eskes C, Ezendam J, Fuchs H, Hayden P, Hegele-Hartung C, Hoffmann S, Hubesch B, Jacobs MN, Jaworska J, Kleensang A, Kleinstreuer N, Lalko J, Landsiedel R, Lebreux F, Luechtefeld T, Locatelli M, Mehling A, Natsch A, Pitchford JW, Prater D, Prieto P, Schepky A, Schuurmann G, Smirnova L, Toole C, van Vliet E, Weisensee D, Hartung T. Integrated Testing Strategies (ITS) for safety assessment. ALTEX. 2015; 32:25–40. http://dx.doi.org/10.14573/altex.1411011. [PubMed: 25413849]

Sailstad DM, Hattan D, Hill RN, Stokes WS. ICCVAM evaluation of the murine local lymph node assay: I. The ICCVAM review process. Regul Toxicol Pharmacol. 2001; 34:249–257. [PubMed: 11754529]

Sakaguchi H, Ryan C, Ovigne JM, Schroeder KR, Ashikaga T. Predicting skin sensitization potential and inter-laboratory reproducibility of a human Cell Line Activation Test (h-CLAT) in the European Cosmetics Association (COLIPA) ring trials. Toxicol In Vitro. 2010; 24:1810–1820. [PubMed: 20510347]

Shen MY, Su BH, Esposito EX, Hopfinger AJ, Tseng YJ. A comprehensive support vector machine binary hERG classification model based on extensive but biased end point hERG data sets. Chem Res Toxicol. 2011; 24:934–949. DOI: 10.1021/tx200099j [PubMed: 21504223]

Strickland J, Zang Q, Kleinstreuer N, Paris M, Lehmann DM, Choksi N, Matheson J, Jacobs A, Lowit A, Allen D, Casey W. Integrated decision strategies for skin sensitization hazard. J Appl Toxicol. 2016; doi: 10.1002/jat.3281

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. 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. 2015; 35(11):1318–1332. DOI: 10.1002/jat.3127 [PubMed: 25820183]

Takenouchi O, Miyazawa M, Saito K, Ashikaga T, Sakaguchi H. Predictive performance of the human Cell Line Activation Test (h-CLAT) for lipophilic chemicals with high octanol-water partition coefficients. J Toxicol Sci. 2013; 38:599–609. [PubMed: 23824015]

Tsujita-Inoue K, Hirota M, Ashikaga T, Atobe T, Kouzuki H, Aiba S. Skin sensitization risk assessment model using artificial neural network analysis of data from multiple in vitro assays. Toxicol In Vitro. 2014; 28:626–639. DOI: 10.1016/j.tiv.2014.01.003 [PubMed: 24444449]

UN. Globally Harmonised Sysem of Classification and Labelling of Chemicals (GHS). United Nations; New York: 2013.

Urbisch D, Mehling A, Guth K, Ramirez T, Honarvar N, Kolle S, Landsiedel R, Jaworska J, Kern PS, Gerberick F, Natsch A, Emter R, Ashikaga T, Miyazawa M, Sakaguchi H. Assessing skin sensitization hazard in mice and men using non-animal test methods. Regul Toxicol Pharmacol. 2015; 71:337–351. DOI: 10.1016/j.yrtph.2014.12.008 [PubMed: 25541156]

van der Veen JW, Rorije E, Emter R, Natsch A, van Loveren H, Ezendam J. Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals. Regul Toxicol Pharmacol. 2014; 69:371–379. DOI: 10.1016/j.yrtph.2014.04.018 [PubMed: 24813372]

Van Och FMM, Slob W, De Jong WH, Vandebriel RJ, Van Loveren H. A quantitative method for assessing the sensitizing potency of low molecular weight chemicals using a local lymph node assay: employment of a regression method that includes determination of the uncertainty margins. Toxicology. 2000; 146:49–59. [PubMed: 10773362]

Varmuza, K., Peter, F. Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press; Boca Raton, Florida: 2009.

White JML, Kullavanijaya P, Duangdeeden I, Zazzeroni R, Gilmour NJ, Basketter DA, McFadden JP. p-Phenylenediamine allergy: the role of Bandrowski's base. Clin Exp Allergy. 2006; 36:1289–1293. [PubMed: 17014438]

Wong CL, Ghassabian S, Smith MT, Lam AL. In vitro methods for hazard assessment of industrial chemicals - opportunities and challenges. Front Pharmacol. 2015; 6:94.doi: 10.3389/fphar.2015.00094 [PubMed: 25999858]

Yang C, Tarkhov A, Marusczyk J, Bienfait B, Gasteiger J, Kleinoeder T, Magdziarz T, Sacher O, Schwab CH, Schwoebel J, Terfloth L, Arvidson K, Richard A, Worth A, Rathman J. New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. J Chem Inf Model. 2015; 55:510–528. DOI: 10.1021/ci500667v [PubMed: 25647539]

Zang Q, Rotroff DM, Judson RS. Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure-activity relationship and machine learning methods. J Chem Inf Model. 2013; 53:3244–3261. DOI: 10.1021/ci400527b [PubMed: 24279462]

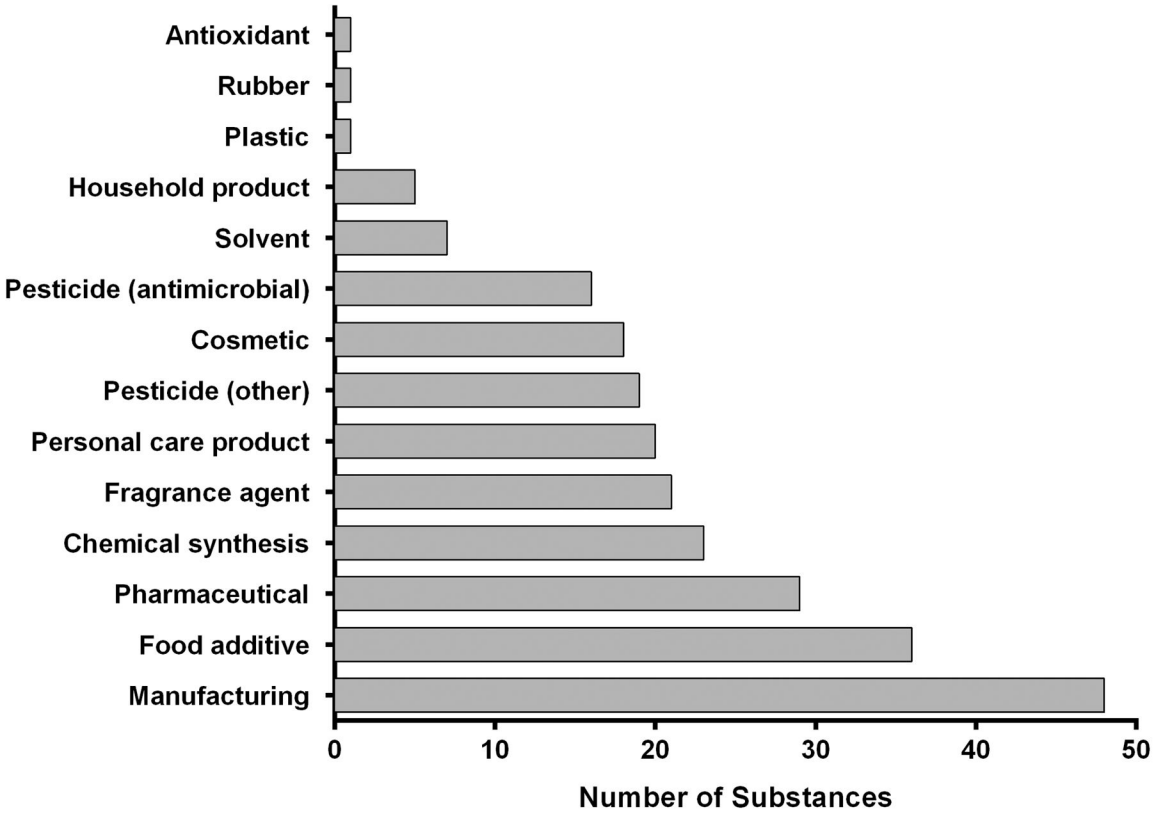

Figure 1.

Product uses for 96 substances in the database. Total number of substances exceeds 96 because most substances were associated with more than one product use.

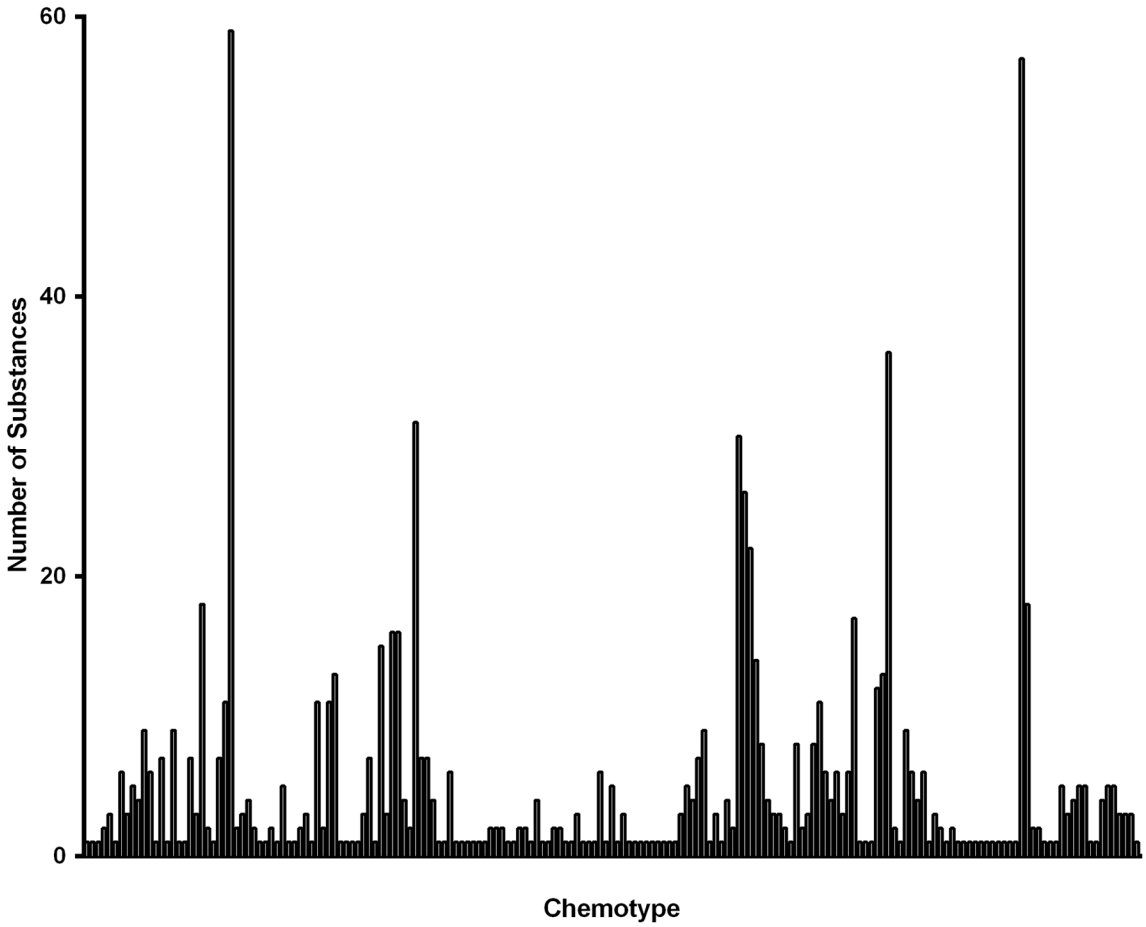

Figure 2.
Frequency of appearance of 183 chemotypes for the 96 substance database. Height of bars represent the number of substances that included each of 183 chemotypes.

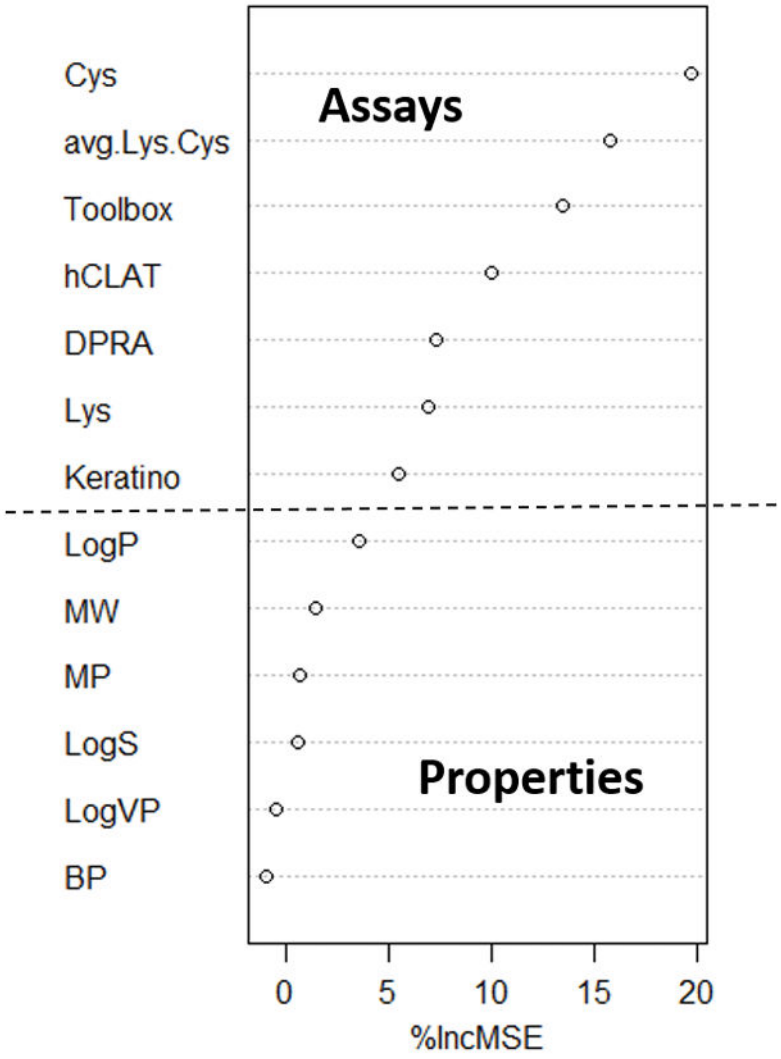

Figure 3.

Ranking of variable importance by random forest algorithm. Avg.Lys.Cys, average percent depletion for lysine and cysteine peptides from the DPRA; BP, boiling point; Cys, average percent depletion of cysteine peptide from the DPRA; DPRA, direct peptide reactivity assay binary result; hCLAT, human cell line activation test; %IncMSE, percent increase in mean squared error; Keratino, KeratinoSens assay; log P, log octanol:water partition coefficient; log S, log water solubility; log VP, log vapor pressure; Lys, average percent depletion of lysine peptide from the DPRA; MP, melting point; MW, molecular weight; Toolbox, read-across prediction from QSAR Toolbox.

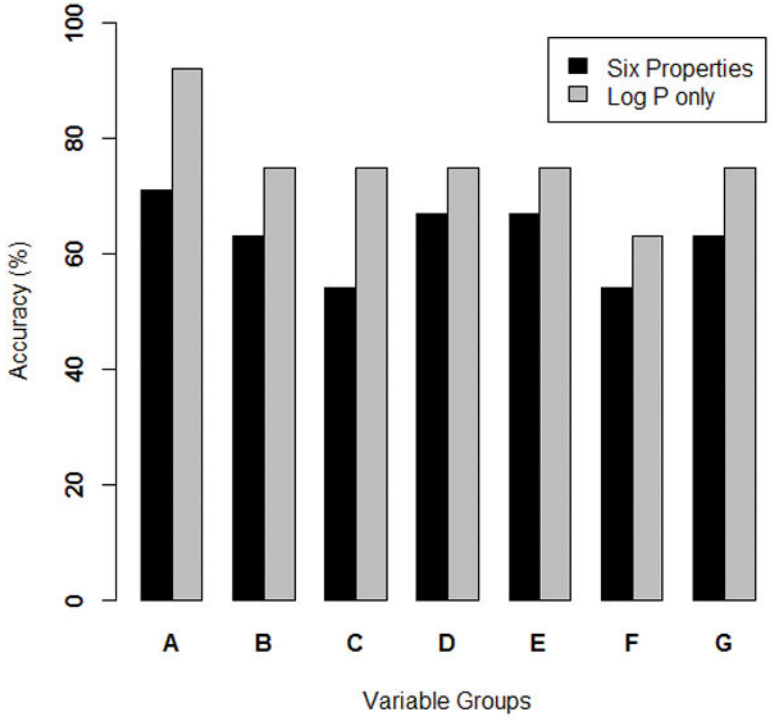

Figure 4.

Logistic regression models: comparison of accuracy for test set for variable groups containing either six physicochemical properties or log P. Test set contained nine nonsensitizers and 15 sensitizers. Variable Groups A–G are defined in Table 3. Log P, log octanol:water partition coefficient.

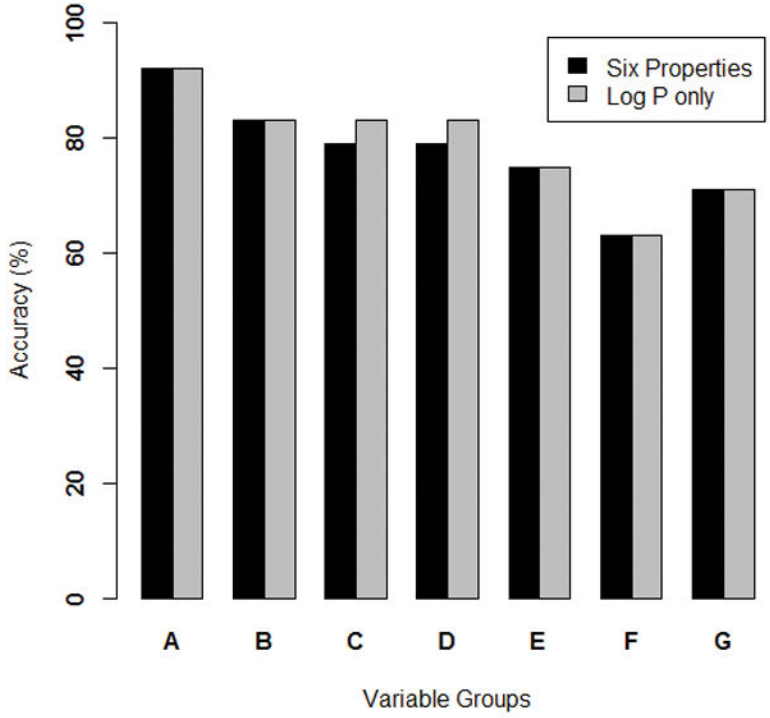

Figure 5.

Support vector machine models: comparison of accuracy for test set for variable groups containing either six physicochemical properties or log P. Test set contained nine nonsensitizers and 15 sensitizers. Variable Groups A–G are defined in Table 3. Log P, log octanol:water partition coefficient.

Table 1

Data sources

Test Method	Reference
DPRA	Bauch *et al.* (2011)
	Bauch *et al.* (2012)
	Gerberick *et al.* (2004)
	Gerberick *et al.* (2007)
	Jaworska *et al.* (2011)
	Jaworska *et al.*(2013)
	Joint Research Centre of the European Union (2013)
	Natsch *et al.* (2013)
	Nukada *et al.* (2013)
KeratinoSens	Ball *et al.* (2011)
	Bauch *et al.* (2011)
	Bauch *et al.* (2012)
	Emter *et al.* (2010)
	Joint Research Centre of the European Union (2014)
	Natsch *et al.* (2013)
h-CLAT	Ashikaga *et al.* (2010)
	Bauch *et al.* (2011)
	Bauch *et al.* (2012)
	Nukada *et al.* (2011)
	Nukada *et al.* (2012)
	Nukada *et al.* (2013)
	Sakaguchi *et al.* (2010)
	Takenouchi *et al.* (2013)
LLNA	Basketter *et al.* (1996) and Estrada *et al.* (2003) (xylene)
	Basketter and Kimber (2006) (diphenylcyclopropenone, maleic anhydride, and propyl gallate)
	NICEATM LLNA database
	Van Och *et al.* (2000) (phthalic anhydride)

DPRA, direct peptide reactivity assay; h-CLAT, human cell line activation test; LLNA, murine local lymph node assay; NICEATM, National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods.

Table 2

Data types and ranges of 13 input variables

Name	Description	Data Type	Value Range
h-CLAT	h-CLAT majority call	Categorical	0/1
DPRA	DPRA majority call	Categorical	0/1
KeratinoSens	KeratinoSens majority call	Categorical	0/1
Toolbox	Read-across prediction using QSAR Toolbox	Categorical	0/1
Avg.Lys.Cys	Average lysine and cysteine depletion measurements from DPRA	Numeric	0 – 95.0
Lys	Average lysine depletion from DPRA	Numeric	0 – 91.0
Cys	Average cysteine depletion from DPRA	Numeric	0 – 100
Log P	Octanol:water partition coefficient	Numeric	−8.28 – 6.46[a]
Log S	Water solubility (mol/L)	Numeric	−6.39 – 1.92[a]
Log VP	Vapor pressure (mm Hg)	Numeric	−28.47 – 5.89[a]
MP	Melting point (°C)	Numeric	−148.5 – 288.0
BP	Boiling point (°C)	Numeric	−19.1 – 932.2
MW	Molecular weight (g/mol)	Numeric	30.03 – 581.57

Abbreviations: DPRA = direct peptide reactivity assay; h-CLAT = human cell line activation test.

[a]Range for base 10 logarithm of these measurements.

Table 3

Variable groups used to build models for predicting human skin sensitization hazard

Variable Group	Input Variables				
	Avg.Lys.Cys from DPRA	KeratinoSens	h-CLAT	QSAR Toolbox	Log P or Six Physicochemical Properties
A	X	X	X	X	X
B	-	-	X	X	X
C	-	X	-	X	X
D	X	-	-	X	X
E	-	-	X	-	X
F	-	X	-	-	X
G	X	-	-	-	X
H	-	-	-	-	X
I	X	X	X	X	-
J	-	X	X	X	-
K	X	-	X	X	-
L	X	X	-	X	-

Avg.Lys.Cys, average depletion for lysine and cysteine peptides from the DPRA; Cys, average depletion of cysteine peptide from the DPRA; DPRA, direct peptide reactivity assay; h-CLAT, human cell line activation test; log P, log octanol:water partition coefficient; log S, log water solubility; log VP, log vapor pressure; Lys, average depletion of lysine peptide from the DPRA; Toolbox, read-across using QSAR Toolbox.

Xs denote the input variables included in each variable group. Dashes indicate that the variable is not included in the designated group. The six physicochemical properties are log octanol:water partition coefficient; log water solubility; log vapor pressure, melting point, boiling point, and molecular weight.

Table 4

Performance of LR and SVM models: test and training sets

Variable Group	Dataset	Accuracy (%)		Sensitivity (%)		Specificity (%)	
		LR	SVM	LR	SVM	LR	SVM
A. Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox + Log P	Training	94	94	94	94	95	95
	Test	92	92	87	93	100	89
B. h-CLAT + Toolbox + Log P	Training	86	88	84	84	91	95
	Test	75	83	73	80	78	89
C. KeratinoSens + Toolbox + Log P	Training	82	85	84	86	76	81
	Test	75	83	73	87	78	78
D. Avg.Lys.Cys + Toolbox + Log P	Training	86	89	88	84	81	100
	Test	75	83	73	73	78	100
E. h-CLAT + Log P	Training	82	82	88	88	67	67
	Test	75	75	80	80	67	67
F. KeratinoSens + Log P	Training	78	78	82	82	67	67
	Test	63	63	60	60	67	67
G. Avg.Lys.Cys + Log P	Training	86	85	82	80	95	95
	Test	75	71	67	67	89	78
H. Log P + Log S + Log VP + MP +BP + MW	Training	81	82	84	80	71	86
	Test	54	58	67	67	33	44
I. Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox	Training	93	93	92	92	95	95
	Test	92	92	93	93	89	89
J. h-CLAT + KeratinoSens + Toolbox	Training	85	90	90	90	71	91
	Test	75	79	87	80	56	78
K. Avg.Lys.Cys + h-CLAT + Toolbox	Training	93	93	92	92	95	95
	Test	92	92	93	93	89	89
L. Avg.Lys.Cys + KeratinoSens + Toolbox	Training	85	90	86	88	81	95
	Test	75	88	73	87	78	89

Avg.Lys.Cys, average depletion for lysine and cysteine peptides from the direct peptide reactivity assay; BP, boiling point; h-CLAT, human cell line activation test; log P, log octanol:water partition coefficient; log S, log water solubility; log VP, log vapor pressure; LR, logistic regression; MP, melting point; MW, molecular weight; SVM, support vector machine; Toolbox, read-across using QSAR Toolbox.

The only variable group that contains all six physicochemical properties is Variable Group H. The only physicochemical property contained in the other variable groups is log P. The three variable groups yielding the highest accuracy are bolded. The training set of 72 substances contains 51 human sensitizers and 21 nonsensitizers. The test set of 24 substances contains 15 human sensitizers and 9 nonsensitizers.

Table 5

Performance of leave-one-out cross validation for LR and SVM models for 96 substances

Variable Group	Accuracy (%)		Sensitivity (%)		Specificity (%)	
	LR	SVM	LR	SVM	LR	SVM
A. Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox + Log P	91	94	89	95	93	90
I. Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox	92	94	92	95	90	90
K. Avg.Lys.Cys + h-CLAT + Toolbox	92	94	92	92	90	97

Avg.Lys.Cys, average depletion for lysine and cysteine peptides from the direct peptide reactivity assay; h-CLAT = human cell line activation test; log P, log octanol:water partition coefficient; LR, logistic regression; SVM, support vector machine; Toolbox, read-across using QSAR Toolbox.

The dataset of 96 substances contains 66 human sensitizers and 30 nonsensitizers.

Table 6

Misclassified substances[a] for the six LR and SVM models with the highest performance

Test Method or Model[b]	Training Set						Test Set		
	2-Methoxy-4-methylphenol	Sulfanil-amide	Streptomycin sulfate	Penicillin G	Benzo-caine	α-Amyl-cinnamal-dehyde	Coumarin	Pentachloro-phenol	Lilial
Human Reference Result	NEG	POS	POS	POS	POS	POS	POS	NEG	POS
DPRA	POS	NEG	NEG	POS	POS	NEG	NEG	POS	POS
KeratinoSens	NEG	NEG	NEG	NEG	POS	NEG	POS	NEG	NEG
h-CLAT	POS	NEG	NEG	POS	POS	POS	NEG	POS	POS
Toolbox	POS	NEG	POS	NEG	NEG	POS	NEG	POS	POS
A. Avg.Lys.Cys + h-CLAT + KeratinoSens + Toolbox + Log P	**POS**	**NEG**	POS	**NEG**	POS/**NEG**[b]	**NEG**/POS[c]	**NEG**	NEG/**POS**[b]	**NEG**/POS[c]
I. Avg.Lys.Cys + h- CLAT + KeratinoSens + Toolbox	**POS**	**NEG**	**NEG**	**NEG**	**NEG**	POS	**NEG**	**POS**	POS
K. Avg.Lys.Cys + h-CLAT + Toolbox	**POS**	**NEG**	**NEG**	**NEG**	**NEG**	POS	**NEG**	**POS**	POS

Avg.Lys.Cys, average depletion for lysine and cysteine peptides from the direct peptide reactivity assay; h-CLAT, human cell line activation test; log P, log octanol:water partition coefficient; LR, logistic regression; NEG, negative; POS, positive; SVM, support vector machine; Toolbox, read-across using QSAR Toolbox.

[a] Misclassifications, which disagree with the human outcomes, are bolded.

[b] Correctly classified in the LR model but misclassified by the SVM model.

[c] Misclassified by the LR model but correctly classified by the SVM model

Table 7

Performance of individual methods and the LLNA for predicting human skin sensitization hazard compared with machine learning approaches

Method	Data Set[a]	Accuracy (%)	Sensitivity (%)	Specificity (%)
Machine learning models[b]	Training	93–94	92–94	95
	Test	92	87–93	89–100
	All	93–94	92–94	94–96
h-CLAT	Training	82	88	67
	Test	79	87	67
	All	81	88	67
DPRA	Training	88	88	86
	Test	71	73	67
	All	83	85	80
KeratinoSens	Training	78	82	67
	Test	63	60	67
	All	74	77	67
Toolbox	Training	81	82	76
	Test	71	73	67
	All	78	80	73
LLNA	Training	83	90	67
	Test	88	100	67
	All	84	92	67
Test Battery 1 (≥ 1 positive = positive)	Training	79	98	33
	Test	75	100	33
	All	78	99	33
Test Battery 2 (≥ 2 positives = positive)	Training	89	96	71
	Test	75	87	56
	All	85	94	67

DPRA, direct peptide reactivity assay; h-CLAT, human cell line activation test; LLNA, murine local lymph node assay; Toolbox, read-across using QSAR Toolbox.

[a]Test set contains 15 sensitizers and nine nonsensitizers. The training set contains 51 sensitizers and 21 nonsensitizers. “All” is the entire dataset of 96 substances: 66 sensitizers and 30 nonsensitizers.

[b]Models with the highest performance from Table 4: support vector machine and logistic regression models with Variable Groups A, I, and K.