Prediction of Skin Sensitization Potency Using Machine Learning Approaches

J Strickland[1], Q Zang[1], M Paris[1], DM Lehmann[2], S Bell[1], D Allen[1], J Matheson[3], A Jacobs[4], W Casey[5], N Kleinstreuer[5]

[1]ILS, RTP, NC, USA; [2]EPA/ORD/NHEERL, RTP, NC, USA; [3]CPSC, Rockville, MD, USA; [4]FDA/CDER, Silver Spring, MD, USA; [5]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA

Introduction

- The replacement of animal use in testing for hazard classification and labeling of skin sensitizers is a top priority for the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM).
- Skin sensitization is a complex process and thus a successful non-animal approach is likely to require integrating data from several methods.
- We previously developed machine learning models that integrate in vitro and in chemico skin sensitization data with physicochemical properties to identify potential skin sensitizers without using animal data (Strickland et al. 2016 a, b).
- Because some regulatory agencies require that sensitizers be classified into potency categories, we have now developed machine learning models using similar input variables to predict skin sensitization potency results for both the murine local lymph node assay (LLNA) and human skin sensitization tests.

Study Design

- The National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) and the ICCVAM Skin Sensitization Working Group compiled publicly available data from non-animal (in chemico and in vitro) tests for 120 substances.
- NICEATM used previously compiled high-quality in vivo test data for most of these substances (Basketter et al. 2014; ICCVAM 2011, NICEATM LLNA Database at http://ntp.niehs.nih.gov/go/40500).
  - All of the substances were evaluated for skin sensitization potency in the LLNA.
  - Human skin sensitization potency test data were also available for 87 of the 120 substances.
- The LLNA and human data were used to classify substances into potency categories according to the Globally Harmonized System of Classification and Labeling of Chemicals (GHS) (United Nations 2015) (**Table 1**).

Table 1. GHS Potency Categories

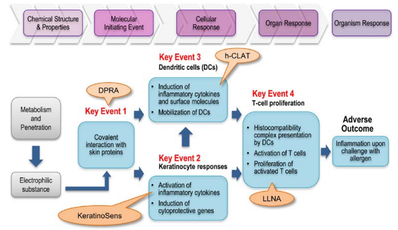
GHS Category	LLNA EC3	Human Threshold
1A (strong)	≤ 2%	≤ 500 µg/cm² skin area
1B (other than strong –"weak"[a])	> 2%	> 500 µg/cm² skin area
Nonsensitizer	Unclassified	Unclassified

EC3 = estimated test substance concentration that produces a stimulation index of 3, the threshold for a substance to be considered a sensitizer in the LLNA; GHS = Globally Harmonized System of Classification and Labeling; LLNA = murine local lymph node assay.

[a] For simplicity, we refer to Category 1B sensitizers as "weak"; this term is not used in the GHS.

- The in chemico and in vitro data were generated using methods recommended for use in a weight-of-evidence approach (OECD 2015a, b; OECD 2016). These methods align with key events in the adverse outcome pathway for skin sensitization (OECD 2012) as shown in **Figure 1**:
  - The direct peptide reactivity assay (DPRA) measures a substance's ability to covalently bind protein; activity (**Table 2**) is measured as depletion of specific peptides in the reaction mix.
  - The KeratinoSens™ (Givaudan) assay assesses the ability of a substance to activate cytokines and induce cytoprotective genes in keratinocytes; activity is measured by induction of luciferase activity in a reporter gene.
  - The human cell line activation test (h-CLAT) assesses the ability of a substance to activate and mobilize dendritic cells in the skin; activity is measured by increase of CD86 and CD54 cell surface markers.
- Data for six physicochemical properties that may impact skin absorption were also collected.
- Ranges for the in chemico and in vitro endpoints and the physicochemical properties are shown in **Table 2**.
- To predict potency outcomes, the in chemico, in vitro, and physicochemical property data were integrated using four machine learning approaches:
  - Support vector machine (SVM)
  - Classification and regression tree
  - Linear discriminant analysis
  - Logistic regression

Figure 1. Adverse Outcome Pathway for Skin Sensitization Produced by Substances That Covalently Bind to Proteins



Abbreviations: DPRA = direct peptide reactivity assay; h-CLAT = human cell line activation test; LLNA = murine local lymph node assay.

Figure adapted from OECD (2012).

Table 2. Descriptions and Ranges of Input Variables

Abbreviated Name	Description	Range
DPRA	Average lysine and cysteine peptide depletion (%)	0 – 95
h-CLAT	Minimum induction threshold (smallest value for EC200 for CD54 and EC150 for CD86) in µg/mL	0.54 – 2001
KeratinoSens	Concentration producing a 1.5-fold induction of luciferase controlled by the antioxidant response element in µM	0.50 – 2001
Log P	Octanol:water partition coefficient	-8.28 – 6.46
Log S (mol/L)	Water solubility	-6.39 – 1.92[a]
Log VP (mm Hg)	Vapor pressure	-28.47 – 5.89[a]
MP (°C)	Melting point	-148.50 – 288.00
BP (°C)	Boiling point	-19.10 – 932.20
MW (g/mol)	Molecular weight	30.03 – 581.57

DPRA = direct peptide reactivity assay; EC150 = estimated concentration inducing a 150% increase for CD86; EC200 = estimated concentration inducing a 200% increase for CD54; h-CLAT = human cell line activation test.

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

Model Building and Evaluation

- The LLNA dataset (120 substances) and human dataset (87 substances) were each divided into training and test sets in an approximate proportion of 75% to 25%, respectively, for building and evaluating the predictive models (**Figure 2**).
- Using the nine input variables in **Table 2**, we defined four variable groups (**Table 3**).
- Each machine learning model was constructed by applying one of the four machine learning approaches and one of the four variable groups in **Table 3** to the training set substances.
- The machine learning models were then evaluated by assessing the accuracy of their predictions of the GHS classifications for LLNA and human potency of the test set substances.
- Performance of the models was also determined by calculating accuracy using a leave-one-out cross-validation (LOOCV) procedure on each dataset (**Figure 3**).
  - *n* - 1 substances from the complete set of *n* substances were used as the training data for building the model and the remaining substance was used for testing the model.
  - The cross-validation process was repeated *n* times with each of the substances used exactly once as the test set.
  - Accuracy was calculated by averaging individual values over the *n* runs.
- Two strategies were applied to model strong potency (1A), weak potency (1B), and nonsensitizers:
  - Strategy A, a one-tier strategy, simultaneously predicted all three classes of substances (**Figure 4a**).
  - Strategy B was a two-tier strategy (**Figure 4b**).
    - Tier 1 predicted a binary classification of sensitizer or nonsensitizer using previously developed SVM models to predict LLNA and human skin sensitization hazard (Strickland et al. 2016 a,b). The variables used in these models are shown in **Table 4**.
      - The LLNA hazard model predicted test set classifications with accuracy of 96%, sensitivity of 95%, and specificity of 100% (Strickland et al. 2016a).
      - The human hazard model predicted test set classifications with accuracy of 92%, sensitivity of 93%, and specificity of 89% (Strickland et al. 2016b).
    - Tier 2 predicted the sensitizers from Tier 1 as either GHS Category 1A or 1B (**Figure 4b**)
  - The performance of the two strategies was assessed by comparing the accuracies of the best performing model (machine learning approach + variable group) for each strategy.

Figure 2. Proportion of Category 1A and 1B Sensitizers and Nonsensitizers for Training and Test Sets

Figure 2a. LLNA Training Set (n=94)

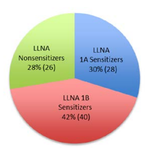


Figure 2b. LLNA Test Set (n=26)

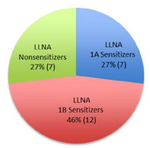


Figure 2c. Human Training Set (n=63)

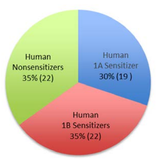


Figure 2d. Human Test Set (n=24)

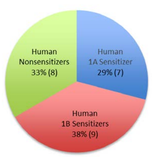


Table 3. Variable Groups Used to Build Models for Predicting Skin Sensitization Potency

Variable Group	Variables Included				
	DPRA	KeratinoSens	h-CLAT	Log P	Six Physicochemical Properties
1	■	■	■		
2					■
3	■	■	■		■
4	■	■	■	■	

Abbreviations: DPRA = direct peptide reactivity assay; h-CLAT = human cell line activation test; LLNA = murine local lymph node assay; log P = log octanol:water partition coefficient.

Colored shading indicates variables included in Variable Groups 1-4.

Figure 3. Leave-one-out Cross-validation Procedure

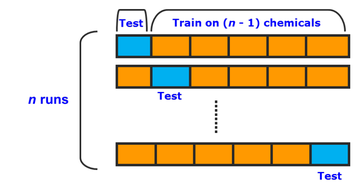


Table 4. Variables Used for Skin Sensitization Hazard Models in Tier 1 of Strategy B

SVM Model	Tier 1 Variables for Strategy B					
	DPRA (continuous)	KeratinoSens (binary)	h-CLAT (binary)	Read-across (binary)	Log P (continuous)	Six Physicochemical Properties (continuous)
LLNA			■	■		■
Human	■	■	■	■	■	

Abbreviations: DPRA = direct peptide reactivity assay; h-CLAT = human cell line activation test; LLNA = murine local lymph node assay; log P = log octanol:water partition coefficient.

From Strickland et al. 2016 a, b.

Figure 4. Classification Strategies for Modeling Skin Sensitization Potency

Figure 4a. Strategy A

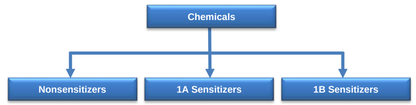
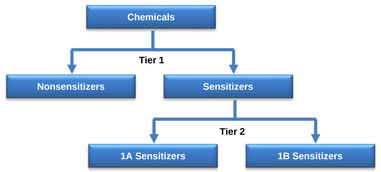


Figure 4b. Strategy B



Strategy A simultaneously models three categories of potency. Strategy B models in two tiers: Tier 1 predicts sensitizers and nonsensitizers and Tier 2 predicts 1A and 1B sensitizers.

Acknowledgements

The Intramural Research Program of the National Institute of Environmental Health Sciences (NIEHS) supported this poster. Technical support was provided by ILS under NIEHS contract HHSN273201500010C.

The views expressed above do not necessarily represent the official positions of any Federal agency. Since the poster was written as part of the official duties of the authors, it can be freely copied.

Results

Performance of the Machine Learning Approaches

- When all three potency categories were predicted simultaneously using Strategy A (**Figure 4a**), models using the SVM approach predicted potency classifications of the test sets with higher accuracy than models using other machine learning approaches: 77% for the LLNA dataset and 71% for the human dataset (**Figure 5a**).
- Tier 1 of Strategy B (**Figure 4b**) used previously developed SVM models for classification of sensitizers and nonsensitizers. Therefore, multiple machine learning approaches were only evaluated for Tier 2.
- Using Strategy B, the SVM approach also predicted GHS 1A and 1B hazard classifications of the test sets with highest accuracy: 89% for the LLNA dataset and 81% for the human dataset (**Figure 5b**).

Performance of the Variable Groups

- Because SVM was the best performing machine learning approach, it was used to evaluate the variable groups for prediction of potency category. LOOCV was used for this effort.
  - For Strategy A, Variable Groups 3 and 4 (**Table 3**) yielded the highest accuracy for predicting hazard classifications for the LLNA and human datasets (**Figure 6a**).
    - Variable Group 3 predicted hazard classifications with 78% accuracy for the LLNA dataset and 75% accuracy for the human dataset.
    - Variable Group 4 predicted hazard classifications with 73% accuracy for the LLNA dataset and 75% accuracy for the human dataset.
    - Variable Group 2 had the lowest accuracy for classifying substances using Strategy A: 61% for the LLNA dataset and 60% for the human dataset (**Figure 6a**).
  - For Strategy B, only Variable Groups 3 and 4 were evaluated for Tier 2 because they yielded the best performance with SVM when the four machine learning approaches were evaluated. Comparing Variable Groups 3 and 4 (**Table 3**) would indicate whether all physicochemical properties are needed (Variable Group 3) or whether log P only was sufficient (Variable Group 4) (**Figure 6b**).
    - Variable Group 3 predicted classification of 1A and 1B sensitizers in the LLNA dataset with an accuracy of 87%, while Variable Group 4 had an accuracy of 84%.
    - Variable Group 3 predicted classification of 1A and 1B sensitizers in the human dataset with an accuracy of 81%, while Variable Group 4 had an accuracy of 79%.

Performance Statistics for Strategy A (One Tier) and Strategy B (Two Tiers)

- The best performing models for both LLNA and human datasets applied SVM to Variable Group 3, which include all nine in chemico, in vitro, and physicochemical variables.
  - The two-tiered approach in Strategy B outperformed the one-tiered approach in Strategy A for both datasets (**Table 5**).
  - The accuracy of the LLNA for predicting human potency was 69% (60/87).

Table 5. Classification Performance of Strategy A and Strategy B using SVM and Variable Group 3

Dataset	Strategy	Classification Rate								Overall Accuracy
		1A (Strong)		1B (Weak)			Nonsensitizers			
		Correct	Under	Over	Correct	Under	Correct	Over		
LLNA	A	83 ± 12% (29/35)	17 ± 12% (6/35)	17 ± 10% (9/52)	69 ± 12% (36/52)	14 ± 9% (7/52)	85 ± 12% (28/33)	15 ± 12% (5/33)		78 ± 7% (93/120)
	B	91 ± 9% (32/35)	9 ± 9% (3/35)	15 ± 10% (8/52)	81 ± 11% (42/52)	4 ± 5% (2/52)	97 ± 6% (32/33)	3 ± 6% (1/33)		88 ± 6% (106/120)
Human	A	85 ± 14% (22/26)	15 ± 14% (4/26)	26 ± 15% (8/31)	61 ± 17% (19/31)	13 ± 12% (4/31)	80 ± 14% (24/30)	20 ± 14% (6/30)		75 ± 9% (65/87)
	B	85 ± 14% (22/26)	15 ± 14% (4/26)	22 ± 15% (7/31)	65 ± 17% (20/31)	13 ± 12% (4/31)	93 ± 9% (28/30)	7 ± 9% (2/30)		81 ± 8% (70/87)
	LLNA	65 ± 18% (17/26)	35 ± 18% (9/26)	16 ± 13% (5/31)	74 ± 15% (23/31)	10 ± 11% (3/31)	67 ± 17% (20/30)	33 ± 17% (10/30)		69 ± 10% (60/87)

Abbreviations: LLNA, murine local lymph node assay; SVM, support vector machine.

Leave-one-out cross-validation accuracy results shown with 95% confidence limits.

Figure 5. Performance of Machine Learning Approaches for LLNA and Human Skin Sensitization Potency

Figure 5a. Strategy A

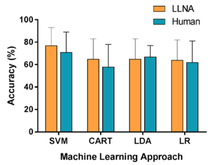


Figure 5b. Tier 2 of Strategy B

Abbreviations: CART = classification and regression tree; LDA = linear discriminant analysis; LR = logistic regression; SVM = support vector machine.

Accuracy of hazard classification predictions for test set of 26 substances for LLNA potency and 24 substances for human potency. Error bars show 95% confidence limits. Accuracy is shown for the best performing variable group for each machine learning approach.

Figure 6. Comparison of Variable Groups

Figure 6a. Strategy A

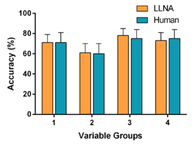
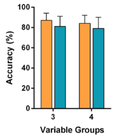


Figure 6b. Tier 2 of Strategy B



Abbreviation: LLNA = murine local lymph node assay; SVM = support vector machine.

Leave-one-out cross validation accuracy for the variable groups in **Table 2**. Error bars show 95% confidence limits.

Misclassified Substances

- The best performing models in Strategy B (**Table 5**) misclassified substances in both datasets.
  - Fourteen substances were misclassified in the LLNA dataset:
    - Tier 1 misclassified three sensitizers (two weak [1B] and one strong [1A]) as nonsensitizers and one nonsensitizer as a sensitizer.
    - Tier 2 misclassified eight weak sensitizers as strong sensitizers and two strong sensitizers as weak sensitizers.
  - Seventeen substances were misclassified in the human dataset:
    - Tier 1 misclassified four sensitizers (weak [1B]) as nonsensitizers and two nonsensitizers as sensitizers.
    - Tier 2 misclassified seven weak sensitizers as strong sensitizers, and four strong sensitizers as weak sensitizers.

Conclusions

- The SVM models performed better than other machine learning approaches in predicting LLNA or human skin sensitization potency.
- The Strategy B two-tiered models performed better than the Strategy A one-tiered models.
  - The best performing Strategy A model predicted hazard classifications for the LLNA dataset with an accuracy of 78% while the best performing Strategy B model had accuracy of 88%.
  - The best performing Strategy A model had predicted hazard classification for the human dataset with an accuracy of 75% while the best performing Strategy B model had accuracy of 81%.
- Variable groups (**Table 3**) that included physicochemical properties with the in chemico and in vitro methods (Variable Groups 3 and 4) performed better than the variable group with only the in chemico and in vitro methods (Variable Group 1).
  - Variable groups including in chemico and in vitro methods (Variable Groups 1, 3, and 4) performed better than the physicochemical properties alone (Variable Group 2).
  - Using all six physicochemical properties in combination with the in chemico and in vitro methods (Variable Group 3) yielded only slightly better predictive performance than using only log P (Variable Group 4). Thus, little predictive power is lost when the only physicochemical property available is log P.
- The LLNA predicts human potency categories with an accuracy of 69% for the 87 substances in the human dataset. The 81% accuracy rate achieved in this study (**Table 5**) for the human dataset using the SVM machine learning approach with Strategy B and variable sets that included physicochemical properties and data from in chemico and in vitro methods suggests that computational approaches may be useful for predicting human skin sensitization potency.

References

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. 2014. Categorization of chemicals according to their relative human skin sensitizing potency. Dermatitis 25:11-21.

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

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

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

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

OECD. 2016. Test No. 442E. *In Vitro* Skin Sensitisation: human Cell Line Activation Test (h-CLAT). In *OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects*. OECD Publishing: Paris.

Strickland J, Zang Q, Kleinstreuer N, Paris M, Lehmann DM, Choksi N, Matheson J, Jacobs A, Lowit A, Allen D, Casey W. 2016a. Integrated decision strategies for skin sensitization hazard. J Appl Toxicol 36(9): 1150-62.

Strickland J, Zang Q, Paris M, Lehmann DM, Kleinstreuer N, Allen D, Choksi N, Matheson J, Jacobs A, Casey W. 2016b. Multivariate models for prediction of human skin sensitization hazard. J Appl Toxicol DOI 10.1002/jat.3366.

United Nations. 2015. *Globally Harmonized System for Classification and Labelling of Chemicals*. United Nations: New York