Chemical Research in Toxicology

pubs.acs.org/crt

Article

The Good, The Bad, and The Perplexing: Structural Alerts and Read-Across for Predicting Skin Sensitization Using Human Data

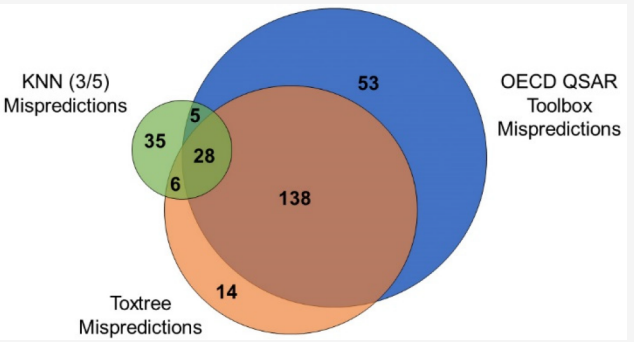
Emily Golden, Daniel C. Ukaegbu, Peter Ranslow, Robert H. Brown, Thomas Hartung, and Alexandra Maertens*

Cite This: *Chem. Res. Toxicol.* 2023, 36, 734−746

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In our earlier work (Golden et al., 2021), we showed 70−80% accuracies for several skin sensitization computational tools using human data. Here, we expanded the data set using the NICEATM human skin sensitization database to create a final data set of 1355 discrete chemicals (largely negative, ∼70%). Using this expanded data set, we analyzed model performance and evaluated mispredictions using Toxtree (v 3.1.0), OECD QSAR Toolbox (v 4.5), VEGA's (1.2.0 BETA) CAESAR (v 2.1.7), and a *k*-nearest-neighbor (kNN) classification approach. We show that the accuracy on this data set was lower than previous estimates, with balanced accuracies being 63% and 65% for Toxtree and OECD QSAR Toolbox, respectively, 46% for VEGA, and 59% for a kNN approach, with the lower accuracy likely due to the higher percentage of nonsensitizing chemicals. Two hundred eighty seven chemicals were mispredicted by both Toxtree and OECD QSAR Toolbox, which was approximately 20% of the entire data set, and 84% of these were false positives. The absence or presence of metabolic simulation in OECD QSAR Toolbox made no overall difference. While Toxtree is known for overpredicting, 60% of the chemicals in the data set had no alert for skin sensitization, and a substantial number of these chemicals were in fact sensitizers, pointing to sensitization mechanisms not recognized by Toxtree. Interestingly, we observed that chemicals with more than one Toxtree alert were more likely to be nonsensitizers. Finally, a kNN approach tended to mispredict different chemicals than either OECD QSAR Toolbox or Toxtree, suggesting that there was additional information to be garnered from a kNN approach. Overall, the results demonstrate that while there is merit in structural alerts as well as QSAR or read-across approaches (perhaps even more so in their combination), additional improvement will require a more nuanced understanding of mechanisms of skin sensitization.

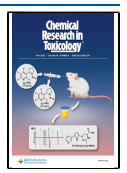


1. INTRODUCTION

Allergic contact dermatitis is the clinical manifestation of skin sensitization.[1] Traditionally, skin sensitization has been assayed by using the guinea pig maximization test (GPMT) and the Buehler test in guinea pigs. More recently, the Local Lymph Node Assay (LLNA) has also been used in mice. However, replacing these animal-based methods with alternative methodologies has been a major focus for the in chemico and in vitro community, as concerns have been raised regarding the reproducibility[2−4] as well as the predictivity of animal tests for humans.[5] In addition, there is significant consumer opposition to animal testing as well as considerable time and expense[6−8] associated with animal-based methods. Several in chemico and in vitro test guidelines have been developed to assay skin sensitization and compiled by the Organisation for Economic Co-operation and Development (OECD), an international organization that sets global standards for toxicity testing. Alternative methods for skin sensitization include, but are not limited to, the Direct Peptide Reactivity Assay (DPRA) (OECD TG 442C[9]), the KeratinoSens assay (OECD TG 442D[10]), and the h-CLAT assay (OECD TG 442E[11]). Most recently, the OECD has, for the first time, validated a test method that includes an in silico approach, the OECD Test Guideline 497. This is a defined approach for assessing skin sensitization that utilizes in silico methods to predict skin sensitization, suggesting that in silico tools are becoming more widely embraced by the regulatory community.[12]

Our previous work has demonstrated that in silico skin sensitization platforms achieved respectable accuracies of around 70−80% (up to 87%) and balanced accuracies of 60−80% (up to 88%) for human skin sensitization data sets[13] and even higher for animal skin sensitization data sets.[4] Others have demonstrated an accuracy of greater than or equal to

Received: November 30, 2022
Published: May 1, 2023



80%, with sensitivity and specificity varying significantly among the different models.[14−17] Although there have been a number of different in silico skin sensitization tools developed,[18] most of them are based on structural alerts, nearest neighbor, or read-across approaches or some combination of these methods.[19]

Structural alerts are the cornerstone of predictive toxicology. They predict toxicity through the identification of reactive functional groups or moieties. In the case of skin sensitization, structural alert models flag functional groups/moieties that are considered to be electrophilic and subsequently bind with proteins, i.e., hapten formation, the so-called molecular initiating event of skin sensitization.[1,20] Read-across or nearest-neighbor approaches extrapolate toxicity from similar chemicals (similar most often being defined based on structure) to a target chemical, a chemical with unknown toxicity. Thus, for skin sensitization, the nearest-neighbor or read-across approaches take advantage of existing experimental sensitization data and apply them to the target chemical to predict its sensitization status.[3,4] Some approaches, such as that used in the OECD QSAR Toolbox Automated Workflow, use both methods grouping chemicals by mechanistic similarity (i.e., structural alert for electrophilicity) followed by a nearest-neighbors approach, based on structural similarity with a threshold of 50%.[21]

While significantly useful for flagging potentially active chemicals, structural alerts by their nature are prone to overpredict and are not themselves considered sufficiently accurate.[20,22] Read-across, on the other hand, is challenged by instances where similar chemicals exhibit dissimilar toxicity, referred to as an "activity cliff". Activity cliffs have commonly been investigated in the pharmaceutical industry as a tool to identify potential toxic functional groups and design away hazards or to identify small changes in chemical structure that may improve the efficacy of a drug.[23−25] For predictive toxicology, identifying activity cliffs is essential. Activity cliffs can illuminate instances where read-across can be misleading as well as detect potential mechanisms of toxicity.

Here, we build upon our previous human skin sensitization data set by adding a subset of the human NICEATM skin sensitization database and investigate how activity cliffs can affect the performance of in silico sensitization tools. We confirm that structural alerts have significant informational value, but consistent with previous findings, structural alerts are not equally valid in all areas of chemical space and are certainly not sufficient to determine skin sensitization status by themselves.[22] At the same time, we show that using a nearest-neighbors approach has both potential and limitations. The results presented here will inform users when to use caution and when to have higher confidence in the results of structural alerts, read-across, and nearest-neighbor approaches.

2. METHODS

Data Set. The data set used in this analysis was generated by combining our previously curated data set (Golden et al., 2021)[13] with the skin sensitization data available in the Integrated Chemical Environment (ICE) compiled by the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) (https://ice.ntp.niehs.nih.gov/DATASETDESCRIPTION; accessed Nov 4, 2021).[26−28] Details regarding the curation of the Golden et al. data set can be found elsewhere.[13] In brief, the Golden et al. data set contains 466 discrete chemicals. For the purposes of this assessment, naturals, metals, and chemicals that did not have a Chemical Abstract Service Registry Number (CASRN), simplified molecular input line entry system (SMILES), and PubChem compound ID (CID) were removed, leaving 393 discrete chemicals with sufficient chemical identifiers.

The NICEATM skin sensitization database contains 24 445 data points for 2176 unique identifiers (i.e., chemicals identified by CASRN (n = 1871), proprietary chemicals identified by generic chemical name only and no CASRN or formulation name (n = 222) and formulations with no CASRN or generic chemical name (n = 83)), including in vitro, in vivo, and human data. For this analysis, we focused on the human data points, which include results from human maximization tests as well as human repeat insult patch tests (HRIPT); this left 11 366 data points for 1378 unique identifiers. These results are reported as either "active/inactive" or quantitative values (e.g., induction dose per skin area). For this analysis, we focused on discrete chemicals with CASRN (i.e., proprietary chemicals and mixtures were excluded) with reported results of active or inactive; this left 2017 data points for 1154 chemicals.

Some chemicals had only one test result, while others had multiple test results. Chemicals that had only one test result were assigned a sensitization status based on that single test result. For chemicals that had multiple results, the following approach was taken.

- Unanimous results: for chemicals that had multiple results that all agreed, the unanimous sensitization status (either active or inactive) was assigned.
- Mixed results: for chemicals that had multiple results that were mixed (i.e., both active and inactive results), a case-by-case assessment was carried out to look at the range of concentration in addition to the weight of evidence of the mixed results, and a sensitization status (either active or inactive) was assigned.
- Split results: for those chemicals that had multiple assay results at the same test concentration but those assay results were split 50/50, an active sensitization status was assigned to be protective of human health.

A more detailed approach is outlined in Supplemental Table 1.

In this set of 1154 chemicals, 142 chemicals were already present in the Golden et al. data set. Therefore, these chemicals were not carried over from the NICEATM database and the Golden et al. data set status was used. This left 1012 chemicals from NICEATM's database. Finally, we removed all chemicals with no PubChem CIDs or SMILES (simplified molecular identification line entry system) strings; SMILES strings for the chemicals were retrieved from ChemIDplus. Using these SMILES strings, a list of PubChem IDs was generated from the PubChem ID Identifier Exchange Service tool (https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi). There were 50 chemicals that lacked SMILES and CID strings, leaving a final subset of 962 chemicals from the NICEATM human skin sensitization database. The processing procedure for this data set is summarized in Supplemental Figure 1. The Golden et al. and NICEATM data sets were combined for a final total of 1355 discrete chemicals.

Similarity Score Calculations. Structural similarities between each chemical in the data set were calculated using PubChem's Score Matrix Service (undated). PubChem compound identifiers (CID) were entered into the Score Matrix Service, and the score type was set to "2-D similarity (substructure keys)". The PubChem Score Matrix Service calculates similarity scores based on its own list of defined substructure fingerprints and Tanimoto distance.[29,30] Once the score matrix was calculated, it was converted into a pairwise distance list using R Studio.

Construction of Chemical Similarity Maps. A chemical similarity map is a tool that aids in the visualization of the chemical similarity of the entire chemical space of a data set. To construct a chemical similarity map, the pairwise distances for all chemicals 80%, 85%, and 90% similar were uploaded to CytoScape 3.9.1[31] and arranged using the degree sorted circle layout.

Predictive Toxicology Models. To predict the skin sensitization potential of all chemicals in the data set, three skin sensitization prediction tools were used: (1) Toxtree v. 3.1.0,[32,33] (2) the

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

Table 1. Summary of Predictive Models[a]

platform/approach (version)	model	methodology	refs
Toxtree (v 3.1.0)	Skin Sensitization Reactivity Domains	structural alert	IdeaConsult, Ltd., 2018; Enoch et al., 2008
OECD QSAR Toolbox (v 4.5)	Automated Workflow for EC3 from LLNA or Skin sensitization from GPMT assays	read-across/structural alert (with metabolism simulation)	LMC, 2021; Yordanova et al., 2019
	Automated Workflow for EC3 from LLNA or Skin sensitization from GPMT assays−metabolism removed	read-across/structural alert	LMC, 2021
VEGA (1.2.0 BETA)	CAESAR (v 2.1.7)	QSAR	Benfenati et al., 2013; IRFMN, 2020

[a]Note: The Automated Workflow with metabolism removed was constructed using the Automated Workflow created by LMC and modified by the authors of this analysis to remove the metabolism commands.

Organisation for Economic Cooperation and Development's (OECD) Quantitative Structure Activity Relationship (QSAR) Toolbox v. 4.5,[21,34] and (3) VEGA's CAESAR Skin Sensitization Model.[35] Each approach is outlined below and summarized in Table 1.

Toxtree is a structural alert tool that scans the chemical structure for the presence of an electrophilic chemical group that is associated with skin sensitization. These groups are referred to as "alerts", of which there are 5 in Toxtree (i.e., SNAr, SN2, Michael addition, Schiff base, and acyl transfer agent). If no alerts are detected, the output is simply "no alert". Within Toxtree, the Skin Sensitization Reactivity Domains[33] decision tree was used.

OECD QSAR Toolbox applies a read-across approach by identifying the protein binding alert in the target chemical structure, gathering chemicals that contain the same protein binding alert and ultimately applying a read-across prediction based on structural similarity.[21] In OECD QSAR Toolbox, we utilized three workflows, two preprogrammed workflows and one workflow that we built. The two preprogrammed workflows that were used in this analysis were the automated workflow for skin sensitization using LLNA and GPMT and the defined approach for skin sensitization (DASS) automated workflow. The third workflow was one that we designed based on the preprogrammed automated workflow for skin sensitization using the LLNA and the GPMT, but we removed the metabolism commands, resulting in a workflow that did not account for metabolism. We created this workflow to determine what impact the metabolism simulators built into the automated workflow for skin sensitization using the LLNA and GPMT have on model accuracy.

VEGA[35,36] is a platform that houses QSAR models for several human health end points, including skin sensitization. Here, we utilized the CAESAR model (v 2.1.7) to predict the skin sensitization potential of our data set. The CAESAR model provides a reliability score with each prediction; here, we used only results with "good reliability" or results that were stated to be based on experimental values.

Finally, we used a k-nearest-neighbor (kNN) classification approach. In the context of skin sensitization, this approach assigns a sensitization status for a target chemical (i.e., a chemical with unknown sensitization status) by using the majority status of k neighboring chemicals.[37] For this analysis, neighbors were assigned based on structural similarity, though other approaches can be used when applying a kNN approach. Here, we assessed k values from 5 to 15 to explore optimal values and accuracy for k. The final kNN analysis focuses on chemicals with 5 nearest neighbors based on a structural similarity metric of 85%.

3. RESULTS

Data Set. After combining the chemicals from Golden et al. (2021)[13] and the NICEATM data set,[26−28] the final data set curated in this assessment consisted of 1355 discrete chemicals (see Supplemental Table 2); the majority were characterized as inactive in human skin sensitization tests (72%), as displayed in Supplemental Figure 2.

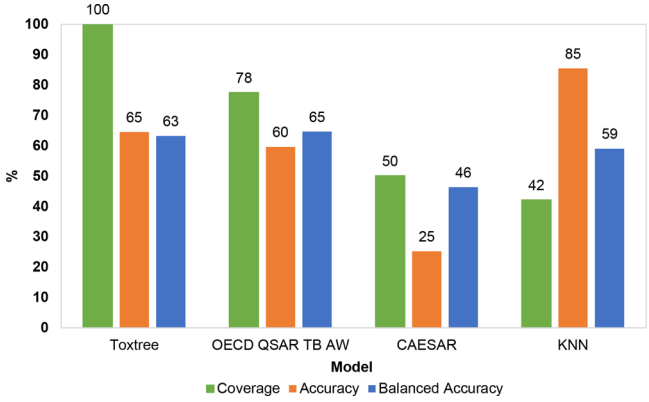


Figure 1. Accuracy and balanced accuracy for the predictive models and a kNN approach. Coverage, accuracy, and balanced accuracy for the three predictive tools—Toxtree, OECD QSAR Toolbox Automated Workflow (OECD QSAR TB AW), and VEGA's CAESAR—as well as a kNN approach. Coverage is the number of predictions made by each model/approach out of the total number of chemicals assessed by the model/approach. Toxtree's coverage made predictions for each chemical, resulting in complete coverage, while OECD QSAR Toolbox's coverage was about 80%. CAESAR and the kNN approach used in this analysis both had reduced coverage of about 50% and 60%, respectively. Toxtree and OECD QSAR Toolbox performed similarly, with accuracies of 65% and 60%, respectively, and balanced accuracies of 63% and 65%, respectively. CAESAR's accuracy and balanced accuracy were low (25% and 46%, respectively); this was mainly attributed to a high rate of false-positive predictions (92%). The kNN approach used in this analysis had the highest accuracy (85%) but a reduced balanced accuracy of 59%, which was due to the high rate of false negative predictions (89%).

The chemical sensitization statuses for the chemicals sourced from the NICEATM database were based on a single assay result, multiple assays with unanimous results, split results, or mixed results, as defined previously and summarized in Supplemental Table 1. The distribution of these "weight of evidence" calls is presented in Supplemental Figure 3. The majority (75%, n = 872) of the skin sensitization calls for the NICEATM data set were based on only one assay result, suggesting that the confidence in these calls may be low. About 10% (n = 108) of the chemicals had mixed or split assay results, again indicating somewhat low confidence results. The remaining 15% (n = 174) of chemicals had multiple assays with unanimous results, indicating that these sensitization calls had higher confidence. Regardless of the number of results available per chemical, the fact that they are based on human results rather than animal data makes them more

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

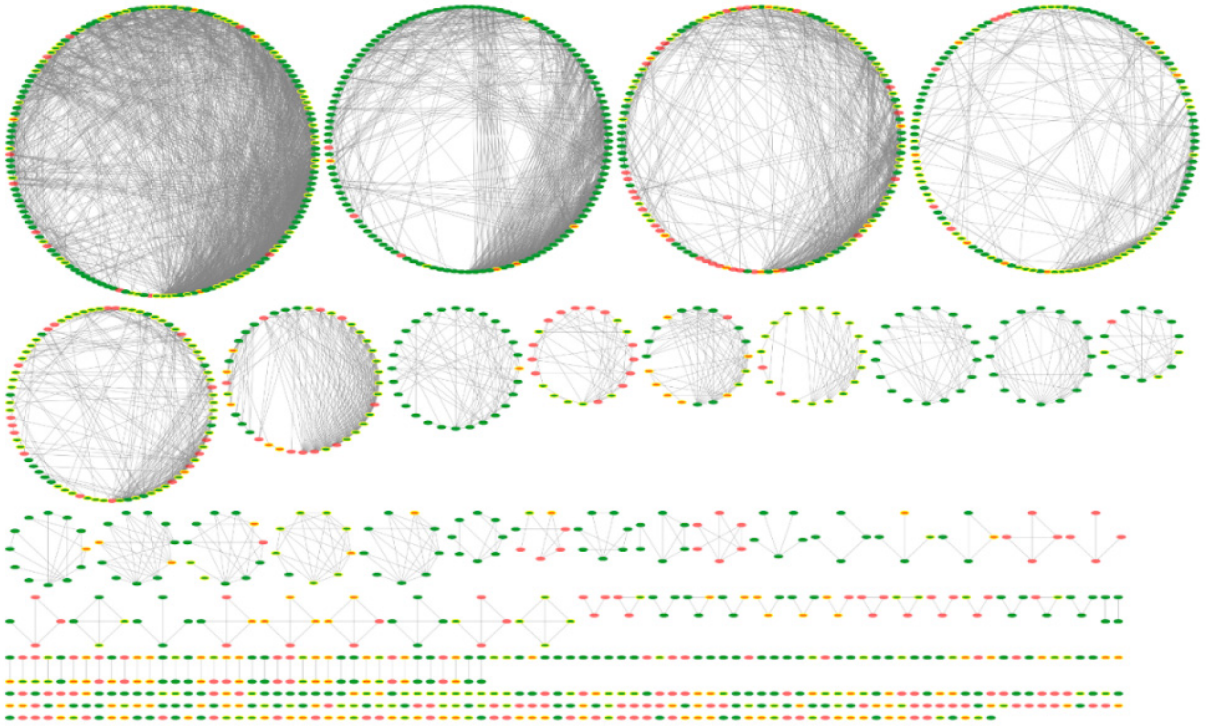


Figure 2. Chemical similarity map for combined data set with Toxtree mispredictions. Chemical similarity map for all chemicals in the combined data set (n = 1355) using a Tanimoto similarity score of 0.85. Node color indicates the experimental sensitization status based on human data: green nodes are nonsensitizers, red nodes are sensitizers. The lines (i.e., edges) between two chemicals (i.e., nodes) indicate that the connected chemicals are structurally similar by 85% or more. Yellow borders around the nodes indicate a misprediction by Toxtree; the overall accuracy for Toxtree for this data set was 65%, while the balanced accuracy was 63%.

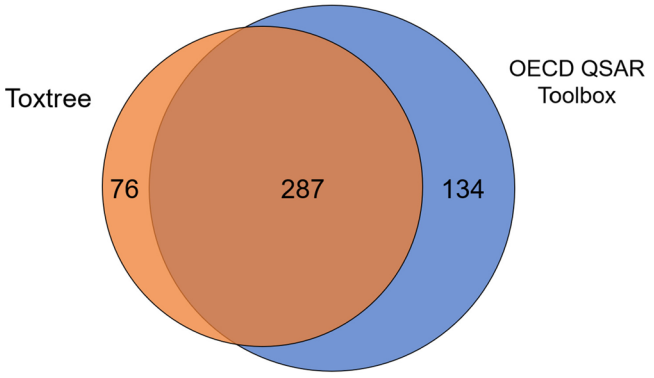


Figure 3. Overlap of mispredictions by QSAR Toolbox and Toxtree. When comparing the mispredictions between OECD QSAR Toolbox and Toxtree, there was a significant overlap of mispredicted chemicals between the two models (n = 287). OECD QSAR Toolbox had about twice the number of unique mispredictions as Toxtree. Note: Some mispredictions for Toxtree were not included in this analysis because predictions for those chemicals were not made by OECD QSAR Toolbox (OECD QSAR Toolbox does not make a prediction unless sufficient analogs are identified, whereas Toxtree makes predictions for all chemicals).

relevant to characterize human skin sensitization outcomes than the LLNA or GPMT results.

Model Performance. Once the final data set was curated, we analyzed the entire combined data set (n = 1355 chemicals) using the three predictive tools (Toxtree, OECD QSAR Toolbox, and CAESAR) as well as a kNN approach. The accuracies and balanced accuracies for the 3 models are depicted in Figure 1. One advantage of the Toxtree approach is

Table 2. Accuracy of QSAR Toolbox with and without Metabolic Simulation[a]

	automated workflow with metabolism		automated workflow without metabolism	
metric	n	%	n	%
accuracy	621/1042	60	662/1079	61
balanced accuracy	n/a	65	n/a	66
sensitivity	208	75	225	77
specificity	413	54	437	55
false positive	353	46	351	45
false negative	68	25	66	23
coverage	1042/1344	78	1079/1344	80

[a]OECD QSAR Toolbox includes simulated chemical metabolism in its skin sensitization workflow; however, overall, accounting for metabolism did not improve accuracy for this data set.

that it can make a prediction for any chemical, while the QSAR Toolbox, CAESAR model, and a kNN approach can only make predictions in the presence of sufficient analogs. Furthermore, for the kNN approach applied in this analysis, the final k value selected was 5, which reduced the number of chemicals predicted to only 571, while OECD QSAR Toolbox predicted 1042, and CAESAR predicted 589 chemicals.

Overall, Toxtree and OECD QSAR Toolbox performed similarly, with 65% and 60% accuracies, respectively, and 63% and 65% balanced accuracies, respectively. This was lower than other estimates and likely reflects some unique features of this data set. Unlike other data sets, the NICETAM data set consisted of mostly inactive chemicals, and unlike data from ECHA, it does not consist of regulatory data for industrial chemicals and thus is skewed to negative results. In other

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

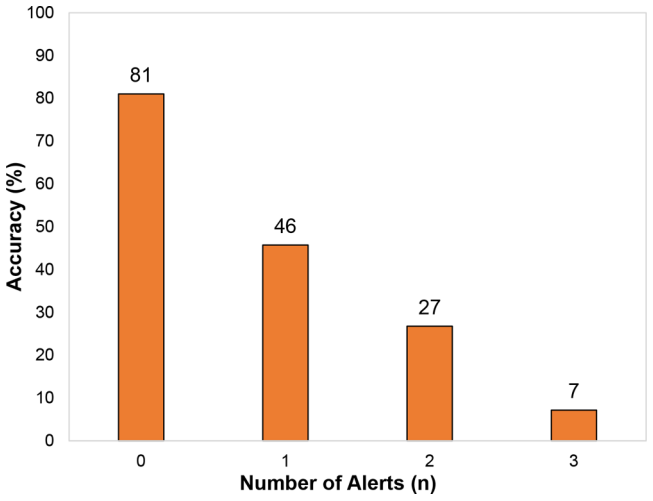


Figure 4. Toxtree accuracy by number of alerts present in a chemical. Toxtree predicted a chemical to be a sensitizer if it identified 1 of 5 alerts within the chemical; if no alerts were identified, it predicted the chemical to be nonsensitizing. As shown in Figure 4, Toxtree's accuracy decreased as the number of alerts in a chemical increased.

studies,[22] Toxtree often had a high percentage of false positives, and our results were consistent. In addition, for our data set, Toxtree also had a high number of false negatives (40%), i.e., chemicals with no known mechanism for skin sensitization yet were positive experimentally. On the other hand, OECD QSAR Toolbox had a high number of false positives (46%). The lowest performing model, CAESAR, similarly suffered from overprediction, as demonstrated by a very high number of false positives (92%), which contributed to its balanced accuracy of 46%. Complete performance metrics for all models as well as performance for individual data sets (i.e., NICEATM and Golden et al. (2021)) can be found in Supplemental Table 3. Taken together, this shows worse performance of all models on a larger data set that overlaps less with the data sets used for training purposes.

Model Mispredictions. Chemical similarity maps are a useful tool to visualize a large chemical space to determine if there are any patterns, such as mispredictions. In Figure 2, we use a chemical similarity map based on structural similarity (Tanimoto distance 0.85) using 2D Pubchem chemical fingerprints to show graphically the mispredictions from Toxtree (n = 481/1355; 35%).

Read-across is predicated on the idea that structural similarity indicates toxicological similarity. When mapping chemicals by structural similarity, we would expect to see each cluster of structurally similar chemicals exhibiting the same sensitization status (i.e., clusters composed of only green or only red nodes); however, in many areas of the chemical similarity map (see Figure 2), clusters have mixed chemical sensitization status, that is, there are clusters that contain both red and green nodes—these are activity cliffs in the data set. Activity cliffs can be defined in many ways, such as based on structural similarity, i.e., chemicals that are highly structurally similar but have very different experimental toxicities (in this case, skin sensitization). In the context of a chemical similarity map, activity cliffs are represented by two chemicals connected on the map, indicating high structural similarity, where one chemical in the connected pair is a sensitizer (red node) and the other is not (green node). This offers some clues as to why both kNN and OECD QSAR Toolbox, whose predictive methodologies rely on structurally similar chemicals, offer no overall advantage over Toxtree: many similar chemicals in this data set have discordant status. Toxtree, on the other hand, which looks exclusively for reactive functional groups, often overpredicted and underpredicted for different clusters of chemicals.

There was a fair amount of overlap in the chemicals mispredicted by both Toxtree and OECD QSAR Toolbox (Figure 3)—approximately 20% of the entire data set (i.e., the same 287 chemicals) was mispredicted by both models, and the overwhelming majority of these were false positives (n = 240/287, 84%).

To determine whether there were any patterns in the mispredictions made by both Toxtree and OECD QSAR Toolbox, we analyzed the molecular weight and log K_{OW}, two common physicochemical properties that are known to affect skin permeability and therefore sensitization potential. However, there was no obvious pattern to the mispredictions based on either parameter (Supplemental Table 4). For example, as Toxtree is only looking for the presence or absence of a reactive functional group and does not consider physicochemical properties, one source of error could be a positive prediction for a chemical that is likely negative due to the chemical's size or inability to penetrate the skin. This confirms our previous finding where in the REACH registration data set[3] we found that a large molecular weight does not preclude skin sensitization owing to low bioavailability: 49 sensitizing chemicals with a molecular weight > 500 Da were found.

Our previous study[13] found a higher percentage of mispredictions in chemicals with ester groups. However, this pattern was not present in these data, as chemicals with or without an ester group were mispredicted at the same rate (see Supplemental Table 5a). Of the chemicals mispredicted by both models, almost one-half had a benzene ring, and globally there was a modest increase (5%) in mispredictions for both models in chemicals with a benzene ring (see Supplemental Table 5b).

Although there was no obvious explanation for the chemicals mispredicted by both models, the differences in model mispredictions pointed to some interesting results. OECD QSAR Toolbox had more unique mispredictions than Toxtree (n = 134 vs n = 76, respectively). Of the 134 chemicals mispredicted (which was 10% of the overall data set) by OECD QSAR Toolbox but correctly predicted by Toxtree,

Table 3. Accuracy of Toxtree and OECD QSAR Toolbox by Structural Alert[a]

SNAr accuracy (%)	Schiff base formation accuracy (%)	Michael acceptor accuracy (%)	acyl transfer agent accuracy (%)	SN2 accuracy (%)	no alerts accuracy (%)
47	31	49	36	30	81

[a]Overall, the performance by structural alert was fairly low, especially for the SN2 alert (30%) and the Schiff base formation alert (31%); however, Toxtree exhibited a relatively high accuracy for chemicals that had no alert (81%).

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

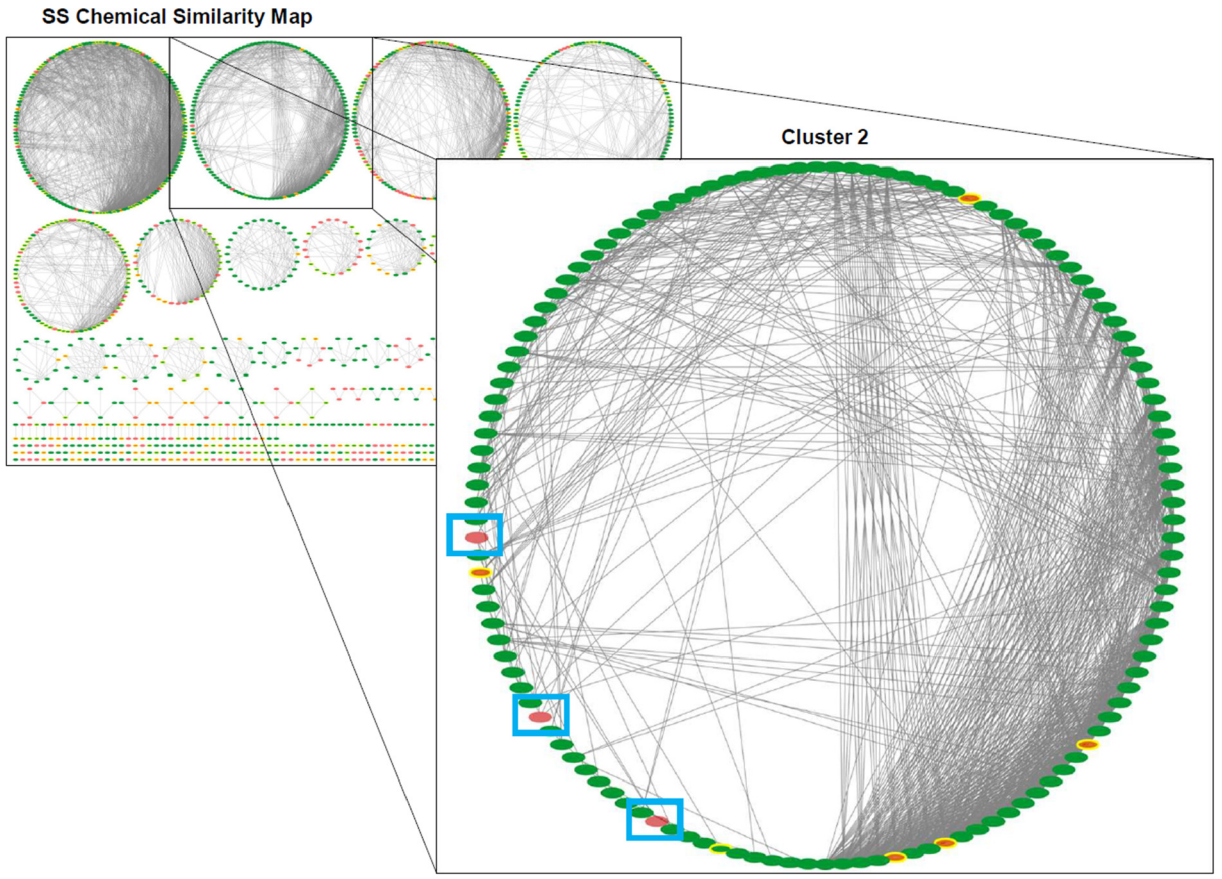


Figure 5. Sensitizers correctly identified by Toxtree with experimentally discordant neighbors: Cluster 2 from the overall chemical similarity map (Figure 2). Green nodes represent nonsensitizers, red nodes represent sensitizers based on experimental status, and yellow outlined nodes indicate mispredictions by Toxtree. The three red nodes identified in the blue squares are examples of instances where Toxtree correctly identified 3 sensitizers within a group of largely negative chemicals with relatively few false positives.

about one-half of the mispredictions (n = 61/134, 46%) were based on the metabolite and the majority of these were false positive (n = 52/61, 85%). As many of the predicted reactive metabolites were attributed to radical formation (n = 30), one possible explanation for this finding is that these chemicals were quenched by antioxidant mechanisms, as supported by their nonsensitizing experimental results. The remaining unique mispredictions by OECD QSAR Toolbox were based on the parent chemical (n = 73/134, 54%). Again, the majority of mispredictions were false positives (n = 61/73, 84%), with about one-half of those being attributed to chemicals containing a ketone functional group (n = 28/61, 46%) (see Supplemental Figure 4).

Metabolism can be one reason for activity cliffs. For example, chemical A may be metabolized from nonsensitizing to sensitizing or vice versa, but its structurally similar neighbor, chemical B, may not be metabolized. If this were the case, OECD QSAR Toolbox should be able to capture these differences, as part of its workflow includes metabolic simulation. However, as seen in Table 2, there was no overall improvement offered by accounting for metabolism in this data set, as the accuracies and balanced accuracies with and without metabolism were essentially the same, 60% vs 61% accuracy with and without metabolism, respectively, and 65% vs 66% balanced accuracy with and without metabolism, respectively. While this does not preclude the possibility that metabolism was important for some chemicals, within this data set it did not suggest that globally metabolism was a significant contributor to mispredictions or its mitigation.

Toxtree had about one-half the number of unique mispredictions as OECD QSAR Toolbox (n = 76). Of these, over one-half (n = 45/76, 60%) were false negatives. Almost one-half of these false negatives contained a sulfur atom (n = 21/45, 47%). These chemicals were flagged as having an SN2 alert in OECD QSAR Toolbox but not in Toxtree, suggesting an interesting blind spot in Toxtree where OECD QSAR Toolbox has an advantage.

Despite its correct identification of this subgroup, OECD QSAR Toolbox offered essentially no improvement over Toxtree in terms of balanced accuracy (i.e., only a 2% improvement, see Figure 1) and, in fact, had reduced coverage, that is, OECD QSAR Toolbox did not make predictions for all chemicals in the data set due to model design (i.e., some chemicals had a limited number of or no structurally similar chemicals available in the model), only predicting 78% of the chemicals in the data set compared to complete coverage (i.e., predictions for all chemicals in the data set) offered by Toxtree. In other words, the added nuance of predicting metabolism and using a read-across-based approach did little to improve accuracy and decreased the number of predictions made by OECD QSAR Toolbox.

Many of the chemicals in the data set (n = 796/1355, 59%) did not have an alert flagged by Toxtree, while about one-third of the chemicals in the data set contained 1 alert. The remaining 10% of chemicals contained either 2 or even 3 alerts

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

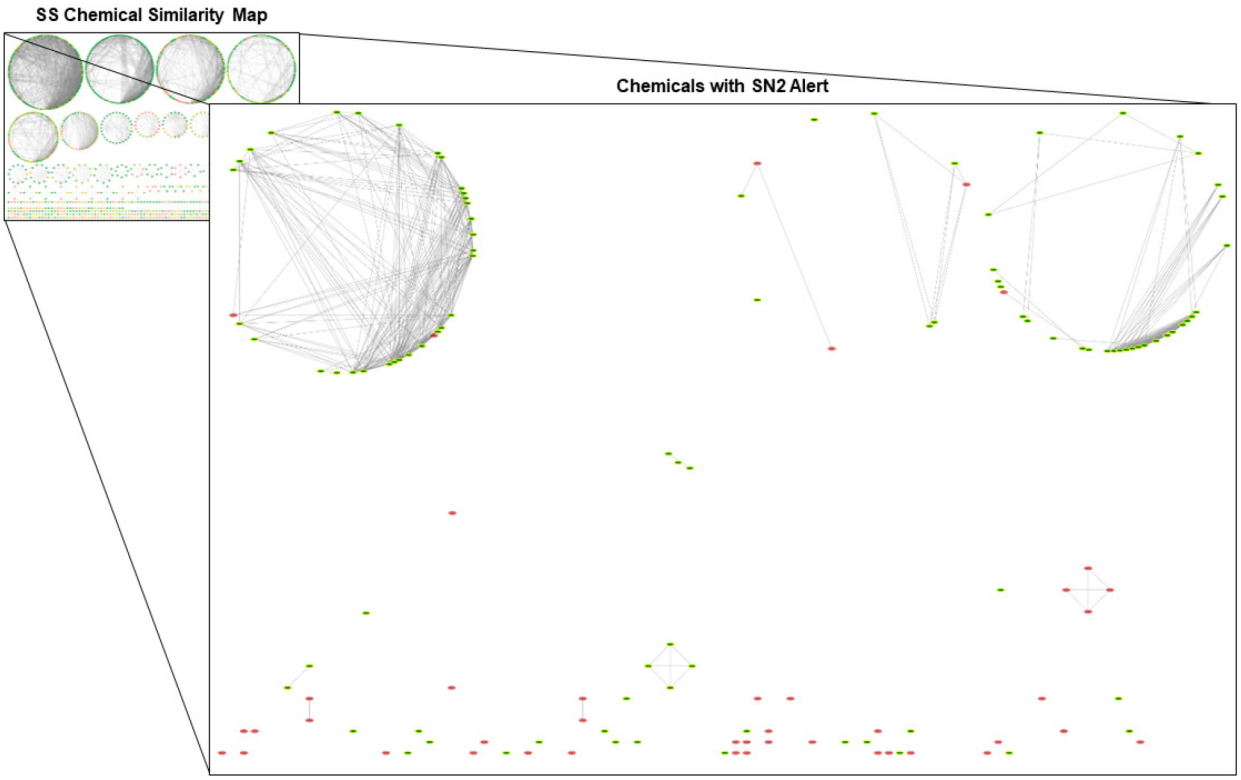


Figure 6. SN2 alerts according to Toxtree. Chemicals flagged as having an SN2 alert for skin sensitization are shown. This alert produced a high number of false positives, as indicated by green nodes (i.e., nonsensitizers based on human experimental data) with yellow outlines (i.e., mispredictions by Toxtree). While many chemicals were correctly flagged as sensitizers, a nearest-neighbor approach would likely not predict these chemicals correctly, as they did not cluster together.

each (see Supplemental Figure 5). Interestingly, many of the false positives in the data set had more than one Toxtree alert (n = 95, 7%). In fact, as the number of alerts within a chemical increased, accuracy decreased (see Figure 4). This may be because the chemical is larger, and the structural alerts may be within the middle of a larger, more complex molecule. Thus, the alert moieties may not be bioaccessible, rendering them nonsensitizers despite containing multiple alerts for sensitization.

Because Toxtree is an alert-based model, another useful approach to identify areas in the chemical space where it was mispredicting was to evaluate performance by alert. These results are summarized in Table 3. The accuracy of each alert was generally poor for this data set, with some alerts especially likely to be misleading, such as the SN2 (only 30% accurate) or Schiff base formation (only 31% accurate) alerts, which were likely, at least in part, attributed to the high number of negative chemicals in the data set. Indeed, Toxtree achieved a balanced accuracy of 63% despite poor alert performance, suggesting that its performance was substantially driven by the accuracy of "no alerts".

Since Toxtree looks at mechanism vs structural similarity (i.e., Tanimoto distance) used in read-across/nearest-neighbor approaches, it was useful to see when and where Toxtree helped identify mechanistically similar chemicals where a similarity-based approach did not (indeed, overall, 144 sensitizers in this data set had at least one highly structurally similar chemical (≥85% structurally similar) with discordant sensitization status). In other words, we wanted to investigate whether some structural activity cliffs were explained by the presence of a chemical moiety that Toxtree identified. In Cluster 2 (see Figure 5), Toxtree successfully identified 3 sensitizers within a group of largely negative chemicals with relatively few false positives.

However, the broader problems with a structural alert system were apparent when looking at all chemicals that had an SN2 alert (n = 141, see Figure 6), which had the lowest accuracy by alert (30%, see Table 3). Most of these chemicals are experimentally nonsensitizing, but the sensitizing chemicals often did not cluster together in a way that would allow a nearest-neighbors approach to identify them correctly, and many of the chemicals, once extracted from the larger, global map, had no sufficiently similar analog. While OECD QSAR Toolbox did not outperform Toxtree with this alert, it did suggest that the alerts can be evaluated based on their neighbors, although that approach was not precise due to the presence of activity cliffs.

Furthermore, alerts also need to be taken in context. For example, the five chemicals shown in Figure 7 were all flagged by the SN2 alert; however, they were experimentally discordant. The only chemical among these confirmed to be sensitizing was styrene-7,8-oxide (PubChem CID 114705), which was not surprising given that the SN2 alert was an epoxide positioned on the end of the molecule, making it easily accessible to react with a protein. On the other hand, the four experimentally nonsensitizing chemicals depicted in Figure 7 had SN2 alerts, but they are more centrally located in the molecule, suggesting that while the alert was present, it may not be accessible.

When looking at the map of chemicals that had no alerts for sensitization in Toxtree (Figure 8), there were many false negatives (shown as red nodes with yellow outlines).

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

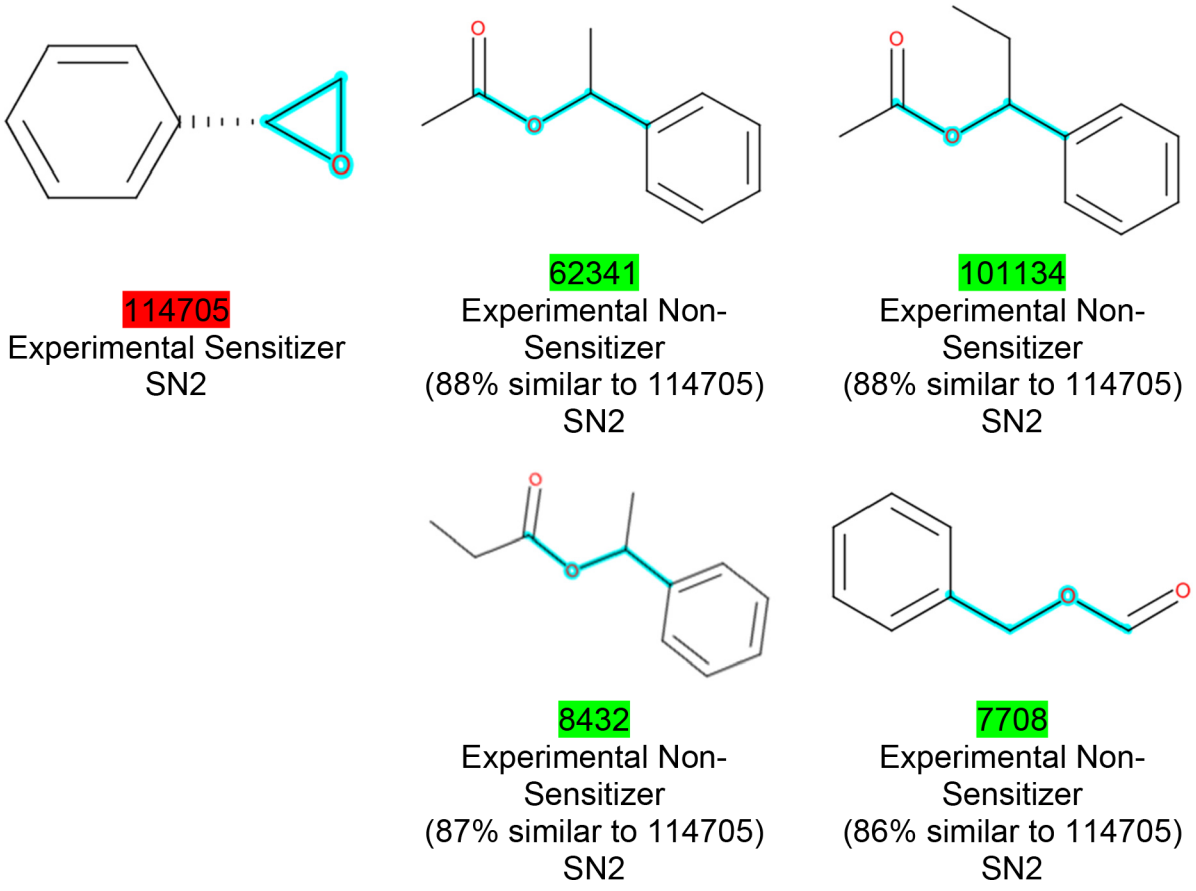


Figure 7. Mechanistically similar chemicals with discordant experimental status. The five chemicals presented were all flagged as sensitizers by the SN2 alert. However, the majority were nonsensitizing based on experimental data (as indicated by green highlighting). The portions of the molecule highlighted in aqua are the mechanistic alert. Notably, the alerts for the nonsensitizers were located in a more internal position of the molecule, while the epoxide group in the sensitizing group was located on the end and thus more accessible. These results suggest that alerts must be taken in context.

However, the false negatives did not cluster in such a way that would make a read-across-based approach work. Nearly 40% of the chemical pairs that lacked an alert had discordant experimental status (Table 4).

As mentioned above, many of these chemicals contained a sulfur group, and updating Toxtree to identify these chemicals would likely improve the overall accuracy. Overall, there were 796 chemicals (about 60% of the entire data set) with no alerts, and 151 of these chemicals were mispredictions (i.e., false negatives).

We investigated all clusters to try to identify the source of these false predictions. Interestingly, we discovered in Cluster 9 that although the chemicals there were structurally very similar (85−99% similarity), there were clear differences in sensitization status. Figure 9 shows the chemicals in Cluster 9 with no alerts. First, all chemicals with no alerts were extracted from the overall chemical similarity map (labeled "SS Chemical Similarity Map"); these chemicals are depicted in the map entitled "Chemicals with No SS Alert". Because these chemicals lacked structural alerts for skin sensitization, theoretically, they should all be experimental nonsensitizers, i.e., green nodes. However, we saw that this map was composed of both green and red nodes, indicating there were experimental sensitizers that lack structural alerts according to Toxtree. This was problematic as these were false negatives and thus posed a risk to human health in the context of skin sensitization. When separating sensitizers from nonsensitizers in Cluster 9, we observed a very nuanced difference: nearly all ortho-substituted chemicals were nonsensitizing, while all para-substituted chemicals were sensitizing. This suggests a mechanism outside Toxtree's prediction repertoire.

Overall, within each grouping of mechanistically similar chemicals, many positive chemicals had at least one structurally similar chemical with a discordant status (Table 4), making any read-across-based approach both prone to mispredictions and highly dependent on available data. When looking at the positive chemicals with no structural alert (n = 151), most had at least one chemical that was highly similar (n = 86) yet discordant sensitization status (n = 40).

This was born out in the overall disappointing balanced accuracy of the kNN approach used in this analysis. Predictions were made using various values for k (i.e., various number of neighbors), beginning with 5 and up to 15. As seen in Table 5, the number of chemicals with 5 neighbors was 571 while the number of chemicals with at least 15 neighbors was about one-half that (n = 240). While the kNN approach was fairly accurate for the subset of chemicals with many neighbors, this was in part because those chemicals were largely negative. Therefore, the balanced accuracy was much lower than the accuracy. Further, as k increased, the accuracy and balanced accuracy essentially plateaued. There was almost no change in

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

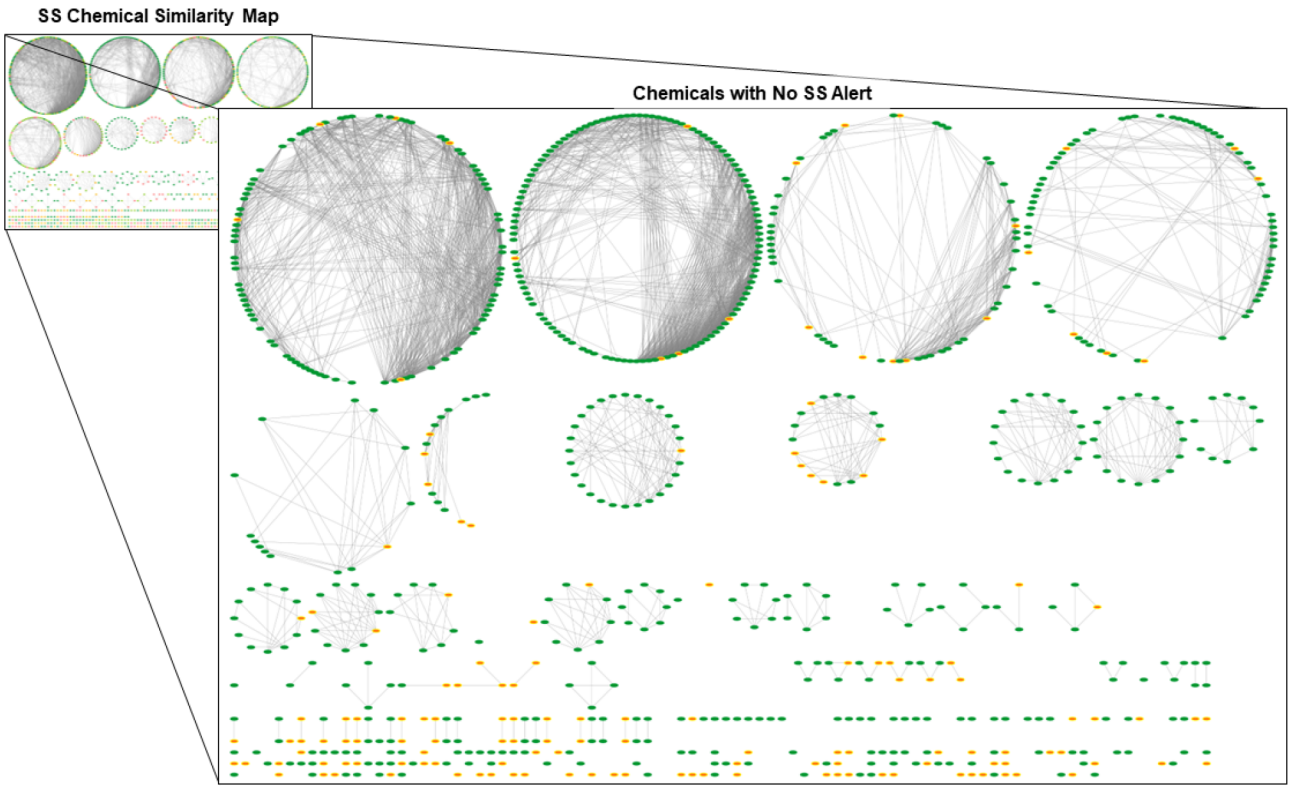


Figure 8. Chemicals without a structural alert for skin sensitization according to Toxtree: the chemicals in the data set that were not associated with a skin sensitization alert in Toxtree. Green nodes are nonsensitizers based on experimental human skin sensitization data, red nodes are sensitizers based on experimental human skin sensitization data, and nodes with yellow circles indicate false predictions by Toxtree. If the Toxtree structural alert model was performing perfectly, we would expect to see only green nodes in this graph, as chemicals with no skin sensitization alerts should all be nonsensitizers. However, we observed many (151/796, 19%) nodes that were red, indicating chemicals that have been confirmed as sensitizers based on human skin sensitization data but lack a structural alert for skin sensitization data. As a result, these chemicals are all misclassified as false negatives in Toxtree.

Table 4. Experimentally Discordant Pairs of Highly Structurally Similar Chemicals (>85%) That Share the Same Mechanism[a]

	experimentally discordant chemical pairs	
shared alert[b]	*n*	%
Schiff base formation	85	9
Michael acceptor	99	10
acyl transfer agent	69	7
SN2	22	2
no alert	357	37

[a]This table demonstrates that there are many chemical pairs that are both structurally and mechanistically similar yet experimentally discordant, suggesting that these chemical pairs could be challenging to predict accurately for both alert- and structural-similarity-based approaches. [b]SNAr alert metrics not depicted, as only 1 experimentally discordant chemical pair was present in the data set.

accuracy, regardless of the number of neighbors, and no change in balanced accuracy when greater than 10 neighbors were considered. Therefore, more data would not necessarily improve the kNN approach in data-rich parts of the chemical universe in the absence of an understanding of mechanism.

Nonetheless, we did note that a kNN approach tended to mispredict different chemicals more than either OECD QSAR Toolbox or Toxtree, suggesting that there was additional information to be gleaned from this method. In this data set, the kNN 5 approach made predictions for 571 chemicals, and Toxtree also made predictions for all 571 of these chemicals. However, OECD QSAR Toolbox made predictions for only 501 of these chemicals. When comparing the mispredictions of these 501 chemicals across the kNN, OECD QSAR Toolbox, and Toxtree approaches, only 28 chemicals (6%) were mispredicted by all 3 approaches (see Figure 10).

Of the chemicals mispredicted by all three approaches, no obvious pattern based on functional groups, molecular weight, or log K_{OW} was readily apparent. The mispredicted chemicals were a mix of chemical functional groups (e.g., aldehydes, esters, phenols), and the molecular weight (between 100 and 300 g/mol) and log K_{OW} values (from −5 to 3) were not particularly skewed in one direction or the other. However, many of these chemicals were considered as weak to mild sensitizers (i.e., they were either considered as Basketter et al. (2014)[38] Category 4 or 5 chemicals or the human sensitization tests on which they were based elicited a positive response but at a relatively high concentration). So, it is possible that the mispredictions here were based on potency. Nevertheless, we cannot know for certain without a thorough potency analysis. It is possible a consensus approach utilizing all three methods may improve overall accuracy, but it should be noted that this analysis was based on a small subset of the data set (n = 501/1355, 37%), as it was limited by both the number of chemicals with a high number of neighbors as well as the coverage of OECD QSAR Toolbox.

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

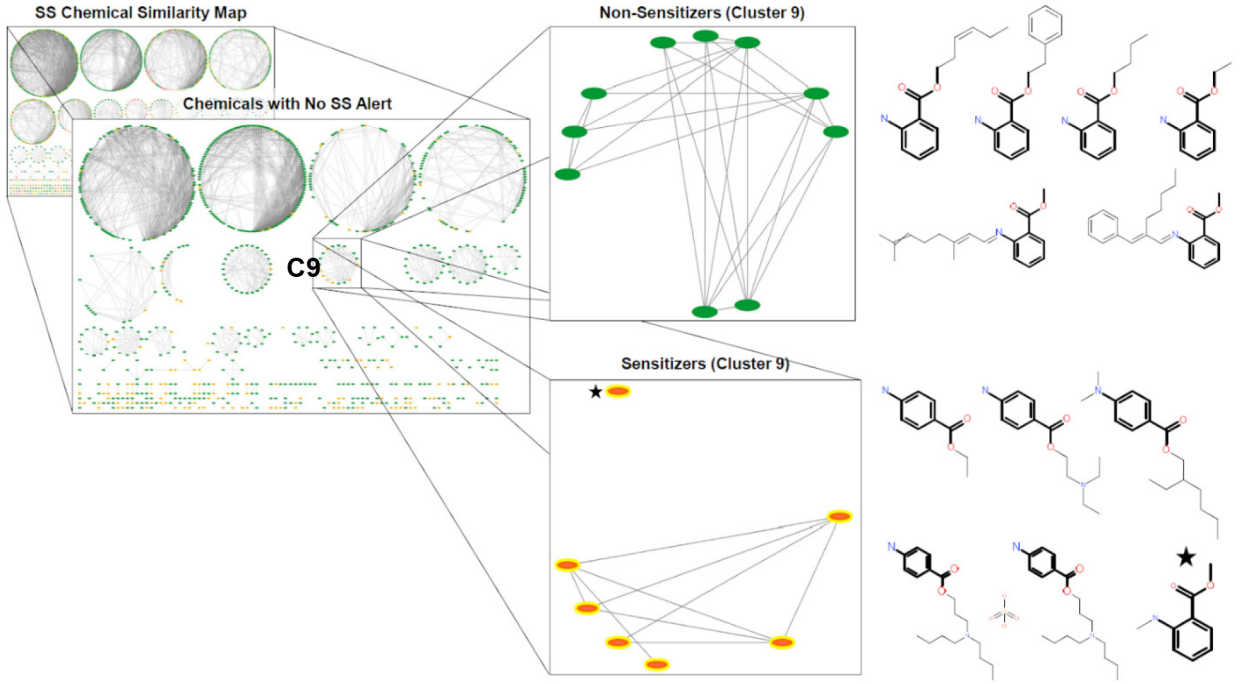


Figure 9. Chemicals in Cluster 9 with no alerts. Depicted in the background of the upper left-hand corner is the chemical similarity map for the entire skin sensitization data set (labeled "SS Chemical Similarity Map"), from which the chemicals with no alerts were pulled (identified as "Chemicals with no SS Alerts"). These maps consist of chemical clusters which were groups of structurally similar chemicals. The cluster of focus for this figure is Cluster 9 (cluster identified by C9), and the challenge of read-across was amply illustrated by this cluster. These 17 chemicals were grouped together because of their high structural similarity, as can be seen from their structures (Tanimoto score ranged from 85% to 99%), but they were discordant in experimental status.

Table 5. Performance Metrics for the kNN Approach To Predict Skin Sensitization Outcomes[a]

	5, n = 571		10, n = 341		11, n = 321		12, n = 290		15, n = 240	
performance metric	3/5	4/5	6/10	7/10	6/11	8/11	7/12	8/12	8/15	11/15
accuracy	85	89	87	88	87	88	88	89	89	89
balanced accuracy	59	55	52	50	50	50	50	50	50	50

[a]kNN was fairly accurate overall (85−89%); however, balanced accuracy was significantly lower (50−59%, depending on k); this is almost certainly based on the fact that the data set is predominantly negative. In addition, as the number of neighbors increases, accuracy and balanced accuracy remain essentially the same, suggesting that adding more neighbors does not increase accuracy. Note: The value of n designates the number of chemicals that have the listed number of neighbors. For example, 571 chemicals have 5 neighbors. However, a prediction may not have been made for all chemicals, as "split calls" (i.e., chemicals that have an even number of active and in active predictions) resulted in no prediction.

4. DISCUSSION

Here, we presented a large human skin sensitization data set curated from the NICEATM ICE database to build upon our previous data set, consisting of the Basketter et al. (2014)[38] data set and our own curated data set from the Hazardous Substances Data Bank (HSDB, a rich toxicity database that contains human data, including from occupational settings). The newly expanded data set proved to be somewhat problematic for the performance of both structural alert and read-across/nearest-neighbor approaches, as supported by relatively modest accuracies of 65% and 60%, respectively, as well as balanced accuracies of 63% and 65%, respectively. Our previous work[13] demonstrated somewhat higher accuracies and balanced accuracies for model performance. However, when we compared the performance of both Toxtree and OECD QSAR Toolbox using only the chemicals that overlapped between this assessment and our last assessment, the accuracies and balanced accuracies for that subset of data were similar between this assessment and our previous one. Although accuracy and balanced accuracy metrics decreased in the current analysis, this may be attributed to the expansion of the data set, as a subanalysis of the individual data sets from our previous work revealed that the performance metrics in this analysis for the chemicals from our previous assessment are nearly identical (see Supplemental Table 6).

Indeed, higher balanced accuracies from previous performance assessments may be attributed to the fact that most available skin sensitization data sets are largely positive while the current data set is largely negative. For example, the Basketter data set, considered the gold standard data set for human skin sensitization, was curated by first devising 6 skin sensitization potency categories and assigning chemicals with human skin sensitization data to these categories; the final data set consisted of 131 chemicals, and of these, the majority (n = 107 (82%)) are chemicals characterized as chemicals positive for skin sensitization, ranging from strong to weak.[38] Similarly, our own previous effort at retrieving human skin sensitization data consisted of a majority of positive chemicals (234 out of 375 chemicals total, or 62%), reflecting perhaps the likelihood of reporting positive vs negative results (i.e., publication

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

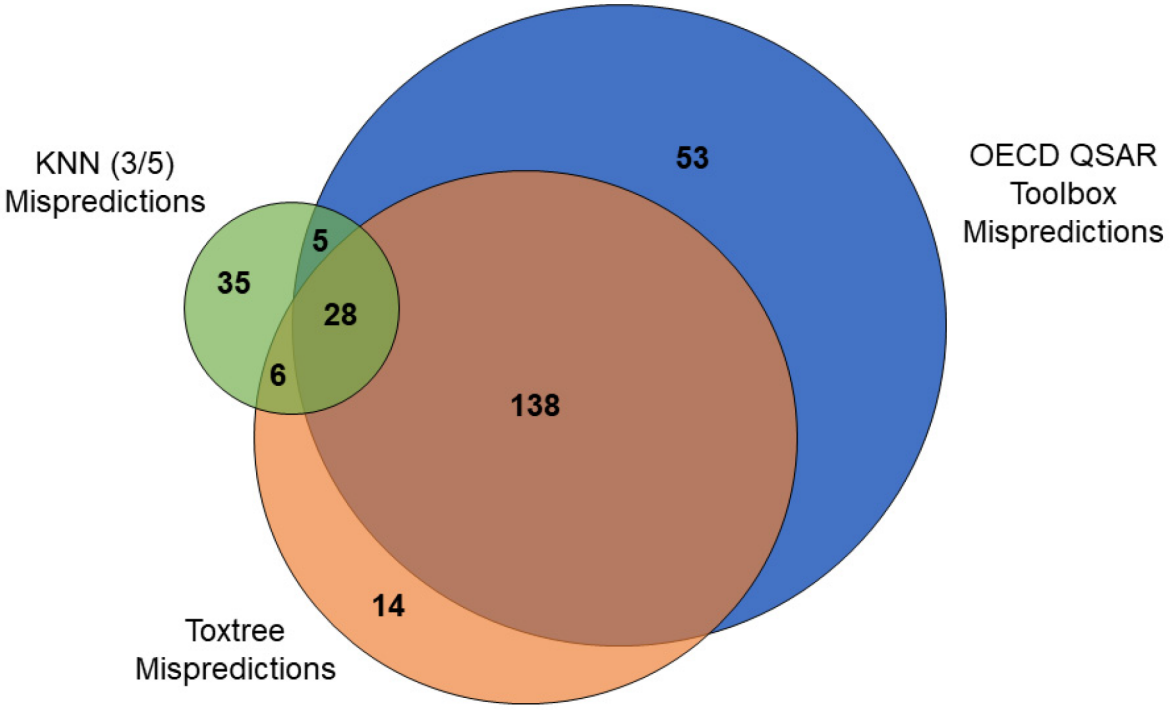


Figure 10. Mispredictions among kNN 3/5 approach, OECD QSAR Toolbox, and Toxtree. Overall, there were relatively few chemicals mispredicted by all 3 approaches (n = 28/501, 6%), and no obvious patterns were detected in these 28 chemicals aside from the fact that most were likely considered weak to moderate sensitizers. It is important to note that this was a relatively small set of chemicals (n = 501).

bias).[13] While the ECHA data set, drawn largely from regulatory testing submitted to ECHA, consisted of 21% positive chemicals, it may, in fact, have its own source of bias, as LLNA tests are more likely to be performed when either a structural alert or a read-across approach indicates a potential for skin sensitization. Therefore, assessing skin sensitization model performance based on a data set that has a high percentage of chemicals that are considered nonsensitizers as well as a high number of structurally diverse organic compounds offers a significant advantage for testing model performance in a real-world application, and the human skin sensitization data set compiled by NICEATM offers such a data set. More fundamentally, computational models are only as good as the data used to develop them. Therefore, when using highly unbalanced data, it is important to determine if and why the data are unbalanced in the first place. The skew toward chemicals with structural alerts and positives in other data sets likely contributed toward an ability to accurately assess the performance of those models and missed some key molecular mechanisms.

Moreover, we showed that Toxtree overpredicted many chemicals—some structural alerts provided remarkably little information value—while also missing a considerable number of potentially positive chemicals with no obvious reactive functional group. We also showed that neither a mechanism-blind read-across (kNN) nor a mechanism-informed read-across (such as OECD QSAR Toolbox) could easily solve this problem. This was readily apparent with chemicals containing an SN2 group: only 30% of these chemicals were positive, and they did not cluster in an obvious way. Consequently, correctly identifying the positives will be both challenging and highly dependent on available data. However, consulting with our colleagues in the field of chemistry can be a useful first step to address this issue.

In addition, the presence of more than one structural alert was more commonly associated with nonsensitizers—a seemingly paradoxical finding that reflects the fact that structural alerts must be taken into context. This perhaps may be attributed to the fact that molecules with multiple alerts could potentially be of larger molecular size. Consequently, the alerts may be related to structures embedded within the molecule. Alternatively, these alerts could be located in the molecule in such a way that balances the electron distribution, making them less likely to be reactive. Additionally, we showed that there was a substantial number of chemicals with no structural alerts that were, in fact, sensitizers, which, in this data set, contributed to the lower balanced accuracy.

Although the role of metabolism in skin sensitization is unclear, the lack of improvement in OECD QSAR Toolbox did not indicate that this was the most likely explanation for most of the observed chemical mispredictions and in fact increased the rate of false positives. In many instances, there was no obvious explanation for potential sensitization, likely pointing to mechanisms that were not covered by Toxtree, which focused exclusively on electrophilic mechanisms. Our analysis suggested that thiol groups in particular were likely to be sensitizers in the absence of other structural alerts in Toxtree.

While read-across remains one of the most common alternatives to animal testing,[3] it is dependent both on available data as well as the definition of "similar". Although previous work has demonstrated a relatively high balanced accuracy for analog-based approaches (especially when a high cutoff is used for similarity), we found that globally such approaches likely had limited accuracy and activity cliffs remained a significant challenge, as indicated by the subset of highly similar chemicals with the same predicted structural alert yet discordant status. In some cases, the discordant status

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746

could be explained by steric effects, showing the importance of considering 3D similarity. Another possibility was metabolism to a less reactive form.

A final challenge to such predictive methodologies was looking at both potency and risk—weak sensitizers are, generally speaking, challenging for in silico methods. They are especially challenging for instances in which sensitization likely reflects both weak potential for sensitization and genetic susceptibility. For example, ethanol was classified as a sensitizer in our data set. However, only 6 of 93 individuals reported sensitization at a 50% concentration, indicating that ethanol requires both a high-exposure scenario as well as some level of individual predisposition to sensitization.[39] This combination of high exposure and few affected individuals illustrated that it was important to consider not only chemical hazard but also overall risk.

Overall, while the models evaluated here demonstrated only modest performance, this may be resolved, in part, by expanding model training sets to include more negative chemicals. Furthermore, we have identified several specific areas where the models can be improved. For example, although globally metabolism did not provide any improvement in accuracy, resolving areas where metabolites are driving false predictions may improve the utility of metabolism. Additionally, by expanding the SN2 alert definition for Toxtree, many false negative predictions could be resolved. Also, reconsidering the way similarity was defined may offer improvement to read-across-based approaches. Although no single chemical feature appeared to be the driver for all mispredictions, through investigative analyses such as this one, the field of computational toxicology will continue to improve. Going forward, models with more data and deep learning approaches, such as the read-across structure−activity relationship (RASAR) approach, will likely be better able to build upon the network effects of a kNN model to improve accuracy, as has been demonstrated in other larger data sets.[4]

■ ASSOCIATED CONTENT

● Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrestox.2c00383.

Definition of weight of evidence categories; performance metrics for Toxtree, OECD QSAR Toolbox, and CAESAR by data set; patterns of physicochemical properties among mispredictions; presence of ester group among mispredictions; presence of benzene group among mispredictions; comparison of model accuracy and balanced accuracy from Golden et al. to equivalent subset in current data set; NICEATM Data set processing procedure; final human skin sensitization data set; distribution of chemicals by weight of evidence designation; breakdown of unique mispredictions by OECD QSAR Toolbox; distribution of chemicals in the data set by number of Toxtree alerts (PDF)

Final human skin sensitization data set with chemical identifiers and sensitization statuses (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Alexandra Maertens − *Center for Alternatives to Animal Testing (CAAT), Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, United States; Consortium for Environmental Risk Management (CERM), Hallowell, Maine 04347, United States;* orcid.org/0000-0002-2077-2011; Email: amaerte1@jhu.edu

Authors

Emily Golden − *Center for Alternatives to Animal Testing (CAAT), Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, United States;* orcid.org/0000-0002-9760-143X

Daniel C. Ukaegbu − *Center for Alternatives to Animal Testing (CAAT), Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, United States*

Peter Ranslow − *Consortium for Environmental Risk Management (CERM), Hallowell, Maine 04347, United States*

Robert H. Brown − *School of Medicine, Johns Hopkins University, Baltimore, Maryland 21287, United States*

Thomas Hartung − *Center for Alternatives to Animal Testing (CAAT), Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, United States; CAAT-Europe, University of Konstanz, 78464 Konstanz, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrestox.2c00383

Funding

This work was funded by the Alternatives Research and Development Foundation (ARDF). E.G. was supported by a NIEHS training grant (T32 ES007141).

Notes

The authors declare the following competing financial interest(s): T.H. consults Underwriters Laboratories (UL) on computational toxicology, especially read-across, and has a share of their respective sales. He also holds stock options and consults ToxTrack LLC.

■ REFERENCES

(1) Organisation for Economic Cooperation and Devleopment (OECD). The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins, 2012; https://www.oecd-ilibrary.org/content/publication/9789264221444-en.

(2) Hoffmann, S. LLNA variability: An essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. *Altex.* **2015**, *32* (4), 379−83.

(3) Luechtefeld, T.; Maertens, A.; Russo, D. P.; Rovida, C.; Zhu, H.; Hartung, T. Analysis of publically available skin sensitization data from REACH registrations 2008−2014. *Altex.* **2016**, *33* (2), 135−48.

(4) Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility. *Toxicol. Sci.* **2018**, *165* (1), 198−212.

(5) Strickland, J.; Zang, Q.; Paris, M.; Lehmann, D. M.; Allen, D.; Choksi, N.; Matheson, J.; Jacobs, A.; Casey, W.; Kleinstreuer, N. Multivariate models for prediction of human skin sensitization hazard. *J. Appl. Toxicol.* **2017**, *37* (3), 347−360.

(6) Akhtar, A. The flaws and human harms of animal experimentation. *Camb Q Healthc Ethics.* **2015**, *24* (4), 407−19.

(7) Hartung, T. Opinion versus evidence for the need to move away from animal testing. *ALTEX - Alternatives to animal experimentation.* **2017**, *34* (2), 193−200.

(8) Meigs, L.; Smirnova, L.; Rovida, C.; Leist, M.; Hartung, T. Animal testing and its alternatives - the most important omics is economics. *Altex.* **2018**, *35* (3), 275−305.

(9) Organisation for Economic Cooperation and Devleopment (OECD). Test No. 442C: In Chemico Skin Sensitisation, 2022;

https://www.oecd-ilibrary.org/content/publication/9789264229709-en.
(10) Organisation for Economic Cooperation and Devleopment (OECD). Test No. 442D: In Vitro Skin Sensitisation, 2022; https://www.oecd-ilibrary.org/content/publication/9789264229822-en.
(11) Organisation for Economic Cooperation and Devleopment (OECD). Test No. 442E: In Vitro Skin Sensitisation, 2022; https://www.oecd-ilibrary.org/content/publication/9789264264359-en.
(12) Organisation for Economic Cooperation and Devleopment (OECD). Guideline No. 497: Defined Approaches on Skin Sensitisation, 2021; https://www.oecd-ilibrary.org/content/publication/b92879a4-en.
(13) Golden, E.; Macmillan, D. S.; Dameron, G.; Kern, P.; Hartung, T.; Maertens, A. Evaluation of the global performance of eight in silico skin sensitization models using human data. *ALTEX* **2020**, *38* (1), 33−48.
(14) Kostal, J.; Voutchkova-Kostal, A. CADRE-SS, an in Silico Tool for Predicting Skin Sensitization Potential Based on Modeling of Molecular Interactions. *Chem. Res. Toxicol.* **2016**, *29* (1), 58−64.
(15) Macmillan, D. S.; Canipa, S. J.; Chilton, M. L.; Williams, R. V.; Barber, C. G. Predicting skin sensitisation using a decision tree integrated testing strategy with an in silico model and in chemico/in vitro assays. *Regul. Toxicol. Pharmacol.* **2016**, *76*, 30−8.
(16) Borba, J. V. B.; Braga, R. C.; Alves, V. M.; Muratov, E. N.; Kleinstreuer, N.; Tropsha, A.; Andrade, C. H. Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. *Chem. Res. Toxicol.* **2021**, *34* (2), 258−267.
(17) Wilm, A.; Garcia de Lomana, M.; Stork, C.; Mathai, N.; Hirte, S.; Norinder, U.; Kühnl, J.; Kirchmair, J. Predicting the Skin Sensitization Potential of Small Molecules with Machine Learning Models Trained on Biologically Meaningful Descriptors. *Pharmaceuticals.* **2021**, *14* (8), 790.
(18) Ta, G. H.; Weng, C. F.; Leong, M. K. In silico Prediction of Skin Sensitization: Quo vadis? *Front Pharmacol.* **2021**, *12*, 655771.
(19) Alves, V. M.; Capuzzi, S. J.; Braga, R. C.; Borba, J. V. B.; Silva, A. C.; Luechtefeld, T.; Hartung, T.; Andrade, C. H.; Muratov, E. N.; Tropsha, A. A Perspective and a New Integrated Computational Strategy for Skin Sensitization Assessment. *ACS Sustainable Chemistry & Engineering.* **2018**, *6* (3), 2845−2859.
(20) Enoch, S. J.; Ellison, C. M.; Schultz, T. W.; Cronin, M. T. D. A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity. *Critical Reviews in Toxicology.* **2011**, *41* (9), 783−802.
(21) Yordanova, D.; Schultz, T. W.; Kuseva, C.; Tankova, K.; Ivanova, H.; Dermen, I.; Pavlov, T.; Temelkov, S.; Chapkanov, A.; Georgiev, M.; Gissi, A.; Sobanski, T.; Mekenyan, O. G. Automated and standardized workflows in the OECD QSAR Toolbox. *Computational Toxicology.* **2019**, *10*, 89−104.
(22) Alves, V.; Muratov, E.; Capuzzi, S.; Politi, R.; Low, Y.; Braga, R.; Zakharov, A. V.; Sedykh, A.; Mokshyna, E.; Farag, S.; Andrade, C.; Kuz'min, V.; Fourches, D.; Tropsha, A. Alarms about structural alerts. *Green Chem.* **2016**, *18* (16), 4348−4360.
(23) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55* (7), 2932−2942.
(24) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (1), 18−28.
(25) Stumpfe, D.; Hu, H.; Bajorath, J. Evolving Concept of Activity Cliffs. *ACS Omega.* **2019**, *4* (11), 14360−14368.
(26) Bell, S. M.; Phillips, J.; Sedykh, A.; Tandon, A.; Sprankle, C.; Morefield, S. Q.; Shapiro, A.; Allen, D.; Shah, R.; Maull, E. A.; Casey, W. M.; Kleinstreuer, N. C. An Integrated Chemical Environment to Support 21st-Century Toxicology. *Environmental Health Perspectives.* **2017**, *125* (5), 054501.
(27) Bell, S.; Abedini, J.; Ceger, P.; Chang, X.; Cook, B.; Karmaus, A. L.; Lea, I.; Mansouri, K.; Phillips, J.; McAfee, E.; Rai, R.; Rooney, J.; Sprankle, C.; Tandon, A.; Allen, D.; Casey, W.; Kleinstreuer, N. An integrated chemical environment with tools for chemical safety testing. *Toxicology in Vitro.* **2020**, *67*, 104916.
(28) Abedini, J.; Cook, B.; Bell, S.; Chang, X.; Choksi, N.; Daniel, A. B.; Hines, D.; Karmaus, A. L.; Mansouri, K.; McAfee, E.; Phillips, J.; Rooney, J.; Sprankle, C.; Allen, D.; Casey, W.; Kleinstreuer, N. Application of new approach methodologies: ICE tools to support chemical evaluations. *Computational Toxicology.* **2021**, *20*, 100184.
(29) National Library of Medicine. *PubChem. Score Matrix Service.*
(30) National Library of Medicine. *PubChem. Substructure Fingerprints v1.3*; 2009.
(31) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13* (11), 2498−504.
(32) Idea Consult, Ltd. *Toxtree. 3.1.0*; 2018.
(33) Enoch, S. J.; Madden, J. C.; Cronin, M. T. Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach. *SAR QSAR Environ. Res.* **2008**, *19* (5−6), 555−78.
(34) Laboratory of Mathematical Chemistry (LMC). *OECD QSAR Toolbox. 4.5*; 2021.
(35) Istituto di Ricerche Farmacologiche Mario Negri IRCCS. *VEGA. 1.2.0*; 2020.
(36) Benfenati, E.; Manganaro, A.; Gini, G. VEGA-QSAR: AI inside a platform for preditive toxicology. *2nd Workshop on Popularize Artificial Intelligence*, Turin, Italy; 2013.
(37) Kramer, O. K-Nearest Neighbors. *Dimensionality Reduction with Unsupervised Nearest Neighbors*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; pp 13−23.
(38) Basketter, D. A.; Alépée, 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* (1), 11−21.
(39) Stotts, J.; Ely, W. J. Induction of human skin sensitization to ethanol. *J. Invest Dermatol.* **1977**, *69* (2), 219−22.

https://doi.org/10.1021/acs.chemrestox.2c00383
Chem. Res. Toxicol. 2023, 36, 734−746