

Leora Vegosen, PhD, MHS¹ and Todd M. Martin, PhD²



¹ U.S. Environmental Protection Agency (EPA); This research was conducted while Dr. Vegosen was an ORISE Postdoctoral Fellow

² Center for Computational Toxicology and Exposure, Office of Research and Development, EPA

OBJECTIVES

This study aims to

- Develop Quantitative Structure Activity Relationship (QSAR) models for predicting skin sensitization as a binary toxicity endpoint
- Assess the performance of different modeling methods and descriptor sets
- Develop consensus models
- Assess methods for defining applicability domain (AD)

APPROACH

- A local lymph node assay (LLNA) dataset of 1355 chemicals was compiled from the NICEATM LLNA Database, OECD QSAR Toolbox, and eChemPortal.
- Records including chemical structures were curated in EPA’s DSSTox Database.
- Using 10 different modeling methods, models were developed in the Online Chemical Database with Modeling Environment (OCHEM) with 25 descriptor sets available in OCHEM and two additional descriptor sets: PaDEL descriptors and descriptors developed for EPA’s Toxicity Estimation Software Tool (T.E.S.T.).
- The best-performing models were used to develop consensus models.
- Java code was used to automate the generation of ADs covering 95% of training set compounds using different measures of distance to model.


RESULTS: Balanced Accuracies (BA %) for the Validation Set (Using Bagging Validation of the Training Set) for the Models with the Best Performing Descriptor Sets (Mean BA > 72%)

Descriptors	ANN	ASNN	KNN	LibSVM	WEKA-J48	WEKA-RF	Mean	STDEV
MORDRED 3D	74	76	71	76	75	75	74.5	1.9
PaDEL	73	76	71	77	75	75	74.5	2.2
PyDescriptor 3D	73	77	70	76	73	77	74.3	2.8
ALogPS, Estate	73	75	70	76	74	75	73.8	2.1
alvaDesc 3D	72	73	70	76	79	73	73.8	3.2
Dragon6 3D	74	74	70	74	76	72	73.3	2.1
SIRMS	72	72	75	74	71	76	73.3	2.0
MOLD2	73	75	69	74	76	72	73.2	2.5
T.E.S.T.	74	76	69	74	74	72	73.2	2.4
CDK2	70	74	73	74	72	71	72.3	1.6
Mean for these descriptors	72.2	73.7	70.8	75.1	74.5	73.8	73.7	2.3

IMPACT

- The best-performing modeling methods -- including Associative Neural Networks (ASNN), Support Vector Machines (SVM), WEKA-J48, and WEKA-RF -- produced validation set balanced accuracies > 75%.
- T.E.S.T. and PaDEL descriptors performed comparably to the best-performing descriptor sets in OCHEM.
- Consensus models performed better than individual models.
- The assessed AD generally did not improve the results, considering the tradeoff between balanced accuracy and prediction coverage, so other methods for defining AD should be explored.
- Models will be made publicly available in OCHEM and T.E.S.T., which will contribute to improving the performance and accessibility of new approach methodologies (NAMs) for predicting skin sensitization.

For more information: Leora Vegosen: vegosen.leora@epa.gov Todd Martin: martin.todd@epa.gov

QSAR Model Development for Skin Sensitization					Vegosen L, Martin TM		
Additional Results: Balanced Accuracies (BA %) and Applicability Domain (AD) Coverage (%) for Bagging Validation Models Used to Develop Consensus Models (ensemble standard deviation was used to assess the Distance to Model (DM))							
Method	Descriptors (or consensus type)	Training Set BA	Validation Set (VS) BA (100% coverage)	AD based on the DM that covers 95% of the training set			
				VS BA	Coverage	Product (BA X Coverage)	
ASNN	Mordred 3D	69	76	75	95	72	
ASNN	PaDEL	70	76	75	96	72	
ASNN	PyDescriptor 3D	68	77	76	94	72	
ASNN	T.E.S.T.	69	76	76	94	71	
LibSVM	ALogPS, Estate	69	76	75	97	73	
LibSVM	alvaDesc 3D	68	76	76	97	73	
LibSVM	Mordred 3D	68	76	75	97	72	
LibSVM	PaDEL	70	77	76	97	73	
LibSVM	PyDescriptor 3D	69	76	76	97	74	
WEKA-RF	PyDescriptor 3D	70	77	79	97	76	
WEKA-RF	SRMS	71	76	76	97	73	
WEKA-J48	Dragon 6 3D	70	76	76	97	73	
WEKA-J48	Mordred 3D	69	76	76	97	74	
WEKA-J48	alvaDesc 3D	70	79	75	97	73	
Consensus	(simple average)	72	79	79	95	75	
Consensus	(optimal combination of models for each property)	72	80	79	91	72	
Consensus	(weighted by model accuracy)	72	78	79	95	75	

**Additional Results: Balanced Accuracies (BA %) for the External Validation Set
(Using Five-Fold Cross-Validation of the Training Set) for the Models with the Best Performing Descriptors (Mean BA $\geq 70\%$)**

Descriptors	ANN	ASNN	DNN	KNN	LibSVM	LSSVMG	RFR	WEKA-J48	WEKA-RF	XGBOOST	Mean	STDEV
ALogPS, Estate	72	73	73	69	76	75	71	70	72	73	72.4	2.1
T.E.S.T.	71	77	72	69	73	77	73	65	72	75	72.4	3.6
alvaDesc (3D blocks 1-30)	75	75	72	70	74	75	75	66	72	70	72.4	3.0
PaDEL	76	76	67	72	73	77	75	60	75	72	72.3	5.2
SIRMS (labels:charge+logp+hb+refractivity)	70	70	68	70	72	77	75	69	77	75	72.3	3.4
RDKIT (3D blocks: 1-11, 15-16)	73	73	73	66	73	77	74	64	76	71	72.0	4.1
MORDRED (All) 3D	72	74	67	71	74	73	75	68	71	72	71.7	2.6
Dragon6 (3D blocks 1-29)	73	70	72	68	73	72	73	66	73	72	71.2	2.4
CDK2 (cons,topol,geom,elec,hybrid) 3D	71	69	72	71	71	75	71	64	74	72	71.0	3.0
ChemaxonDescriptors (pH 0 - 14:1) 3D	72	74	70	67	67	75	72	67	73	70	70.7	3.0
RDKIT (AVALON)	69	73	68	68	72	73	72	66	74	72	70.7	2.7
PyDescriptor 3D	68	71	72	69	69	75	74	63	75	68	70.4	3.8
MOLD2	74	76	68	70	71	72	71	58	72	70	70.2	4.8
Fragmentor (length:2 - 4)	69	70	71	65	70	73	74	67	70	71	70.0	2.6
Mean for these descriptors	71.8	72.9	70.4	68.9	72.0	74.7	73.2	65.2	73.3	71.6	71.4	2.7

Models with balanced accuracies in bold were used to develop consensus models.

ANN = Artificial Neural Network; ASNN = Associative Neural Networks; DNN = Deep Neural Network; KNN = k Nearest Neighbors; LibSVM = Grid-Search Parameter Optimization Support Vector Machine; LSSVMG = Least Squares Support Vector Machine; RFR = Random Forest; WEKA-J48 = Weka C4.5 decision trees for classification; WEKA-RF = Weka Random Forest; XGBOOST = Scalable and Flexible Gradient Boosting

Additional Results: Usefulness of Defining the Applicability Domain (AD) Based on Different Measures of Distance to Model (DM) Covering 95% of Compounds from the Training Set for ASNN Five-fold Cross-Validation Models

Descriptors	AD Covering 100% of the Training Set	PROB-STD (95%)			CLASS-LAG (95%)			ASNN-STDEV (95%)			ASNN-CORREL (95%)		
		VS BA (%)	VS Cov (%)	Product	VS BA (%)	VS Cov (%)	Product	VS BA (%)	VS Cov (%)	Product	VS BA (%)	VS Cov (%)	Product
	VS BA (%)												
T.E.S.T.	77	76	95	72	76	95	72	77	91	70	76	92	70
Mold2	76	76	93	71	76	93	71	77	91	70	77	91	70
PaDEL	76	76	92	69	76	91	69	75	84	63	76	91	69

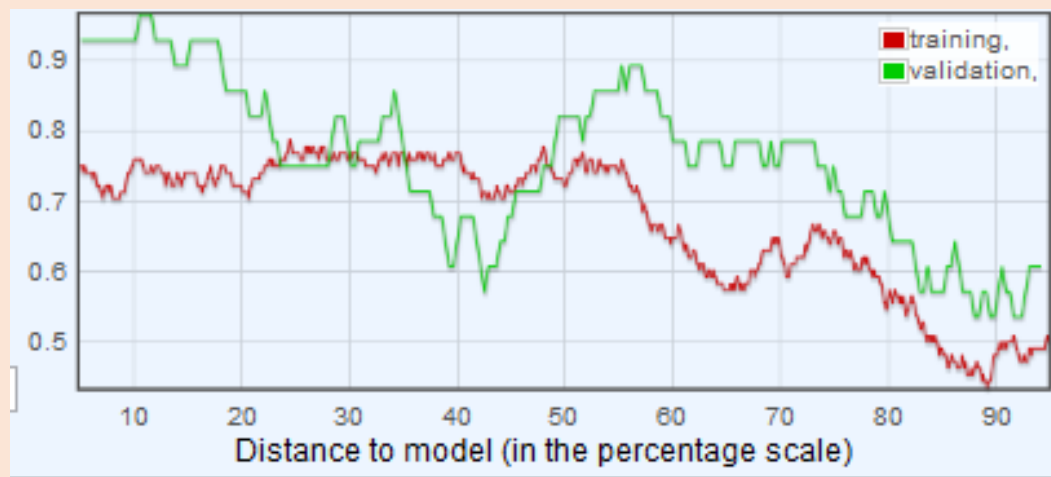
VS = Validation Set; BA = Balanced Accuracy; Cov = Coverage

Distance to Model (DM) Measures used to Calculate the Applicability Domain

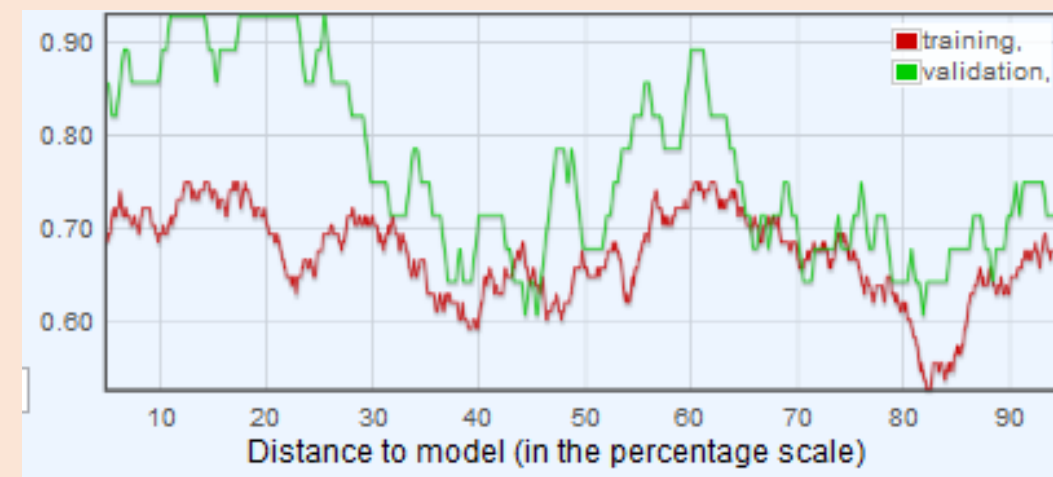
- Prob-STD:** combines information about uncertainty from CLASS-LAG and ASNN-STDEV
- CLASS-LAG:** The absolute value of the difference between a continuous prediction value and the nearest of the binary labels [-1, 1]
- ASNN-STDEV:** The sample standard deviation of an ensemble of models is used as an estimator of the model uncertainty for a given compound
- CORREL:** based on the correlation of vectors of an ensemble’s predictions for the target compound and compounds from the training set

Additional Results: Accuracy vs. Four Different Measures of Distance to Model for the ASNN model developed using T.E.S.T. Descriptors and five-fold cross-validation

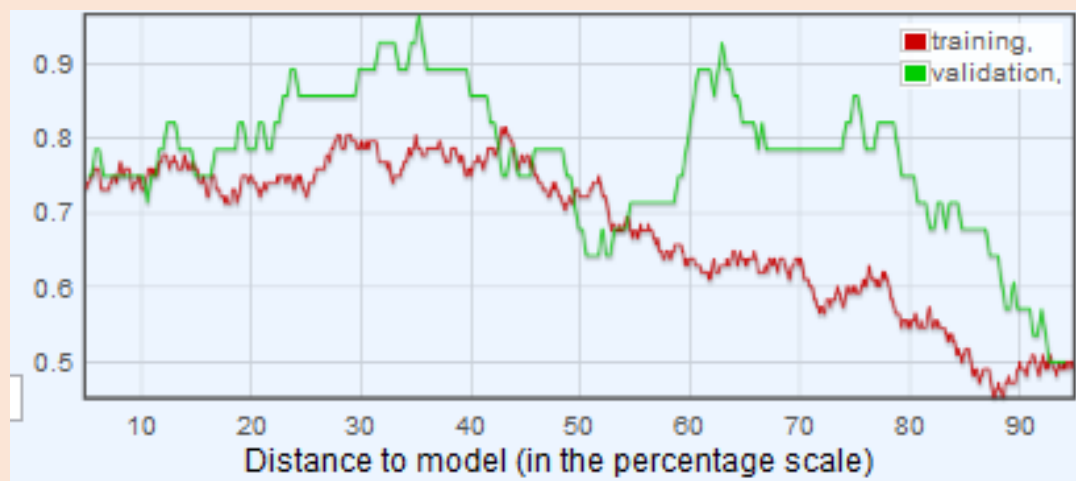
Prob-STD



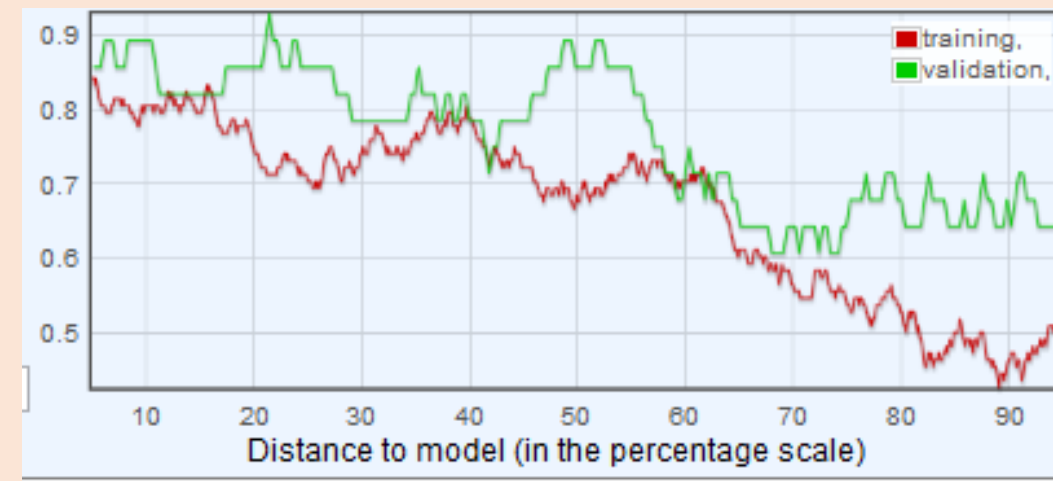
ASNN-STDEV



CLASS-LAG



ASNN-CORREL



REFERENCES

- Online Chemical Database with Modeling Environment (OCHEM): <https://ochem.eu/home/show.do>
- Sushko et al. 2010, Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set, *Journal of Chemical Information and Modeling* 50 (12): 2094–2111. <https://doi.org/10.1021/ci100253r>.
- Sushko et al. 2011, Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information, *Journal of Computer-Aided Molecular Design* 25 (6): 533–54. <https://doi.org/10.1007/s10822-011-9440-2>.

AUTHOR AFFILIATIONS

Leora Vegosen is currently an Epidemiologist in Prioritization and Informatics Branch 1 (PIB1), Data Gathering and Analysis Division (DGAD), Office of Pollution Prevention and Toxics (OPPT), Office of Chemical Safety and Pollution Prevention (OCSP), U.S. Environmental Protection Agency (EPA). This research was conducted while Dr. Vegosen was an ORISE Postdoctoral Fellow in the Computational Chemistry and Cheminformatics Branch (CCCB), Chemical Characterization and Exposure Division (CCED), Center for Computational Toxicology and Exposure (CCTE), Office of Research and Development (ORD), EPA. Todd Martin is a Research Chemist in CCCB, CCED, CCTE, ORD, EPA.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the creators of OCHEM, especially Igor Tetko, who took the time to meet virtually in December 2020 to describe OCHEM and to provide detailed answers to questions posed during the meeting and by email.

The authors gratefully acknowledge colleagues at EPA, particularly Grace Patlewicz for her thoughtful and detailed input on the selection of data sources to include in this study; Chris Grulke, Ann Richard, Indira Thillainadarajah, and Tony Williams for providing advice and assistance with data curation in DSSTOX; Charlie Lowe for assistance with downloading PaDEL descriptors; and Christian Ramsland for double checking the data presented in the Tables.

This research benefited from insights gained through discussions with the Chemistry Modeling Team of the Computational Chemistry and Cheminformatics Branch, CCED, CCTE, ORD, EPA and the authors thank the team participants including Dan Chang, Paul Harten, Gabriel Sinclair, Matt Boyce, Scott Kolmar, Mahmoud Shobair, Tia Tate, Ryan Lougee, and others mentioned above.

Dr. Vegosen gratefully acknowledges support by an appointment to the Research Participation Program at the EPA, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and the EPA.

DISCLAIMER

The views expressed in this presentation are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.