

ICCVAM Integrated Decision Strategy for Skin Sensitization

J Matheson¹, Q Zang², J Strickland², N Kleinstreuer², D Allen², A Lowit³, A Jacobs⁴, W Casey⁵

¹CPSC, Rockville, MD, USA; ²ILS/NICEATM, RTP, NC, USA; ³EPA/OPP, Washington, DC, USA; ⁴FDA/CDER, Silver Spring, MD, USA; ⁵NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA

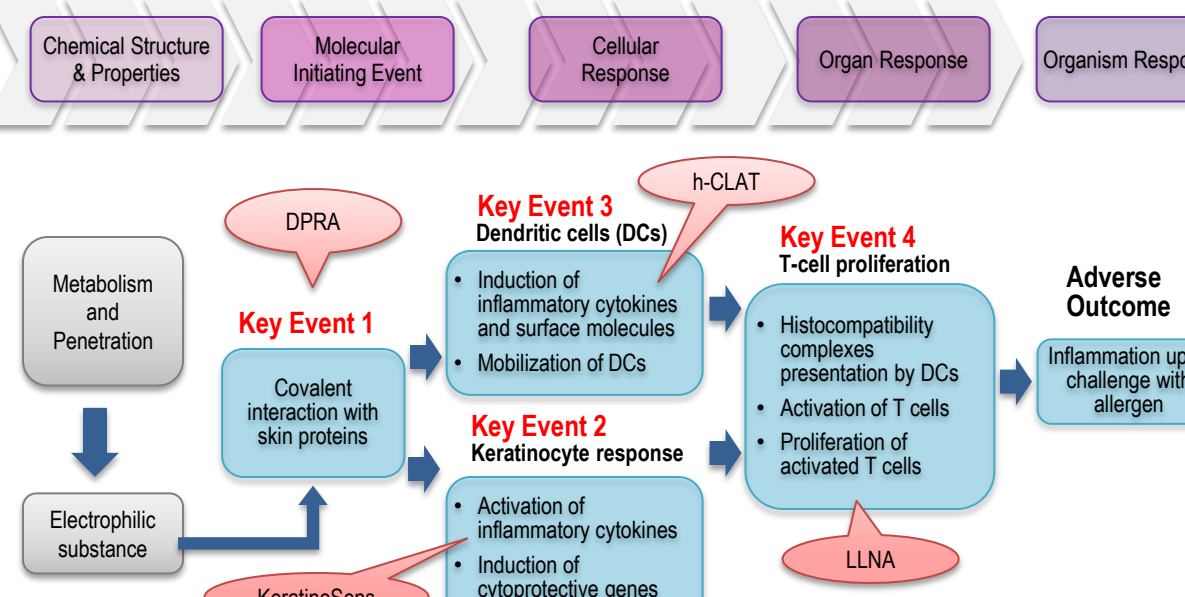
Abstract

One of the top priorities currently being addressed by ICCVAM is the identification and validation of non-animal alternatives for skin sensitization testing. Although it is a complex process, the key biological events leading to skin sensitization have been well characterized in an adverse outcome pathway (AOP) proposed by OECD. Accordingly, ICCVAM is working to develop an integrated decision strategy based on the OECD AOP using *in vitro*, *in chemico*, and *in silico* information on skin sensitization. Data were compiled for 120 chemicals tested in the local lymph node assay (LLNA), direct peptide reactivity assay (DPRA), human cell line activation test (h-CLAT), and KeratinoSens assay. Data for six physicochemical parameters (octanol:water partition coefficient, water solubility, vapor pressure, molecular weight, melting point, and boiling point) were collected and OECD QSAR Toolbox predictions for skin sensitization were calculated for each chemical. These data were combined into a variety of potential integrated decision strategies to predict LLNA outcomes using a training set of 94 chemicals and an external test set of 26 chemicals. Fifty-four models were built using multiple combinations of machine learning approaches and predictor variables. The seven models with the highest accuracy for predicting LLNA outcomes used the support vector machine (SVM) approach with different combinations of the predictor variables. The performance statistics of the SVM models were higher than any of the *in vitro*, *in chemico*, or *in silico* tests alone and higher than a simple test battery approach using these methods. These data suggest that computational approaches are promising tools to effectively integrate data sources to identify potential skin sensitizers without testing animals. (Data in poster abstract have been updated to reflect the most recent analyses.)

Introduction

- Allergic contact dermatitis (ACD) is a skin reaction, characterized by localized redness, swelling, blistering, or itching, that can develop after repeated direct contact with a skin allergen.
- U.S. regulatory agencies establish hazard categories to determine appropriate labeling to warn consumers and workers of potential skin sensitization hazards. Data used to assign substances to appropriate hazard categories are generated using 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 existing non-animal skin sensitization data and physicochemical properties to identify potential skin sensitizers.

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. 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.

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 murine local lymph node assay (LLNA) outcomes (sensitizer or nonsensitizer) for 120 substances.
 - The *in chemico* and *in vitro* data were obtained from methods recommended by the European Union Reference Laboratory for Alternatives to Animal Testing (JRC 2013, JRC 2014, JRC 2015). The methods align with 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 IDS used both categorical (sensitizer/nonsensitizer) and quantitative (average cysteine depletion, average lysine depletion, and average cysteine and lysine depletion) results.
 - The KeratinoSens™ (Givaudan) assay measures activation of cytoprotective genes in keratinocytes (Key Event 2). The IDS used categorical results.
 - The human cell line activation test (h-CLAT) measures activation and mobilization of dendritic cells in the skin (Key Event 3). The IDS used categorical results.
- Additional data compiled included:
 - Six physicochemical properties that may impact skin absorption (Table 1).
 - Categorical *in silico* predictions of skin sensitization hazard produced using a read-across approach with QSAR Toolbox (OECD 2014) (see Strickland et al. at Poster Board 109 [Abstract 422]).

Study Design (cont'd)

Table 1. Ranges of Physicochemical Properties for 120 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
Boiling point (°C)	-19.1 to 932.2
Molecular weight (g/mol)	30.03 to 581.57

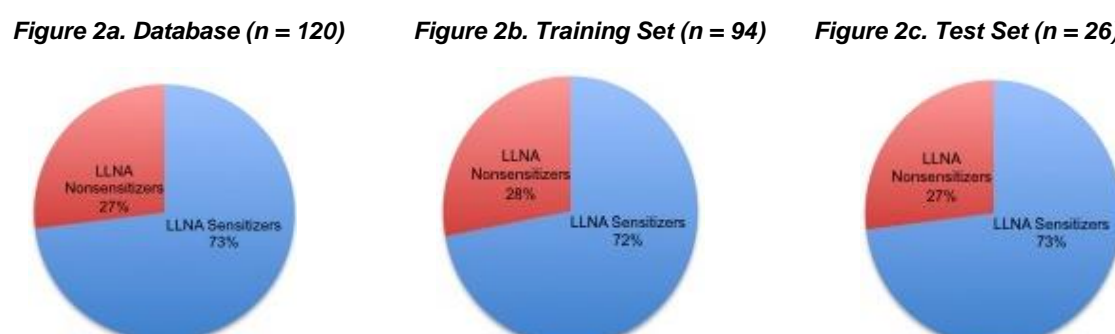
^a Range for base 10 logarithm of these measurements.

- To predict LLNA outcomes, the *in chemico*, *in vitro*, and *in silico* data and physicochemical properties were integrated using the following methods:
 - Artificial neural network (ANN)
 - Naïve Bayes algorithm
 - Classification and regression tree (CART)
 - Linear discriminant analysis (LDA)
 - Logistic regression (LR)
 - Support vector machines (SVM)
 - Test battery approach

Definition of Training and Test Sets

- The database of 120 substances included 73% LLNA sensitizers and 27% LLNA nonsensitizers (Figure 2a).
- The substances were divided into test and training sets with similar characteristics (Figures 2b and 2c). The training set was used to build models to predict LLNA outcome and the test set was used to test the models.

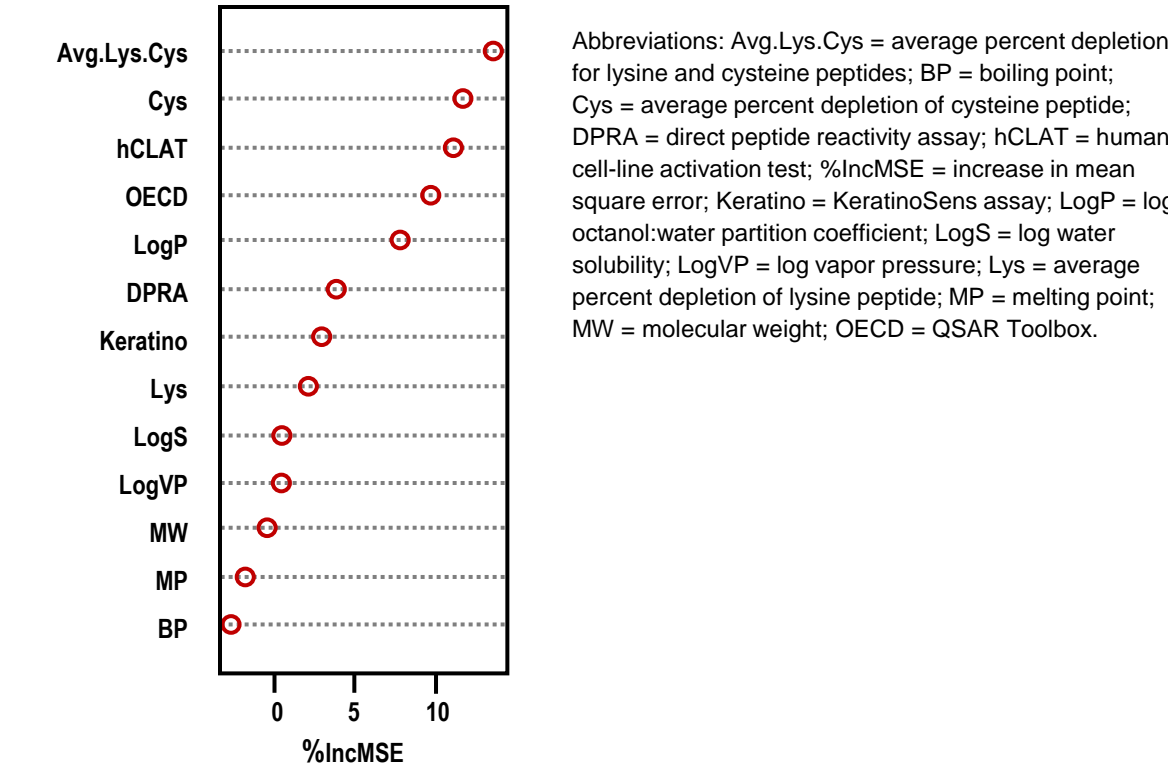
Figure 2a. Proportion of LLNA Sensitizers and Nonsensitizers



Analysis of Variable Importance

- A random forest analysis was conducted to assess the relative importance of the 13 variables (non-animal test data and physicochemical characteristics) for predicting LLNA outcome (Figure 3).
 - The importance of each independent variable to the model is assessed by measuring the increase in the prediction error (mean squared error) when each variable, in turn, is replaced with random noise while the others are left unchanged.
- The most important variables were average lysine and cysteine depletion (Avg.Lys.Cys) and average cysteine depletion (Cys) from DPRA, the h-CLAT classification, and the QSAR Toolbox prediction.

Figure 3. Ranking of Variable Importance by Random Forest Algorithm



Abbreviations: Avg.Lys.Cys = average percent depletion for lysine and cysteine peptides; BP = boiling point; Cys = average percent depletion of cysteine peptide; DPRA = direct peptide reactivity assay; hCLAT = human cell-line activation test; %IncMSE = increase in mean square 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; MP = melting point; MW = molecular weight; OECD = QSAR Toolbox.

Model Building

- Six variable sets (Table 2) were defined using all (Set A) or subsets (Sets B–F) of the 13 variables. The machine learning approaches were then applied to the training set of 94 substances using the six variable sets.

Table 2. Six Variable Sets Used to Build Models for Predicting LLNA Outcome

Variable	Set A	Set B	Set C	Set D	Set E	Set F
DPRA	X		X	X		
KeratinoSens	X		X	X	X	
h-CLAT	X		X	X	X	
Toolbox	X		X	X	X	X
Lys		X				
Cys	X	X		X		
Avg.Lys.Cys	X	X		X	X	X
Log P	X	X		X	X	X
Log S	X	X		X	X	X
Log VP	X	X		X	X	X
Melting Point	X	X			X	X
Boiling Point	X	X			X	X
Molecular Weight	X	X			X	X

Abbreviations: Avg.Lys.Cys = average depletion for lysine and cysteine peptides; Cys = average depletion of cysteine peptide; 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; Log S = log water solubility; Log VP = log vapor pressure; Lys = average depletion of lysine peptide; Toolbox = QSAR Toolbox.

Results

- Table 3 shows performance statistics for the ability of the individual *in chemico*, *in vitro*, and *in silico* methods and two test battery approaches to predict LLNA outcomes for the entire 120-substance database.
 - 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.
- Of the individual test methods, h-CLAT had the highest sensitivity (84%), QSAR Toolbox had the highest specificity (76%), and DPRA had the highest accuracy (79%).
- Test Battery 1 had similar accuracy and higher sensitivity to the individual test methods, but much lower specificity.
- Test Battery 2 had higher accuracy and sensitivity than the individual test methods, but similar specificity.

Table 3. Performance of Individual Methods for Predicting LLNA Outcomes for 120 Substances^a

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)
DPRA	83	70	79
KeratinoSens	76	64	73
h-CLAT	84	64	78
Toolbox	77	76	77
Battery 1	98	30	79
Battery 2	91	64	83

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

^a Full substance set: 87 LLNA sensitizers and 33 LLNA nonsensitizers. **Bolded red text** shows the highest values for each performance statistic.

- Performance statistics for the ability of the machine learning methods to predict LLNA outcomes are shown in Table 4.
- Based on overall accuracy for predicting LLNA outcomes, the modeling approaches ranked as follows: SVM > ANN > LR > LDA > CART = NB.

Table 4. Performance of Machine Learning Methods Predicting LLNA Outcomes for Training and Test Sets^a

Approach	Variable Set ^b	Data Set ^c	Sensitivity (%)	Specificity (%)	Accuracy (%)
SVM	A	E Training	99	96	98
SVM	A	E Test	90	100	92
ANN	D	Training	93	89	93
ANN	D	Test	90	86	89
LR	A	Training	93	85	90
LR	A	Test	84	100	89
LDA	A	Training	93	85	90
LDA	A	Test	84	86	85
CART	A B D E F	Training	87	89	87
CART	A B D E F	Test	74	86	77
NB	F	Training	87	89	87
NB	F	Test	74	86	77

Abbreviations: ANN = artificial neural network; NB = naïve Bayes algorithm; CART = classification and regression tree; LDA = linear discriminant analysis; LLNA = murine local lymph node assay; LR = logistic regression; SVM = support vector machines.

^a Table reports statistics from the best performing variable set(s) for each machine learning approach. **Bolded red text** shows the best performing machine learning approach and variable sets.

^b Color codes match Table 2, which contains descriptions of the variable sets.

^c The training set of 94 substances contains 68 LLNA sensitizers and 26 LLNA nonsensitizers. The test set of 26 substances contains 19 LLNA sensitizers and 7 LLNA nonsensitizers (Figure 2).

Additional Analyses

- After identifying the machine learning approach that produced the highest accuracy (Table 4), data inputs were optimized by testing additional SVM models with 18 different variable sets (Table 5).
- In these analyses, DPRA results were represented only by the average lysine and cysteine peptide depletion values (Avg.Lys.Cys), because this measurement was more highly correlated to LLNA outcomes than DPRA measures (average cysteine peptide depletion, average lysine peptide depletion, and categorical DPRA result).
- Table 5 provides the performance statistics for these analyses.
 - The variable set that included h-CLAT, QSAR Toolbox, and the six physicochemical properties (Variable Set 1) achieved the highest average accuracy for the test and training sets (97%).
 - The variable set with only physicochemical properties (Variable Set 8) produced the lowest accuracy, 73% for both test and training sets.
 - Of all the models tested, the seven SVM models with the highest accuracies (average of the training and test set statistics) were:
 - h-CLAT + QSAR Toolbox + 6 physicochemical properties (97%) (Variable Set 1)
 - DPRA + KeratinoSens + h-CLAT + QSAR Toolbox + Lys + Cys + Ave.Lys.Cys + 6 physicochemical properties (95%) (Variable Set A, Table 4)
 - KeratinoSens + h-CLAT + QSAR Toolbox + Ave.Lys.Cys + 6 physicochemical properties (95%) (Variable Set E, Table 4)
 - KeratinoSens + QSAR Toolbox + Ave.Lys.Cys + 6 physicochemical properties (94%) (Variable Set 2)
 - KeratinoSens + h-CLAT + Ave.Lys.Cys + 6 physicochemical properties (92%) (Variable Set 3)
 - h-CLAT + QSAR Toolbox + Ave.Lys.Cys + 6 physicochemical properties (92%) (Variable Set 4)
 - KeratinoSens + h-CLAT + QSAR Toolbox + 6 physicochemical properties (92%) (Variable Set 5)

Table 5. Classification Results for SVM Models with 18 Additional Variable Combinations^a

Set No.	Variable Set	Data Set ^b	Sensitivity (%)	Specificity (%)	Accuracy (%)
1	h-CLAT + Toolbox + 6 properties	Training	97.1	96.2	96.8
		Test	94.7	100	96.2
2	KeratinoSens + Toolbox + Avg.Lys.Cys + 6 properties	Training	98.5	100	98.9
		Test	84.2	100	88.5
3	KeratinoSens + h-CLAT + Avg.Lys.Cys + 6 properties	Training	97.1	92.3	95.7
		Test	89.5	85.7	88.5
4	h-CLAT + Toolbox + Avg.Lys.Cys + 6 properties	Training	95.6	96.2	95.7
		Test	84.2	100	88.5
5	KeratinoSens + h-CLAT + Toolbox + 6 properties	Training	95.6	96.2	95.7
		Test	89.5	85.7	88.5
6	h-CLAT + KeratinoSens + 6 properties	Training	94.1	88.5	92.6
		Test	89.5	85.7	88.5
7	h-CLAT + Avg.Lys.Cys + KeratinoSens + Toolbox + LogP	Training	91.2	96.2	92.6
		Test	89.5	85.7	88.5
8	h-CLAT + Avg.Lys.Cys + 6 properties	Training	95.6	92.3	94.7
		Test	84.2	85.7	84.6
9	Avg.Lys.Cys + Toolbox + 6 properties	Training	91.2	100	93.6
		Test	78.9	100	84.6
10	h-CLAT + 6 properties	Training	86.8	88.5	87.2
		Test	89.5	85.7	88.5
11	h-CLAT + Toolbox + LogP	Training	80.9	92.3	84.0
		Test	84.2	100	88.5
12	Avg.Lys.Cys + KeratinoSens + 6 properties	Training	92.6	96.2	93.6
		Test	73.7	85.7	76.9
13	Avg.Lys.Cys + KeratinoSens + Toolbox + LogP	Training	88.2	92.3	89.4
		Test	78.9	85.7	80.8
14	Avg.Lys.Cys + 6 properties	Training	85.3	100	89.4
		Test	73.7	100	80.8
15	Toolbox + 6 properties	Training	89.7	80.8	87.2
		Test	84.2	71.4	80.8
16	KeratinoSens + Toolbox + 6 properties	Training	91.2	84.6	89.4
		Test	73.7	85.7	76.9
17	KeratinoSens + 6 properties	Training	79.4	88.5	81.9
		Test	73.7	85.7	76.9
18	6 properties only	Training	67.6	88.5	73.4
		Test	73.3	71.4	73.1

Abbreviations: 6 properties = molecular weight, log octanol:water partition coefficient, log water solubility, log vapor pressure, melting point, and boiling point; Avg.Lys.Cys = average depletion for lysine and cysteine peptides from the DPRA; Cys = average depletion of cysteine peptide; DPRA = direct peptide reactivity assay categorical response; h-CLAT = human cell line activation test; LLNA = murine local lymph node assay; Lys = average depletion of lysine peptide from the DPRA; NEG = negative; POS = positive; SVM = support vector machines; Toolbox = QSAR Toolbox.

^a **Bolded red text** shows the best model (Variable Set 1) based on the average accuracy of the test and training sets. The models with the next highest accuracies (Variable Sets 2–5) are in red text but not bolded.

^b The training set of 94 substances contains 68 LLNA sensitizers and 26 LLNA nonsensitizers. The test set of 26 substances contains 19 LLNA sensitizers and 7 LLNA nonsensitizers (Figure 2).

Misclassified Substances

- The training set substances misclassified by the seven SVM models with the highest accuracies are shown in Table 6. The results from the individual *in chemico/in vitro/in silico* test methods are also shown for reference.
 - None of the false negatives were prehaptens (n = 2), which must oxidize to produce skin sensitization.
 - The three models with the highest accuracies correctly classified the 12 prohaptens, which must be metabolized to produce skin sensitization.
 - h-CLAT correctly classified more misclassified substances (6) than any of the other test methods (3–4 substances).
- The test set substances misclassified by the seven SVM models with the highest accuracies are shown in Table 7.
 - None of the false negatives were prehaptens (n = 1).
 - The three models with the highest accuracies correctly classified the four prohaptens.
 - h-CLAT and OECD Toolbox correctly classified more misclassified substances (5) than any of the other test methods (2–4 substances).

Table 6. Misclassified Substances for the Seven SVM Models with the Highest Accuracy – Training Set^a

Test Method or Model ^b	3-Phenacyl-pyridine	2-Acetylpyridine	Pyridine ^c	Nonanoic acid	3,4-Dihydrocoumarin ^d	Benzylidene acetone	Xylene	2-Hydroxy-ethyl acrylate	Eugenol ^e
LLNA	NEG	NEG	POS	POS	POS	POS	POS	POS	
DPRA (79%)	1	1	0	0	1	1	0	1	1
KeratinoSens (73%)	0	1	0	0	0	1	0	1	1
h-CLAT (78%)	1	1	1	1	1	1	0	1	0
Toolbox (77%)	1	0	0	0	1	0	0	0	1
h-CLAT + Toolbox + 6 properties (97%)	POS	NEG	POS	NEG	POS	NEG	POS	POS	POS
DPRA + KeratinoSens + h-CLAT + Toolbox + Lys + Cys + Ave.Lys.Cys + 6 properties (95%)	POS	NEG	POS	NEG	POS	POS	POS	POS	POS
KeratinoSens + h-CLAT + Toolbox + Ave.Lys.Cys + 6 properties (95%)	POS	NEG	POS	NEG	POS	POS	POS	POS	POS
KeratinoSens + Toolbox + Ave.Lys.Cys + 6 properties (94%)	NEG	NEG	POS	NA	NEG	POS	POS	POS	POS
KeratinoSens + h-CLAT + Ave.Lys.Cys + 6 properties (92%)	POS	POS	POS	NEG	POS	POS	POS	POS	NEG
h-CLAT + Toolbox + Ave.Lys.Cys + 6 properties (92%)	POS	NEG	NEG	NEG	POS	POS	NEG	POS	POS
KeratinoSens + h-CLAT + Toolbox + 6 properties (92%)	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS

Abbreviations: 6 properties = molecular weight, log octanol:water partition coefficient, log water solubility, log vapor pressure, melting point, boiling point; Avg.Lys.Cys = average depletion for lysine and cysteine peptides from the DPRA; Cys = average depletion of cysteine peptide; DPRA = direct peptide reactivity assay categorical response; h-CLAT = human cell line activation test; LLNA = murine local lymph node assay; Lys = average depletion of lysine peptide from the DPRA; NEG = negative; POS = positive; SVM = support vector machines; Toolbox = QSAR Toolbox.

^a Misclassifications are shaded in gray.

^b Parentheses show the accuracy for the test methods for all 120 substances and the average accuracy of the test and training sets for the SVM models.

^c Pyridine, 3,4-dihydrocoumarin, and eugenol are prohaptens.

Table 7. Misclassified