Machine Learning Approaches for Predicting Human Skin Sensitization Hazard

J Strickland[1], Q Zang[1], M Paris[1], N Kleinstreuer[1], DM Lehmann[2], D Allen[1], N Choksi[1], J Matheson[3], A Jacobs[4], A Lowit[5], W Casey[6]

[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]EPA/OCSPP/OPP, Washington, DC, USA; [6]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA

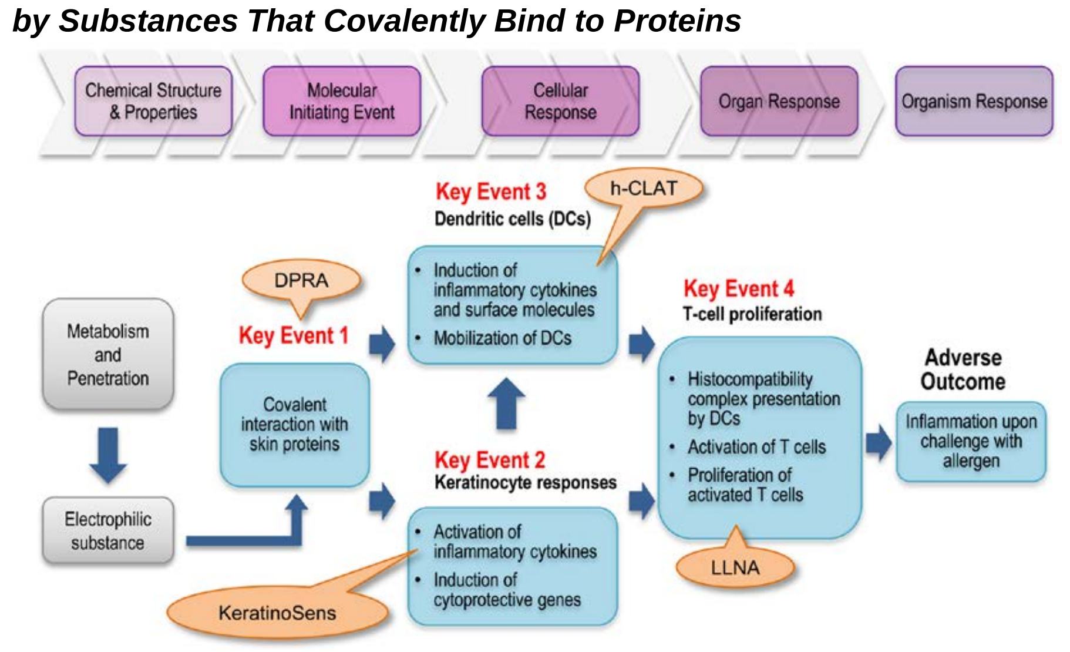
Introduction

- Allergic contact dermatitis (ACD) is an adverse skin reaction, characterized by localized redness, swelling, blistering, or itching, that can develop after repeated direct contact with a skin sensitizer.
- U.S. regulatory agencies establish hazard categories to determine appropriate labeling to warn consumers and workers of potential skin sensitization hazards. Historically, these categories are assigned based on the results of animal tests.
- Since its inception, the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) has given a high priority to replacing, reducing, and refining the use of animals for skin sensitization testing.
- Skin sensitization is a complex process, and it is likely that no single non-animal test can replace animal use for this testing. A more promising approach involves integrating data from several non-animal methods using an integrated decision strategy (IDS).
- This poster describes an IDS developed by ICCVAM that integrates non-animal skin sensitization data and physicochemical properties to identify potential human skin sensitizers.

Study Design

- The National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) and the ICCVAM Skin Sensitization Working Group compiled non-animal (in chemico and in vitro) test data and human skin sensitization hazard data (sensitizer or nonsensitizer) for 96 substances.
- The in chemico and in vitro data were obtained from methods recommended for use in a weight-of-evidence approach (OECD 2015a,b,c). The methods align with key events in the adverse outcome pathway for skin sensitization (OECD 2012) (**Figure 1**).
  - The direct peptide reactivity assay (DPRA) measures covalent interaction with proteins (Key Event 1).
  - The KeratinoSens™ (Givaudan) assay measures activation of genes, controlled by the antioxidant response element (ARE), that protect against oxidative stress in keratinocytes (Key Event 2).
  - The human cell line activation test (h-CLAT) measures activation and mobilization of dendritic cells in the skin (Key Event 3).
  - For DPRA, we evaluated both binary (sensitizer/nonsensitizer) and quantitative (average cysteine depletion, average lysine depletion, and average cysteine and lysine depletion) results, whereas for KeratinoSens and h-CLAT, we evaluated only binary results because quantitative data were available for only a small number of chemicals.

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


Chemical Structure & Properties
Molecular Initiating Event
Cellular Response
Organ Response
Organism Response

Metabolism and Penetration
Electrophilic substance

DPRA
Key Event 1
Covalent interaction with skin proteins

Key Event 3
Dendritic cells (DCs)
h-CLAT
- Induction of inflammatory cytokines and surface molecules
- Mobilization of DCs

Key Event 2
Keratinocyte responses
KeratinoSens
- Activation of inflammatory cytokines
- Induction of cytoprotective genes

Key Event 4
T-cell proliferation
- Histocompatibility complex presentation by DCs
- Activation of T cells
- Proliferation of activated T cells
LLNA

Adverse Outcome
Inflammation upon challenge with allergen

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

Figure adapted from OECD (2012). While the figure shows the assays aligned only to specific key events, a positive response in any of these assays requires completion of all prior events in the pathway.

- Additional data compiled for the 96 substances included:
  - Six physicochemical properties that may impact skin absorption (**Table 1**)
  - Binary in silico predictions of skin sensitization hazard produced using a read-across approach with QSAR Toolbox (OECD 2014), a software package developed by the Organisation for Economic Co-operation and Development (OECD).

Table 1. Ranges of Physicochemical Properties for 96 Substances

Physicochemical Property	Range of Values
Octanol:water partition coefficient	-8.28 to 6.46[a]
Water solubility (M)	-6.39 to 1.92[a]
Vapor pressure (mm Hg)	-28.47 to 5.89[a]
Melting point (°C)	-148.5 to 288.0
Boiling point (°C)	-19.1 to 932.2
Molecular weight (g/mol)	30.03 to 581.57

[a] Range for $\log_{10}$ of these measurements.

- To predict human outcomes, the in chemico, in vitro, and in silico data and physicochemical properties were integrated using a test battery approach and two machine learning approaches, logistic regression (LR) and support vector machine (SVM).

Definition of Training and Test Sets

- The database of 96 substances included 69% (66/96) human sensitizers and 31% (30/96) human nonsensitizers (**Figure 2a**).
- We divided the substances into test and training sets with similar characteristics (**Figures 2b** and **2c**). We used the training set to build models to predict human hazard and the test set to test the models.

Figure 2. Proportion of Human Sensitizers and Nonsensitizers

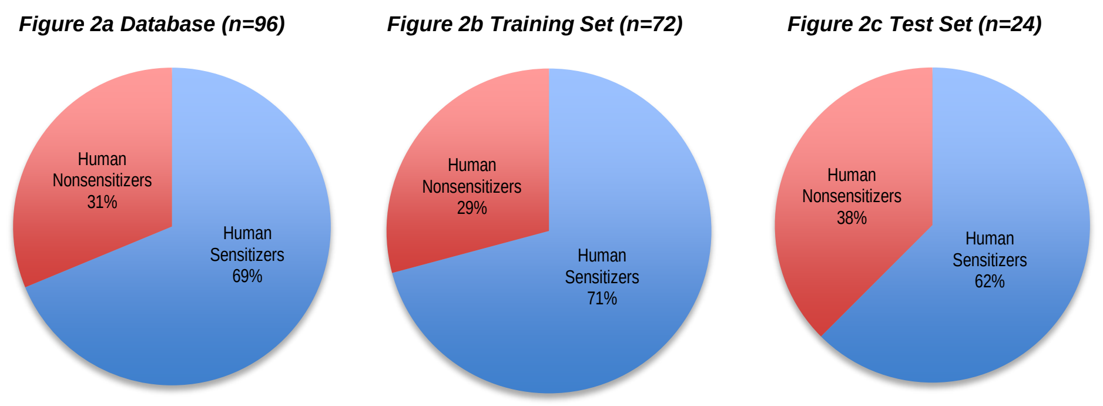


Figure 2a Database (n=96)

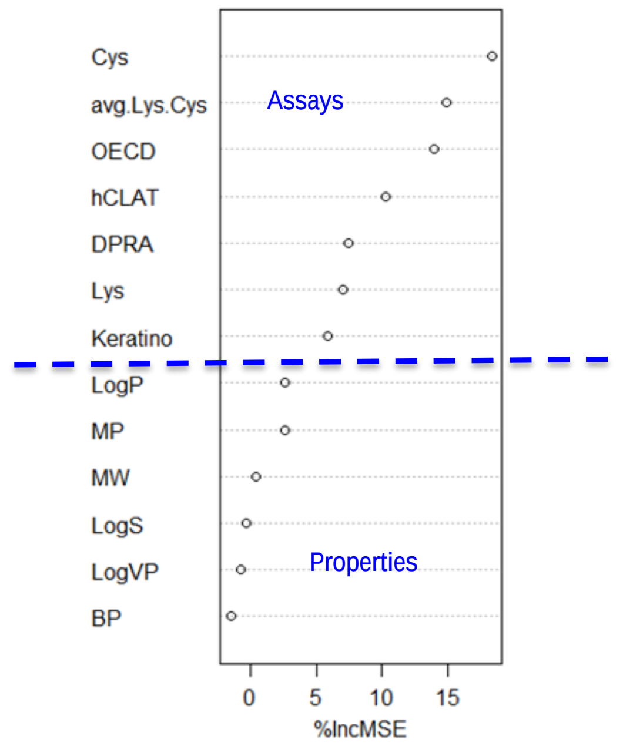
Figure 2b Training Set (n=72)

Figure 2c Test Set (n=24)

Analysis of Variable Importance

- A total of 13 variables (non-animal test data and physicochemical characteristics) were available for predicting human skin sensitization hazard. A random forest analysis (Diaz-Uriarte 2007; Hao et al. 2011) was conducted to assess the relative importance of the variables (**Figure 3**).

Figure 3. Ranking of Variable Importance by Random Forest Algorithm



Abbreviations: 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; LogP = log octanol:water partition coefficient; LogS = log water solubility; LogVP = log vapor pressure; Lys = average percent depletion of lysine peptide from the DPRA; MP = melting point; MW = molecular weight; OECD = read-across prediction from OECD QSAR Toolbox.

Model Building

- We defined twelve variable groups, A–L in **Table 2,** as follows:
  - Groups A–G used different combinations of the non-animal methods with either log P, the most important physicochemical property according to the random forest analysis, or all six physicochemical properties.
  - Group H used log P or the six physicochemical properties only.
  - Groups I-L used different combinations of the non-animal methods without any of the physicochemical properties.
- The machine learning models were constructed by applying one of two approaches, LR or SVM, to the data for training set of 72 substances for each of the variable groups A–L (**Table 2**). The models were then tested by assessing performance for predicting human skin sensitization hazard using data for the test set of 24 substances.

Table 2. Variable Groups Used to Build Models for Predicting Human Skin Sensitization Hazard

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

Abbreviations: Avg.Lys.Cys = average depletion for lysine and cysteine peptides (DPRA); DPRA = direct peptide reactivity assay; h-CLAT = human cell line activation test; log P = log octanol:water partition coefficient; QSAR Toolbox = read-across prediction from OECD QSAR Toolbox.

Results

Comparison of the Machine Learning Models with Other Approaches

- In addition to the machine learning models, we evaluated two test battery approaches for prediction of human skin sensitization hazard:
  - **Test Battery 1:** If any test method classified the substance as a sensitizer (i.e., positive), the substance is classified as a sensitizer.
  - **Test Battery 2:** If two or more tests classified the substance as a sensitizer (i.e., positive), the substance is classified as a sensitizer.
- For comparison with the results from the machine learning methods, **Table 3** shows performance statistics for prediction of human skin sensitization hazard for the test set of 24 substances using (1) the individual non-animal methods, (2) the two test battery approaches, and (3) the LLNA.

Table 3. Performance of Individual Methods and the LLNA for Predicting Human Skin Sensitization Hazard for the Test Set of 24 Substances[a]

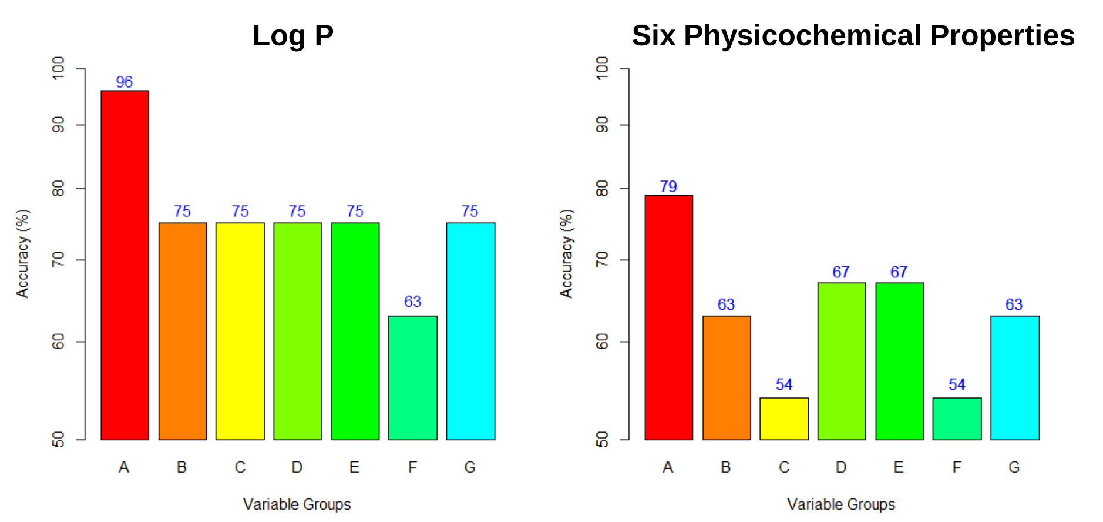
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
DPRA	71	73	67
KeratinoSens	63	60	67
h-CLAT	75	80	67
Toolbox	71	73	67
Battery 1 (≥ **1 method positive)**	75	100	33
Battery 2 (≥ 2 methods positive)	75	87	56
LLNA	88	100	67

Abbreviations: 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] Nine nonsensitizers and 15 sensitizers.

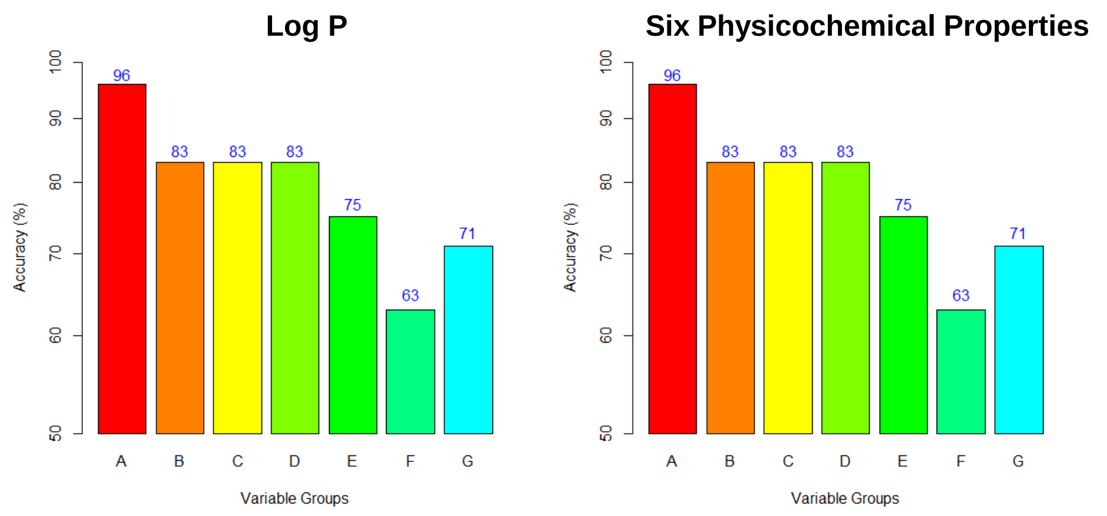
- **Figures 4** and **5** show the accuracies of the machine learning models using Variable Groups A–G listed in **Table 2.**
  - For the LR models, all seven variable groups with log P produced higher accuracy for the test set of 24 substances than those with all six physicochemical properties (**Figure 4**).
  - For the SVM models, all seven variable groups with log P produced equivalent or higher accuracy for the test set compared to those with all six physicochemical properties (**Figure 5**).

Figure 4. Comparison of Accuracy for Log P vs Six Physicochemical Properties Using Logistic Regression[a]


Log P

Six Physicochemical Properties

Figure 5. Comparison of Accuracy for Log P vs Six Physicochemical Properties Using Support Vector Machines[a]


Log P

Six Physicochemical Properties

[a] Data for test set of nine nonsensitizers and 15 sensitizers. Variable groups are defined in **Table 2**.

- Because all variable groups with log P had equivalent or higher accuracy than those with all six physicochemical properties, subsequent models included only log P if physicochemical properties were included.

Acknowledgements

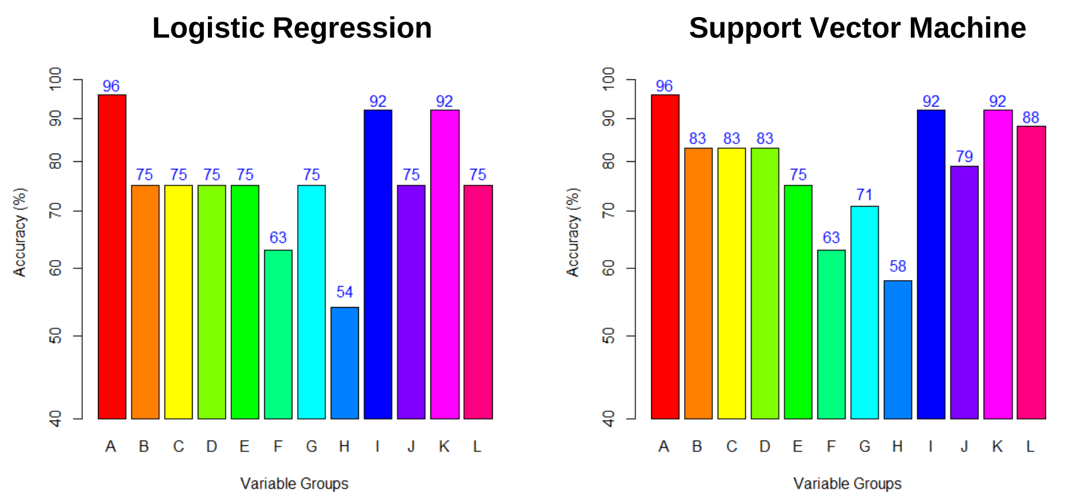
The Intramural Research Program of the National Institute of Environmental Health Sciences (NIEHS) supported this poster. Technical support was provided by ILS under NIEHS contract HHSN273201500010C.

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.

Performance Statistics for the Machine Learning Models

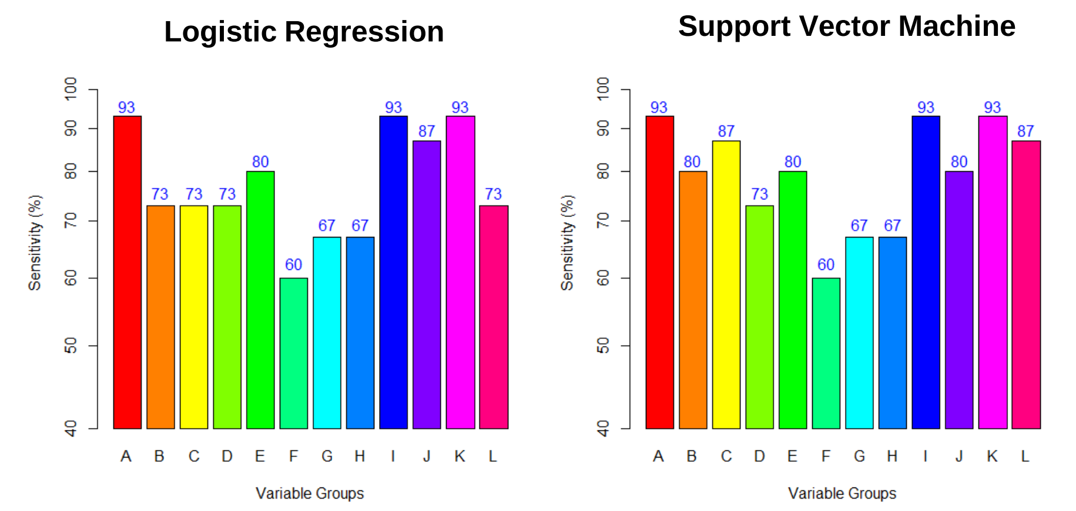
- **Figures 6-8** compare the accuracy, sensitivity, and specificity of the SVM and LR models for 12 variable groups (A–L in **Table 2**).
  - The three variable groups with the highest performance for the test set were the same for the LR and SVM models. For all of these, accuracy was ≥ 92% (**Figure 6**), sensitivity was 93% (**Figure 7**), and specificity was ≥ 89% (**Figure 8**).
    - Variable Group A: all non-animal tests + log P
    - Variable Group I: Avg.Lys.Cys + KeratinoSens + h-CLAT + QSAR Toolbox
    - Variable Group K: Avg.Lys.Cys + h-CLAT + QSAR Toolbox
  - The training set performance of these variable groups were also the same for the LR and SVM models (**Table 4**).
  - The variable group with the worst performance was H, which included only the six physicochemical properties. Accuracy was ≤ 58% (**Figure 6**), sensitivity was 67% (**Figure 7**), and specificity was ≤ 44% (**Figure 8**).

Figure 6. Accuracy for Machine Learning Models[a]


Logistic Regression

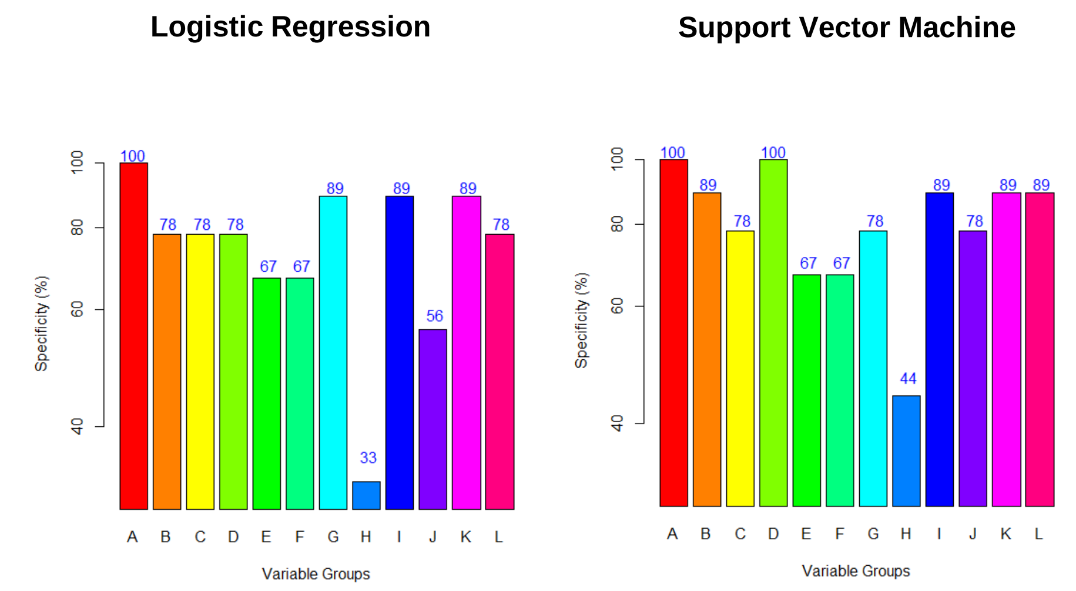
Support Vector Machine

Figure 7. Sensitivity for Machine Learning Models


Logistic Regression

Support Vector Machine

Figure 8. Specificity for Machine Learning Models[a]


Logistic Regression

Support Vector Machine

[a] Data for test set of nine nonsensitizers and 15 sensitizers. Variable groups are defined in **Table 2**.

Table 4. Highest Performing Machine Learning Models: Performance Statistics for Test and Training Sets[a]

Variable Group[a]	Data Set[b]	LR Accuracy (%)	SVM Accuracy (%)	LR Sensitivity (%)	SVM Sensitivity (%)	LR Specificity (%)	SVM Specificity (%)
All non-animal methods + Log P (A)	Training	99	99	98	98	100	100
	Test	96	96	93	93	100	100
Avg.Lys.Cys + KeratinoSens + h-CLAT + Toolbox (I)	Training	93	93	92	92	95	95
	Test	92	92	93	93	89	89
Avg.Lys.Cys + h-CLAT + Toolbox (K)	Training	93	93	92	92	95	95
	Test	92	92	93	93	89	89

Abbreviations: 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 machines; Toolbox = read-across using QSAR Toolbox.

[a] Table reports statistics from the best performing variable groups for each machine learning approach.

[b] The training set of 72 substances contains 51 human sensitizers and 21 nonsensitizers. The test set of 24 substances contains 15 human sensitizers and nine nonsensitizers (**Figure 2**).

Misclassified Substances

- The five training set substances and two test set substances misclassified by any of the three LR and SVM models with the highest performance are shown in **Table 5**. The results from the individual non-animal methods are shown for reference.
  - In the training set, there were four false negatives and one false positive.
  - Although the in chemico and in vitro methods do not consistently yield correct classifications for prehaptens, which must be oxidized to produce skin sensitization, or prohaptens, which must be metabolized to produce skin sensitization, the machine learning methods overcome this limitation. None of the four false negatives were prehaptens or prohaptens.
    - The two prehaptens, 10 prohaptens, and two pre/prohaptens in the training set were correctly classified.
  - In the test set there was one false negative and one false positive. The false negative was not a pre- or pro-hapten.
    - The one prehapten, four prohaptens, and one pre/prohapten in the training set were correctly classified, which overcomes a limitation of the in chemico and in vitro methods.
- KeratinoSens correctly classified more misclassified substances (4) than any of the other test methods (1-2 substances).

Table 5. Misclassified Substances for the Highest Performing Machine Learning Models[a]

Test Method or Model	Training Set					Test Set	
	2-Methoxy-4-methylphenol	Streptomycin sulfate	Penicillin G	Sulfan-ilamide	Benzocaine	Pentachloro-phenol	Coumarin
Human Reference Result	NEG	POS	POS	POS	POS	NEG	POS
DPRA	POS	NEG	POS	NEG	POS	POS	NEG
KeratinoSens	NEG	NEG	NEG	NEG	POS	NEG	POS
h-CLAT	POS	NEG	POS	NEG	POS	POS	NEG
Toolbox	POS	POS	NEG	NEG	NEG	POS	NEG
All non-animal methods + Log P (A)	NEG	POS	POS	NEG	POS	NEG	NEG
Avg.Lys.Cys + KeratinoSens + h-CLAT + Toolbox (I)	POS	NEG	NEG	NEG	NEG	POS	NEG
Avg.Lys.Cys + h-CLAT + Toolbox (K)	POS	NEG	NEG	NEG	NEG	POS	NEG

Abbreviations: Avg.Lys.Cys = average depletion for lysine and cysteine peptides from the DPRA; DPRA = direct peptide reactivity assay; log P = log octanol:water partition coefficient; h-CLAT = human cell line activation test; NEG = negative; POS = positive; Toolbox = read-across using QSAR Toolbox.

[a] Shaded cells indicate results that are discordant with the reference values.

Conclusions

- The LR and SVM machine learning models performed better in predicting human skin sensitization hazard (accuracy ≥ 92% for the highest performing models) than individual non-animal methods (accuracy ≤ 79%) and test batteries (accuracy = 75%). The machine learning models also achieved a better balance between sensitivity and specificity.
- Models using only log P performed better than analogous models with all six physicochemical properties.
- The highest performing LR and SVM models had the same performance using the same three variable groups. Accuracies were ≥92% for the training and test sets.
  - The models using the variable group with all of the non-animal test methods (Avg.Lys.Cys from DPRA + KeratinoSens + h-CLAT + QSAR Toolbox) + log P achieved the highest accuracy (99% for the training set and 96% for the test set) (**Table 4**).
  - Two variable sets used a combination of two or three non-animal test methods (Avg.Lys.Cys from DPRA, h-CLAT, or KeratinoSens) with QSAR Toolbox and without any physicochemical properties. The models that used Avg.Lys.Cys + h-CLAT + QSAR Toolbox would save resources because they require only two non-animal test methods.
  - The LR and SVM models with these three variable groups correctly classified all prehaptens and prohaptens in the dataset.
- The in chemico/in vitro/in silico methods were more informative than physicochemical properties (**Figures 6-8).**
- Future work will explore the use of continuous variables for h-CLAT and KeratinoSens for the development of models to predict skin sensitization potency.

References

Diaz-Uriarte R. 2007. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics 8: 328.

Hao M, Li Y, Wang Y, Zhang S. 2011. A classification study of respiratory Syncytial Virus (RSV) inhibitors by variable selection with random forest. Int J of Mol Sci 12(2): 1259-1280.

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. Paris:OECD Publishing. Available: http://www.oecd.org/env/ehs/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm.

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

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