JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

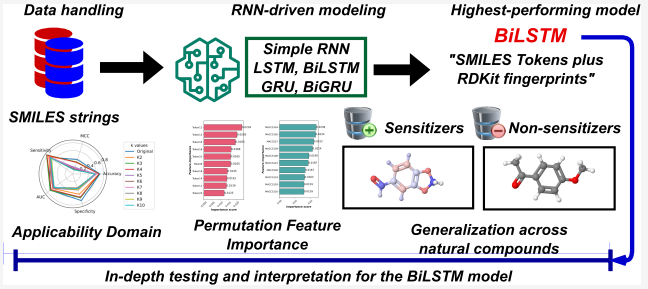
pubs.acs.org/jcim
Article

Bidirectional Long Short-Term Memory (BiLSTM) Neural Networks with Conjoint Fingerprints: Application in Predicting Skin-Sensitizing Agents in Natural Compounds

Huynh Anh Duy and Tarapong Srisongkram*

Cite This: J. Chem. Inf. Model. 2025, 65, 3035−3047
Read Online

ABSTRACT: Skin sensitization, or allergic contact dermatitis, represents a critical end point in toxicity assessment, with profound implications for drug safety and regulatory decision-making. This study aims to develop a robust deep-learning-based quantitative structure−activity relationship framework for accurately predicting skin sensitization toxicity, particularly in the context of natural-product-derived compounds. To achieve this, we explored advanced recurrent neural network architectures, including long short-term memory (LSTM), bidirectional LSTM (BiLSTM), gated recurrent unit (GRU), and bidirectional GRU, to model the intricate structure-toxicity relationships inherent in molecular compounds. We aim to optimize and improve predictive performance by training a cohort of 55 models with a diverse set of molecular fingerprints. Notably, the BiLSTM model, which integrates SMILES tokens with RDKit fingerprints, achieved superior predictive performance, underscoring its capability to effectively capture key molecular determinants of skin sensitization. An extensive applicability domain analysis coupled with an in-depth evaluation of feature importance provided new insights into the key molecular attributes that influence sensitization propensity. We further evaluated the BiLSTM model using a natural product data set, where it demonstrated exceptional generalization capabilities. The model achieved an accuracy of 86.5%, a Matthews correlation coefficient of 75.2%, a sensitivity of 100%, an area under the curve of 88%, a specificity of 75%, and an F1-score of 88.8%. Remarkably, the model effectively categorized natural products by discriminating sensitizing from non-sensitizing agents across various natural product subcategories. These results underscore the potential of BiLSTM-based models as powerful *in silico* tools for modern drug discovery efforts and regulatory assessments, especially in the field of natural products.



INTRODUCTION

Skin sensitization is most commonly associated with allergic contact dermatitis (ACD). ACD is a type IV hypersensitivity reaction arising from an activation of allergen-specific T cells in sensitized individuals, leading to inflammation within 48 h of allergen contact.[1] The evaluation of skin sensitization is a cornerstone of chemical hazard assessment which is crucial for drug development, regulatory toxicology, and safety evaluation. Historically, animal-based methods such as the guinea pig maximization test (GPMT)[2] and the local lymph node assay (LLNA)[3] have been the primary tools for assessing this end point. However, the reliance on these methods is increasingly challenged due to ethical concerns, legal restrictions, and questions about their relevance to human outcomes.[4]

Emerging new approach methodologies (NAMs), including *in silico*, *in chemico*, and *in vitro* techniques, are driving a paradigm shift toward alternative testing methods.[5] Regulatory agencies, including the U.S. Environmental Protection Agency (EPA) and the Consumer Product Safety Commission (CPSC), are actively fostering the adoption of NAMs, supporting initiatives that replace or reduce animal testing.[6] For instance, the EPA, in collaboration with the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM), has initiated efforts to establish computational and laboratory-based alternatives for skin sensitization testing.[7] Moreover, the CPSC has recognized non-animal alternatives, including prior human experience and CPSC-approved *in vitro* or *in silico* methodologies, as reliable and ethical approaches to safety testing.[8] The shift toward computational toxicology aligns with global regulatory goals and offers a pathway for improved human-relevant chemical safety assessments.

Quantitative structure−activity relationship (QSAR) modeling has emerged as a promising *in silico* approach for predicting skin sensitization.[9] By correlating molecular structure with biological activity, QSAR models offer rapid, cost-effective

Received: January 7, 2025
Revised: February 19, 2025
Accepted: February 20, 2025
Published: March 3, 2025

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

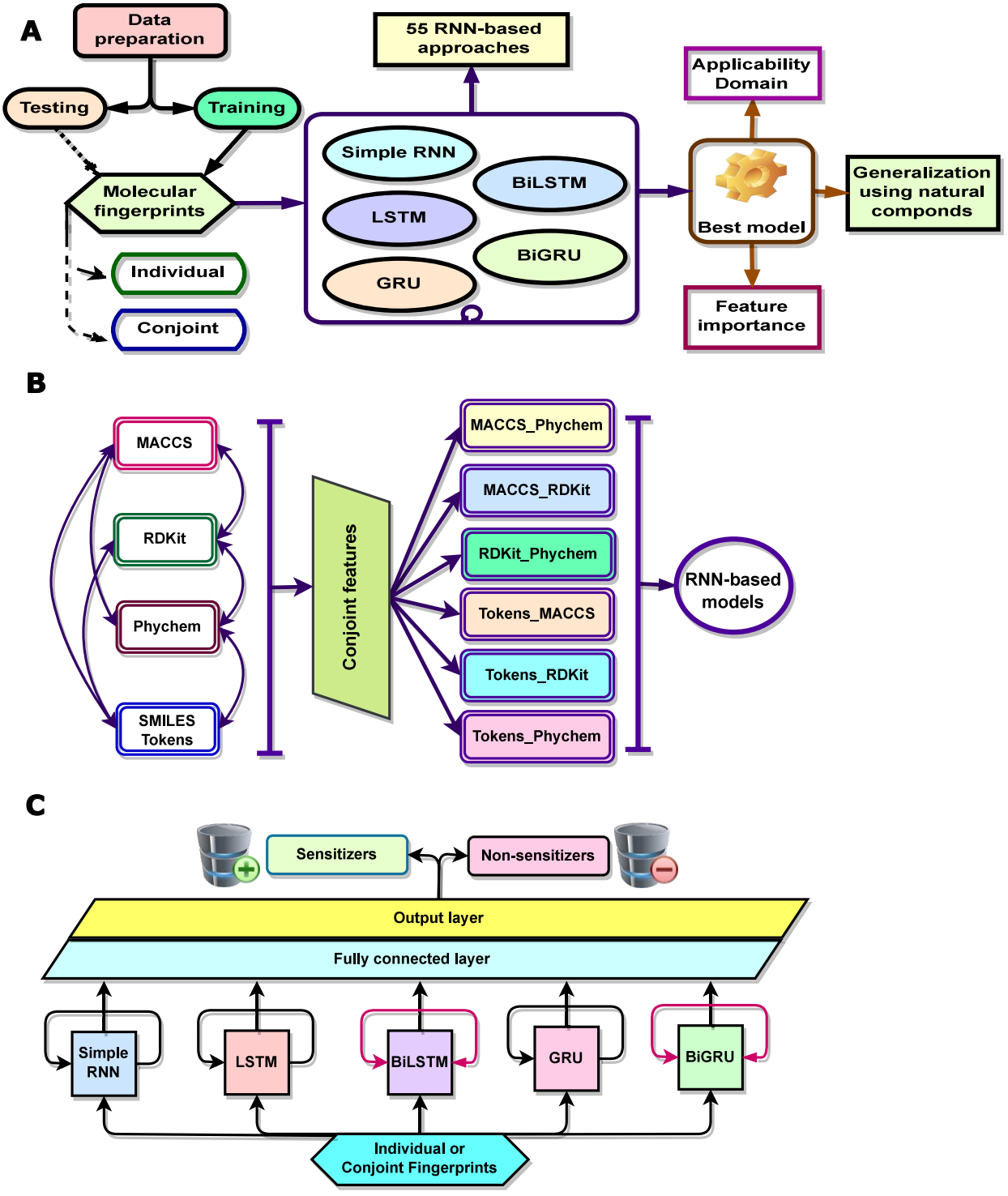


Figure 1. Schematic diagram of this study. A) Flowchart of model construction. B) Conjoint fingerprints construction. C) RNN architectures.

alternatives to animal model. Recently, advancements deep learning (DL) model, particularly in recurrent neural network (RNN) architectures have significantly enhanced QSAR predictive accuracy by offering its ability to capture long-term dependencies between molecular representations and theirs biological activities.[10]

The RNN models contains five major variants: simpleRNN, long short-term memory (BiLSTM), bidirectional LSTM (BiLSTM), gated recurrent unit (GRU), and bidirectional GRU (BiGRU). The simpleRNN is a basic neural network for sequential data processing, where the output from the previous time step is fed back into the network along with the current input, allowing it to keep information from past inputs.[11] In contrast, LSTM enhances this capability by maintaining information over extended periods through three gating mechanisms: the forget gate, input gate, and output gate, which regulate the flow of information within the network.[11] Additionally, BiLSTM is an architectural extension of LSTM that processes input sequences bidirectionally using two LSTM layers: forward and backward LSTMs. This network enables the

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

model to capture both past and future information at each time step, improving its pattern recognition.[12]

Moreover, the GRU was developed as an advanced variant of the simpleRNN that incorporates two key gates: the update gate and the reset gate. The update gate regulates the retention of past information, enabling the model to preserve relevant historical data. The reset gate, conversely, controls the degree to which prior information is discarded, facilitating the integration of new inputs into the memory of the model.[11] Lastly, the BiGRU architecture enhances the standard GRU by processing input sequence in both forward and reverse directions, granting the network access to both past and future information for each time step, similar to the BiLSTM model.[13]

The current state-of-the-art RNN-based models have excel in analyzing sequential data such as molecular representation (e.g., SMILES) or molecular fingerprints,[14−16] making them promise in addressing complex toxicological end points. In 2015, Xu et al. introduced a DL-based model designed to predict drug-induced liver injury (DILI), a toxic adverse drug reaction that leads to liver damage. In their approach, molecular structures were initially represented as small undirected graphs and subsequently transformed into acyclic graphs for processing in RNN. At the time of publication, the model demonstrated the highest performance in terms of sensitivity and specificity.[14] Peng et al. introduced the TOP model which is a DL-based approach designed for chemical toxicity prediction using the Tox21 database. This model integrates a BiGRU-based RNN with fully connected neural networks to enable end-to-end molecular representation learning. TOP can capture a hybrid molecular representation by integrating both SMILES contextual information and physicochemical properties, making it achieved a significant enhancement in toxicity prediction accuracy.[15] Our recent BiLSTM models that utilize MACCS key fingerprints and physicochemical descriptors also have shown performance improvement in skin corrosion prediction compared to the state-of-the-art model.[16]

Building upon these advancements, our study extends the application of RNN architectures to the prediction of skin sensitization toxicity. We aim to develop a robust RNN framework for assessing the skin sensitization potential as well as deriving key structural determinants to provide mechanistic insights into this critical end point. This work aligns with the regulatory vision of transitioning to alternative testing approaches, contributing to the growing field of computational toxicology and advancing the safety assessment of chemical compounds. The flowchart of this study is illustrated in Figure 1A. The principal contributions of this research are as follows:

1. We newly designed and rigorously evaluated 55 RNN-based predictive models, leveraging Simple RNN, LSTM, BiLSTM, GRU, and BiGRU architectures specifically for skin sensitization prediction. These models utilized both individual and conjoint molecular features derived from physicochemical descriptors, atomic environments, predefined substructures, and character-level encodings of SMILES strings, ensuring a comprehensive multidimensional representation of molecular structures relevant to the skin sensitization.
2. We improved skin sensitization prediction reliability and generalizability by using Euclidean distance-based applicability domain (AD). This assessed structural similarity between new compounds and a subset k of training data, ensuring trustable predictions within a defined chemical space.
3. We extracted the key molecular determinants using permutation feature importance, highlighting the most influential features for predicting skin sensitization and providing insights into its underlying mechanisms.
4. Finally, we explored the applicability of RNN-based models to predict the skin sensitization potential of untested natural compounds. The BiLSTM model achieved 86.5% accuracy, showcasing the effectiveness of RNNs in computational toxicology and their potential for chemical safety assessment.

EXPERIMENTAL METHODS

Data Preparation. This study utilized a comprehensive data set comprising chemical compound names and SMILES strings, with skin sensitization data obtained from two authoritative references: (a) the National Toxicology Program's Interagency Center for the Evaluation of Alternative Toxicological Methods, and (b) the publicly available Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH) study results database. This comprehensive data set was meticulously extracted from a previous study.[17] Skin sensitization outcomes were classified into binary toxicity categories of sensitizers and non-sensitizers, based on the Globally Harmonized System of Classification and Labeling of Chemicals (GHS) hazard classes.[18] Data preprocessing involved several steps, including the generation of canonical SMILES representations, removal of inorganic compounds, exclusion of compound mixtures, and elimination of duplicate entries. Following data preprocessing, the 996 chemical entries were divided into training and testing sets in a 70:30 ratio, respectively. A scaffold-based splitting strategy was employed to generate the training and test sets, ensuring a robust evaluation by accounting for molecular diversity. This split was implemented using the astartes.molecules module within the open-source Python library *Astartes*.[19] The sampling parameter was configured as scaffold to prioritize structural dissimilarity between subsets, while the seed parameter was fixed at 0 to guarantee reproducibility across different experimental runs.

Molecular Feature Encoding. Initially, five individual molecular fingerprints were utilized to construct RNN-based models, including Extended Connectivity Fingerprints (ECFP), physicochemical descriptors (such as Molecular Weight, LogP, Number of Hydrogen Bond Donors, Number of Hydrogen Bond Acceptors, Topological Polar Surface Area, Number of Rotatable Bonds, Number of Aromatic Rings, Number of Saturated Rings, Number of Heteroatoms, Ring Count, Heavy Atom Count, and Number of Aliphatic Rings), MACCS keys, RDKit fingerprints, and SMILES token representations. All of these features were computed by using RDKit Python package.[20]

We further conducted the conjoint descriptors based on the model performance of each fingerprint. A molecular fingerprint that exhibited the lowest predictive performance were eliminated from the analysis. The remaining high-performing fingerprints were then combined to create six conjoint features aimed at enhancing the overall predictive accuracy of the models. This approach allowed for an optimized feature set that leveraged the complementary representation across different high performance fingerprint types. As such, we developed six conjoint fingerprints from MACCS, RDKit, Physicochemical

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

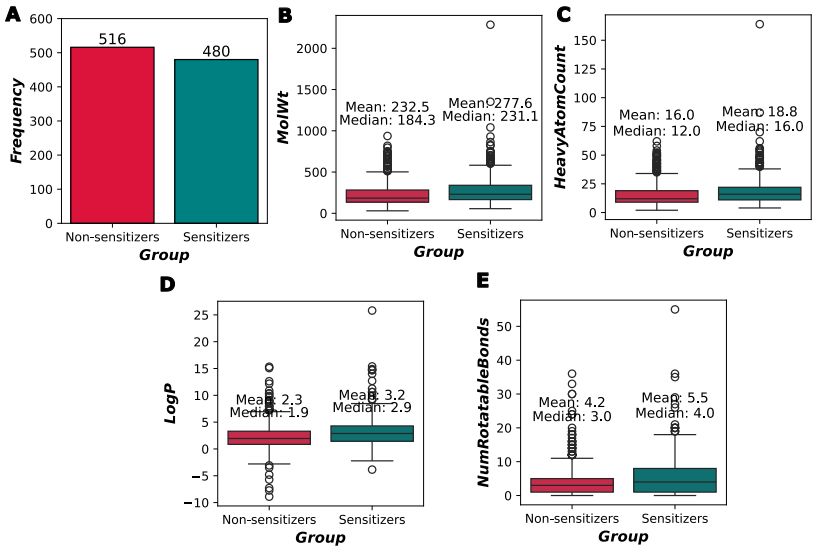


Figure 2. Chemical distribution between sensitizers and non-sensitizers used in this study. A) Class distribution between sensitizers and non-sensitizers. B) Molecular weight (MolWt) distribution between sensitizers and non-sensitizers. C) Non-hydrogen atoms (HeavyAtomCount) distribution between sensitizers and non-sensitizers. D) LogP distribution between sensitizers and non-sensitizers. D) Number of Rotatable Bounds (NumRotatableBonds) distribution between sensitizers and non-sensitizers.

descriptors (PhyChem), and SMILES tokens as illustrated in Figure 1B.

RNN Architectures. A simple RNN model was constructed with a single RNN layer, consisting of an input layer, a hidden layer, and an output layer. The hidden layer contained 64 units, while the output layer consisted of a single neuron with a sigmoid activation function and a threshold of 0.5 for binary classification. For GRU and LSTM models, we used two recurrent layers, utilizing either GRU or LSTM units. The first layer consisted of 64 units, followed by a second recurrent layer with 32 units. After the recurrent layers, a fully connected layer with 100 neurons was implemented, using the ReLU activation function to introduce non-linearity. The output layer contained a single neuron with a sigmoid activation function, applying a threshold of 0.5 to classify the predictions. For the BiGRU and BiLSTM models, the network architectures are similar to the GRU and LSTM but we used BiGRU and BiLSTM instead of those GRU and LSTM layers. For all models, the learning rate was set to 0.001 and the binary crossentropy was employed as the loss function. To optimize the model, the training set was split into a subtraining set and a subvalidation set with a 7:3 ratio using the *validation_split* parameter available in TensorFlow/Keras. The splitting process was carried out randomly, without stratification. As no specific random seed was assigned during the split, the resulting subsets can vary between different executions. The Adam Optimizer, an Adaptive Moment Estimation (ADAM) algorithm, was used for gradient descent optimization. The model was trained over 50 epochs to ensure adequate learning time. The model configuration is represented in Figure 1C.

All models share a similar architecture, consisting of an input layer, recurrent layers, a fully connected layer, and an output layer, but differ in the structure of their recurrent layers. The simple RNN uses basic recurrent units to propagate information sequentially but is limited in learning long-term dependencies due to the vanishing gradient problem. GRU and LSTM address this limitation through gating mechanisms. GRU uses update and reset gates for efficient information flow, while LSTM introduces forget, input, and output gates along with a cell state, allowing it to capture more complex patterns. Although LSTM is more expressive, GRU offers computational simplicity. BiGRU and BiLSTM enhance GRU and LSTM by processing sequences in both forward and backward directions, thereby capturing richer contextual information from the entire sequence, which improves performance on complex sequence classification tasks. The simple RNN serves as a baseline model, while GRU, LSTM, BiGRU, and BiLSTM are designed to handle increasingly complex sequential dependencies. In addition, the methodologies for performance analysis, statistical assessment, applicability domain (AD) analysis, and permutation feature importance are described in the Supporting Information.

RESULTS

Representation of Chemical Space. We analyzed 12 physicochemical properties of compounds classified as skin sensitizers (n = 480) and non-sensitizers (n = 516). The Kruskal–Wallis test identified four parameters—molecular weight (MolWt), number of non-hydrogen atoms (HeavyAtomCount), logarithm of the octanol–water partition coefficient (LogP), and the number of rotatable bonds (NumRotatableBonds)—as statistically significant (p-values <0.05) (Supplementary Table S1). The median values of these parameters were significantly higher in sensitizers compared to non-sensitizers: 231.1 vs 184.3 for MolWt, 16.0 vs 12.0 for HeavyAtomCount, 2.9 vs 1.9 for LogP, and 4.0 vs 3.0 for NumRotatableBonds, respectively (Figure 2). These findings suggest that specific physicochemical features, such as increased molecular size and complexity, are associated with skin sensitization potential.

To explore the chemical distribution of compounds, we applied t-distributed stochastic neighbor embedding or t-SNE to visualize various molecular fingerprints and descriptors, as shown in Figure 3. The results demonstrated that SMILES tokens provided the clearest separation between sensitizers and non-sensitizers, as depicted in Figure 3D, suggesting a strong potential of SMILES tokens to differentiate sensitizer and non-sensitizer compounds. In contrast, other molecular features, such as ECFP, MACCS, RDKit fingerprints, and physicochem-

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

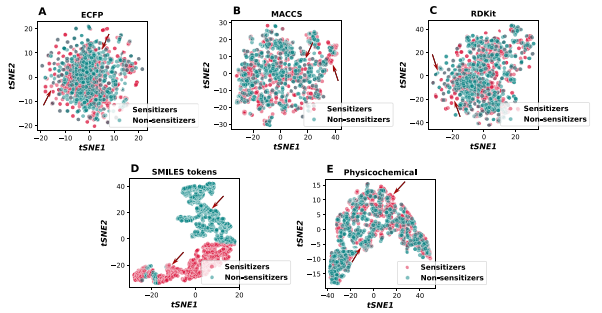


Figure 3. Molecular features distribution of the data set using A) Extended circular fingerprints (ECFP). B) MACCS keys fingerprints. C) RDKit fingerprints. D) SMILES tokens. E) Physicochemical descriptors. Red arrows indicate the unique non-overlap island of chemicals in each group.

ical descriptors, showed less distinct separation between these two classes (Figures 3A–C and 3E), indicating that it may show low classifying potential compared to the SMILES tokens. These preliminary findings underscore the need for further analysis to determine the most suitable molecular feature for optimal predictive performance.

Predictive Performance of RNN Models on the Test Data Set. To investigate predictive performance, we initially developed 25 distinct prediction models leveraging five individual molecular descriptors and five RNN-based architectures. The models' performance based on individual fingerprints was thoroughly analyzed using test data sets comprising both in-domain and out-of-domain compounds. The performance metrics, summarized in Figure 4A, reveal a range of outcomes: accuracy between 0.558 and 0.642, MCC between 0.135 and 0.265, sensitivity from 0.437 to 0.707, AUC from 0.588 to 0.676, and specificity between 0.451 and 0.773. While the models demonstrated a moderate predictive capability, the results were suboptimal in which accuracy and MCC values remaining low across all models. Notably, the LSTM model using the ECFP fingerprint achieved the highest sensitivity (0.707), but exhibited the lowest values for accuracy (0.558), AUC (0.588), and specificity (0.451), alongside a modest MCC of 0.16. Furthermore, all RNN models utilizing ECFP consistently showed the weakest performance in terms of accuracy scores. Given these findings, we sought to improve model performance by constructing conjoint features derived from the combination of multiple individual fingerprints, excluding ECFP due to its limited predictive contribution. This approach resulted in the creation of 30 new predictive models based on six conjoint feature sets, incorporating five RNN architectures as depicted in Figure 1B.

Adopting a similar approach, models employing conjoint fingerprints were extensively evaluated using test data sets spanning both in-domain and out-of-domain chemical spaces. The predictive performance of the models incorporating conjoint features is presented in Figure 4B, with accuracy ranging from 0.551 to 0.701, MCC from 0.116 to 0.385, sensitivity from 0.515 to 0.74, AUC from 0.574 to 0.714, and specificity from 0.509 to 0.729. The integration of conjoint features resulted in notable improvements in predictive performance, yielding increases of 2.5%, 13.5%, 4.9%, 2.0%, and 1.2% across the respective metrics compared to models using individual features. This finding further reinforces the effectiveness of conjoint features in enhancing the robustness of computational prediction models. Among the various approaches, the BiLSTM model using the conjoint combination of SMILES tokens and RDKit fingerprints achieved the highest accuracy (0.701) and MCC (0.385), positioning it as the most effective model. This result indicates that the model correctly predict 70% of skin sensitizers and non-sensitizers. Additionally, with sensitivity and specificity values of 0.74 and 0.72, respectively, the model accurately identifies 74% of skin sensitizers and 72% of non-sensitizers. The AUC value of 0.714 indicates a moderate ability to rank between positive and negative instances, demonstrating reasonable discriminatory power.

Furthermore, the BiLSTM model demonstrated the highest sensitivity (0.74) among all evaluated models. In the context of toxicity prediction, higher sensitivity is indicative of the model's enhanced ability to correctly identify the positive class (sensitizers), thus minimizing false negatives—where a

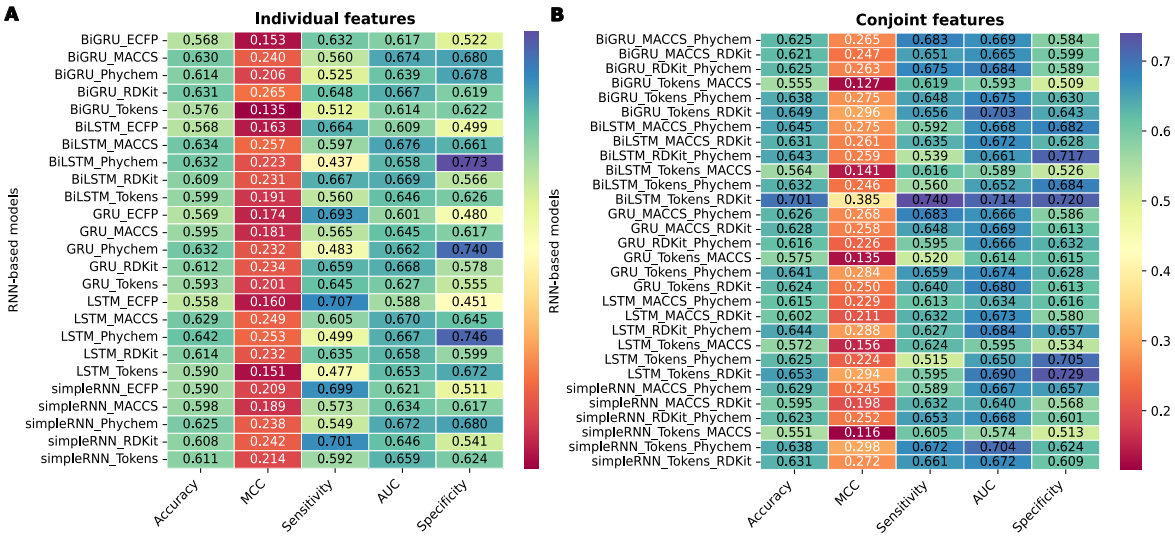


Figure 4. Predictive performance of recurrent neural network (RNN) models for skin sensitization: A) Individual features. B) Conjoint features. All metrics are retrieved from the mean of three separate experiments. The full evaluation model performance of all models is shown in Supplementary Table S5.

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

sensitizer is misclassified as a non-sensitizer. Clinical data from 2013 to 2015 reveal that about 25% of drug development failures during Phase II and Phase III trials were attributed to unpredictable toxicity, highlighting the critical need of accurate toxicity detection.[21] These findings are particularly relevant in safety-critical applications, such as drug development, chemical safety assessments, and environmental monitoring. Additionally, the BiLSTM model's predictive performance aligns with chemical distribution analysis data, where SMILES tokens effectively differentiated sensitizers from non-sensitizers (Figure 3D). Based on these results, we propose the BiLSTM model with SMILES tokens and RDKit fingerprints, as the most effective tool for skin sensitization prediction.

AD Analysis. The AD analysis is a fundamental concept in computational toxicology and QSAR modeling.[22−24] The AD defines the chemical space where the predictive model works reliably, ensuring that predictions apply only to compounds similar to the training data. This strengthens the model's validity and usefulness. In this study, the AD was meticulously assessed using an iterative optimization of the *k*-nearest neighbors (*k*NN) technique, based on Euclidean distance metrics. The performance of the AD was rigorously evaluated by varying the *k*-value from 2 to 10, with the AD defined based on the training set. This approach enabled accurate differentiation of out-of-domain compounds in the independent test set while ensuring that only structurally relevant in-domain chemicals were retained for subsequent predictive analysis. This refined test set was subsequently analyzed using our BiLSTM model, facilitating a direct comparison of the predictive performance between the optimized and unrefined data sets. The *k*-value yielding the highest predictive efficacy was identified as optimal for defining the AD. Figure 5A provides a comprehensive visualization of the AD assessment of skin sensitization, demonstrating the impact of varying *k*-values on model reliability and robustness.

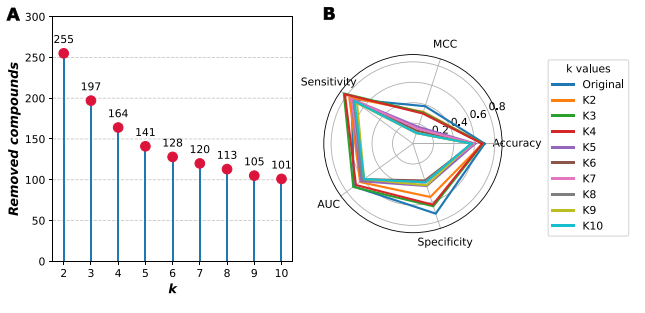


Figure 5. Performance evaluation of the BiLSTM model using within applicability domain compounds from test set. A) Number of removed compounds. B) Evaluation metrics with various *k*-values.

Figure 5B illustrates the evaluation metrics obtained from the within AD test set across various *k*-values, providing insights into the AD determination of the BiLSTM model. At $k = 1$, all test compounds are excluded, so no parameters for $k = 1$ are displayed, indicating that the AD is overly restrictive and unsuitable for practical predictions. At $k = 3$, with 197 compounds excluded, the BiLSTM model achieves an accuracy of 0.683, a MCC of 0.319, and a specificity of 0.644, reflecting declines of 2.6%, 17.1%, and 10.6%, respectively, compared to the original test set. Conversely, the model exhibits a notable increase in sensitivity of 12.9% (reaching 0.835) and a modest improvement in AUC by 1.0% (achieving 0.72). This substantial increase in sensitivity is particularly impactful, as it reflects the model's enhanced ability to identify true skin sensitizers. The slight gain in AUC further supports the model's improved ranking ability across varying thresholds. These findings underscore that the $k = 3$ configuration effectively refines the model's AD, excluding 197 compounds that likely lie outside the training data's chemical space. This exclusion reduces the risk of erroneous predictions for structurally dissimilar compounds while preserving a strong capacity for accurate sensitizer identification.

Key Chemical Substructures Connected to Skin Sensitization. The permutation feature importance technique was utilized for measuring the change in model performance when the values of a single feature are randomly shuffled. This strategy can extract features that are important for the machine learning (ML) model. Also, a major strength of this method is its model-agnostic nature,[25] enabling its application to any trained estimator, regardless of the underlying model structure. In this analysis, the importance score was derived from the average decrease in accuracy after shuffling each feature 30 times. Features with a small accuracy drop were considered less important, whereas those that caused a significant reduction in performance were regarded as more influential in the model's predictions. This approach ensures a comprehensive and reliable assessment of feature relevance.

We systematically analyzed the importance scores to identify and prioritize the top 10 influential features contributing to the predictive performance of the BiLSTM model which integrates SMILES tokens and RDKit fingerprints as conjoint features. Remarkably, all top-ranked features were derived exclusively from the SMILES token group, underscoring their critical role in the predictive framework (Figure 6A). Among these, SMILES

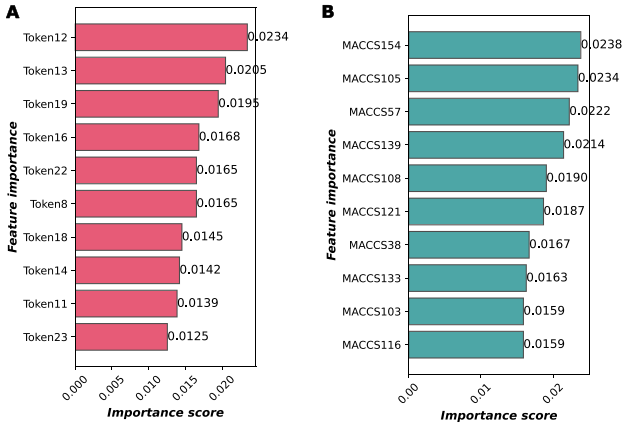


Figure 6. Feature importance for skin sensitization prediction. A) Essential SMILES Tokens linked to skin sensitization prediction. B) Key substructures associated with skin sensitization (based on SMILES Token12 analysis).

token 12 emerged as the most impactful feature, exhibiting the highest importance score across all conjoint features (Figure 6A). Notably, 83.2% of the compounds within the AD contained this token. We identified chemical compounds that contain all 10 key SMILES tokens: 12, 13, 19, 16, 22, 8, 18, 14, 11, and 23. This rigorous filtering process identified 8 compounds, classified as 5 sensitizers and 3 non-sensitizers, designated as compounds (1) through (8) in Figure 7.

To further elucidate the significance of SMILES token No. 12, its associated substructures were investigated, revealing functional groups with the highest importance scores, as illustrated in

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

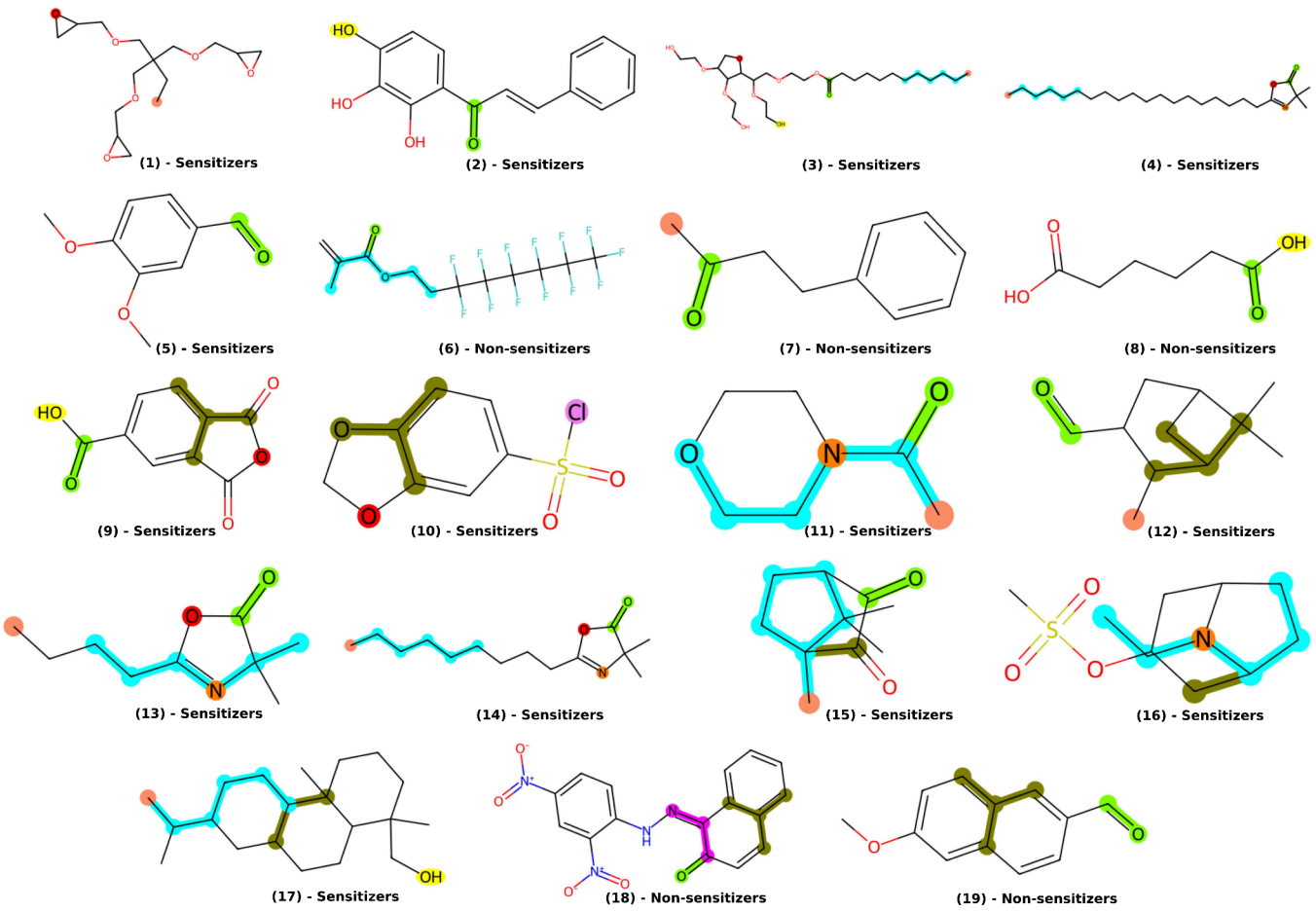


Figure 7. Compounds within the domain containing key substructures. Notes: MACCS154: Chartreuse (yellow-green) color; MACCS105: Olive color; MACCS57: Red color; MACCS139: Yellow color; MACCS108: Cyan color; MACCS121: Orange color; MACCS38: Crimson color; MACCS133: Magenta color; MACCS103: Violet color; MACCS116: Salmon color.

Figure 6B. A subset of in-domain compounds from the test data set containing SMILES token No. 12 was systematically filtered and mapped to their respective MACCS fingerprints. This refined data set underwent analysis using a BiLSTM model combined with the permutation feature importance method, enabling the identification and ranking of MACCS fingerprints with the highest significance. These findings highlight the crucial molecular determinants underpinning skin sensitization. We summarizes the top 10 key substructures identified as most influential in predicting skin sensitization in the Supplementary Table S2. The MACCS fingerprint, specifically the 154th bit, emerges as the most significant feature in the model, associated with the presence of carbonyl groups in the molecular structure. Additionally, oxygen-containing heterocycles (MACCS 57), hydroxyl groups (MACCS 139), nitrogen-containing functional groups (MACCS 121, 38, and 133), and chlorine-containing compounds (MACCS 103) play notable roles in the skin sensitization prediction. Additionally, chirality or stereochemical configuration (MACCS 105), methyl and methylene groups in both linear configurations (MACCS108) and cyclic structures (MACCS116) further emphasize the diversity of chemical features impacting skin sensitization. These results illustrate the model's capacity to accurately identify key molecular determinants relevant to skin sensitization mechanisms. Finally, 11 compounds with the most frequent MACCS substructures were selected to additional illustration in Figure 7, including 9 sensitizers (No. 9−17) and 2 non-sensitizers (No. 18 and 19).

In summary, a comprehensive set of 19 compounds, including 14 sensitizers and 5 non-sensitizers, was curated to visualize the importance features in Figure 7. Importantly, all selected compounds fall within the AD of the BiLSTM model, reinforcing the robustness and predictive accuracy of the model in identifying skin sensitization risks. Among the identified compounds, two natural sensitizers were recognized: (2) − 2′,3′,4′-trihydroxychalcone (a flavonoid), and (15) − camphorquinone (a monoterpenoid). These findings suggest the BiLSTM model's utility in predicting sensitizing agents, including natural compounds, and its potential for further analyses.

Generalization Evaluation of the BiLSTM Model on Test Data Generated by Other Models. Generalization is a fundamental property of ML models, representing their ability to make accurate predictions on unseen data, thereby ensuring their applicability to real-world problems.[26] In this study, we assess the generalization performance of our BiLSTM model by applying it to predict skin sensitization using test sets generated from other ML and DL models. These data sets are sourced from relevant studies published in the last five years, selected based on data available through PubMed database. To ensure consistency, all test sets undergo the same preprocessing and normalization procedures. The BiLSTM model is then employed to generate predictions, with its performance thoroughly evaluated using the conjoint features of SMILES Tokens and RDKit. This methodology allows us to rigorously

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

assess the BiLSTM model's ability to generalize to diverse data and its potential utility in predicting skin sensitization. The results of the generalization evaluation of the model are shown in Table 1.

Table 1. Predictive Capabilities of the BiLSTM Model with the External Data Derived from Other Models

No.	Models	ACC (%)	MCC (%)	Sens. (%)	AUC (%)	Spec. (%)	F1 score (%)
1	BiLSTM using the Pred-Skin's test set[27]	**79.1**	**59.3**	**87.3**	**85.8**	**71.5**	**80.1**
	The Pred-Skin model	66	-	81	-	52	77.9
2	BiLSTM using the Skin Doctor CP's test set[28]	**76.7**	**57.1**	**90.7**	**83.4**	**66.7**	**76.4**
	The Skin Doctor CP model	75	51	81	-	70	72.1
3	BiLSTM using the StopTox's test set[17]	**70.1**	**38.5**	**74**	**71.4**	**72**	**68.5**
	The StopTox model	70	-	66	-	75	68.4
4	BiLSTM using test set from Yuxuan Hu et al, 2023[29]	**72**	**37.5**	**77.4**	**70.8**	**66.7**	**81.3**
	Support Vector Machine (SVM)	98	94	-	95	-	99
	Random Forest (RF)	85	51	-	66	-	91
	Logistic Regression (LR)	82	39	-	60	-	90
	XGBoost (XGB)	79	20	-	53	-	88
	Chemprop	78	19	-	64	-	87

The comparative analysis between our BiLSTM model and the Pred-Skin baseline[27] model demonstrates a significant improvement in predictive performance for skin sensitization classification. The BiLSTM model achieves 79.1% accuracy, surpassing the 66% accuracy of Pred-Skin, indicating superior overall classification capability. This improvement is particularly notable in sensitivity (87.3%), which highlights the BiLSTM's strong ability to identify sensitizers, compared to the 81% sensitivity of the Pred-Skin model. The BiLSTM also excels in specificity (71.5%), effectively identifying non-sensitizers and reducing false positives, while the Pred-Skin model shows 52% specificity. The F1 score of 80.1% for the BiLSTM, compared to 77.9% for PredSkin, reflects a more balanced model with better precision and sensitivity. Furthermore, the AUC of 85.8% demonstrates excellent ranking power, and the MCC of 59.3% reflects a moderate correlation between predicted and actual values, indicating good model reliability. In contrast, the Pred-Skin model does not report these critical metrics, making it harder to assess its full performance. These findings position the BiLSTM model as a more reliable and promising tool for predicting skin sensitization in toxicity screening.

Next, our BiLSTM model outperforms SkinDoctor CP[28] across several key metrics. The BiLSTM achieves an accuracy of 76.7%, slightly higher than SkinDoctor CP's 75%. Importantly, the BiLSTM excels in sensitivity (90.7%), indicating a strong ability to identify sensitizers, which is critical for toxicological predictions. However, its specificity of 66.7% is lower than that of SkinDoctor CP, which has a specificity of 70%, suggesting that SkinDoctor CP is slightly more effective in correctly identifying non-toxic compounds. The AUC of 83.4% further highlights the model's strong ranking performance. The BiLSTM's F1 score of 76.4% reflects a well-balanced trade-off between precision and sensitivity. In contrast, SkinDoctor CP has lower sensitivity (81%) and its AUC is unavailable for comparison. The F1 score of 72.1% suggests a weaker balance between precision and sensitivity. Overall, our BiLSTM model outperforms SkinDoctor CP in sensitivity and AUC, making it a more reliable and effective model for toxicity prediction, particularly in applications where accurate identification of sensitizers is crucial.

Subsequently, we compared the BiLSTM model with StopTox,[17] emphasizing their comparable overall accuracy while revealing notable disparities in their individual performance metrics. The BiLSTM model achieved an accuracy of 70.1%, marginally surpassing StopTox's accuracy of 70%. It exhibited a sensitivity of 74%, a critical metric for minimizing false negatives in toxicity predictions. In terms of specificity, the BiLSTM scored 72%, reflecting a reasonably balanced ability to identify non-toxic compounds. In contrast, StopTox demonstrated lower sensitivity at 66%, suggesting a reduced capacity to correctly identify toxic compounds, although it excelled in specificity with a score of 75%, indicating superior accuracy in detecting non-toxic compounds. The BiLSTM's F1 score of 68.5% highlighted a strong balance between precision and sensitivity, whereas StopTox's F1 score was slightly lower at 68.4%. However, the absence of MCC and AUC data for StopTox restricts a comprehensive assessment of its discriminative performance. Conversely, the BiLSTM model displayed a MCC of 38.5% and an AUC of 71.4%, indicating reasonable ranking capability and reliability in distinguishing between sensitizers and non-sensitizers across various thresholds.

The performance of the BiLSTM model was evaluated alongside several ML models from Yuxuan Hu et al.,[29] including support vector machine (SVM), random forest (RF), logistic regression (LR), extreme gradient boosting (XGB), and Chemprop (a message-passing neural network), using a range of performance metrics. The BiLSTM model achieved an accuracy of 72%, which, although lower than that of SVM (98%), RF (85%), LR (82%), XGBoost (79%), and Chemprop (78%), exhibited a noteworthy MCC of 37.5%. This suggests that the BiLSTM model demonstrates a robust overall ability to differentiate between actutal and predicted instances, despite its lower accuracy with the Yuxuan Hu data set. The MCC value, though not the highest, is more balanced compared to models such as XGBoost (20%) and Chemprop (19%). Additionally, the BiLSTM model displayed a sensitivity of 77.4% and specificity of 66.7%, highlighting its strong capability in classifying sensitizers and non-sensitizers within the data set. In contrast, other models did not report these critical metrics, limiting the ability to conduct a comprehensive performance comparison. Sensitivity and specificity are pivotal for toxicity prediction models due to their relevance in identifying both false positives and false negatives. In terms of the AUC, the BiLSTM model achieved 70.8%, reflecting a reasonable ranking performance for sensitizers and non-sensitizers. This AUC score is competitive when compared to the other models: RF (66%), LR (60%), XGBoost (53%), and Chemprop (64%). The BiLSTM's F1 score of 81.3% underscores its balance between precision and sensitivity, though it remains lower than the scores of other models, suggesting a fine-tune with this new test set may require for further study. In summary, although the BiLSTM model may not surpass all other models in specific performance metrics, it stands out for its ability to deliver balanced predictions for both sensitizers and non-sensitizers. The positive results from the

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

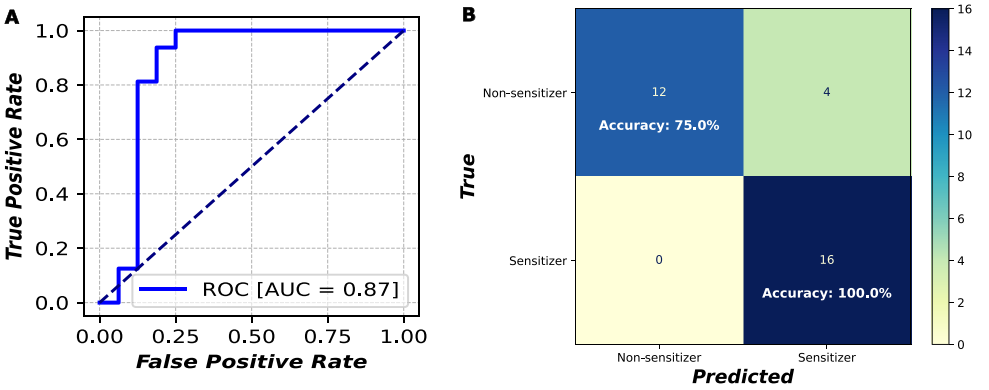


Figure 8. Performance of BiLSTM model using conjoint feature of SMILES Tokens and RDKit fingerprints on the test set from natural compounds. A) The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC). B) Confusion matrix.

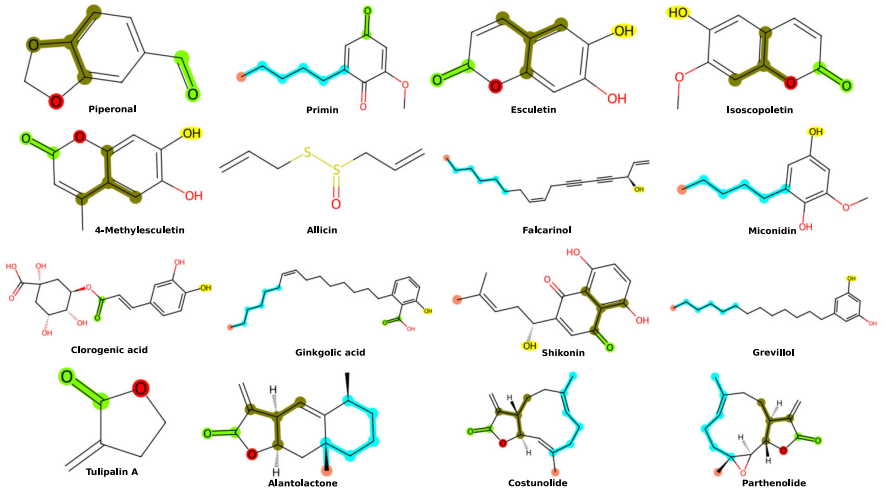


Figure 9. Natural compounds recognized as skin sensitizers with essential substructures. Notes: MACCS154: Chartreuse (yellow-green) color; MACCS105: Olive color; MACCS57: Red color; MACCS139: Yellow color; MACCS108: Cyan color; MACCS121: Orange color; MACCS38: Crimson color; MACCS133: Magenta color; MACCS103: Violet color; MACCS116: Salmon color.

generalization analysis strengthen the model's potential for continued application in further experiments.

Using the BiLSTM Model to Predict Skin Sensitization from Natural Products. We employed the BiLSTM model to predict the skin sensitization properties of natural compounds that carefully curated from previous studies.[30−33] This data set included compounds that the model had not previously encountered, ensuring a comprehensive evaluation of its generalizability and usability. The compounds were classified into sensitizers and non-sensitizers based on well-established *in vivo* experimental data, providing a robust benchmark for performance assessment. In total, 32 natural compounds representing diverse subclasses were selected, comprising 16 sensitizers and 16 non-sensitizers. These compounds were processed and normalized according to standardized methods. The predicted probabilities for each compound were calculated, with compounds exhibiting probabilities greater than 0.5 classified as sensitizers and those below 0.5 as non-sensitizers. The results of these predictions for the natural compounds are summarized in Supplementary Table S3, offering valuable insights into the model's performance.

We also presented the results of applying the BiLSTM model to predict the toxicity of natural compounds in Supplementary Table S4, pointing out its strong predictive performance across various evaluation metrics. With an accuracy of 86.5%, the model demonstrates its ability to correctly classify a significant proportion of both sensitizers and non-sensitizers. Notably, the MCC of 75.2% reflects a well-balanced classification performance, suggesting that the model not only excels in predicting one class but also discriminates effectively between both positive and negative cases. A perfect sensitivity of 100% stresses the importance of the model's exceptional ability to correctly identify all sensitizers, thereby minimizing the risk of false negatives. However, the specificity of 75% suggests that a quarter of the non-toxic compounds were misclassified as toxic, presenting an opportunity for further refinement to reduce false positives. The F1 score of 88.8% offers a comprehensive reflection of the model's ability to balance precision and sensitivity. A high F1 score signifies a balanced trade-off between these two metrics, ensuring that the model performs well in both detecting toxic compounds and minimizing misclassifications. The BiLSTM model demonstrated strong ranking ability in distinguishing natural sensitizers from non-sensitizers, achieving an impressive AUC score of 0.87. The ROC curve, along with the corresponding AUC value depicted in Figure 8A. The per-class predictive outcomes, illustrated by the confusion matrix (Figure 8B), reaffirm its effectiveness in identifying compounds. These findings demonstrate that the BiLSTM model serves as a reliable instrument for predicting the skin sensitization of natural compounds.

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

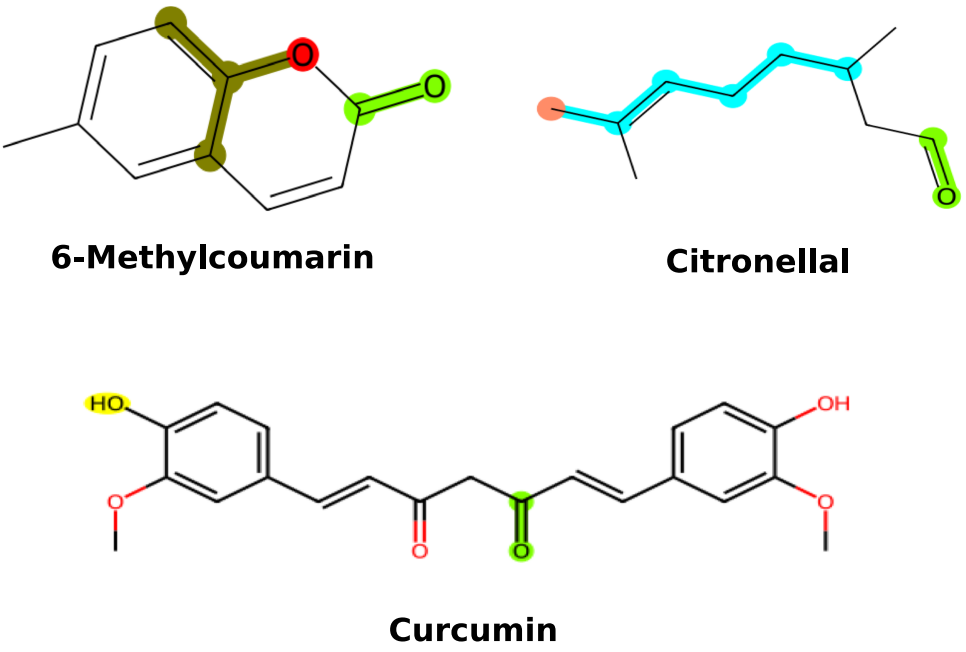


Figure 10. 6-Methylcoumarin, citronellal, and curcumin, highlighting their major substructures. Notes: MACCS154: Chartreuse (yellow-green) color; MACCS105: Olive color; MACCS57: Red color; MACCS139: Yellow color; MACCS108: Cyan color; MACCS121: Orange color; MACCS38: Crimson color; MACCS133: Magenta color; MACCS103: Violet color; MACCS116: Salmon color.

The classification results of the BiLSTM model for natural compounds were further analyzed by visualizing the structures of those identified as skin sensitizers, focusing on the presence of key substructures known to contribute to skin sensitization (Figure 9). Interestingly, while most natural sensitizers contained the major substructures identified by the model, an exception was observed in Allicin, a compound flagged by the BiLSTM model. Despite lacking the key substructures, Allicin's corrected classification highlights the model's ability to capture the sensitization using SMILES and RDKIT fingerprints, emphasizing its potential for uncovering previously unrecognized sensitizers in natural compounds.

DISCUSSION

In this study, we developed a BiLSTM model leveraging conjoint fingerprints derived from SMILES tokens and RDKit-generated molecular descriptors. This innovative approach enabled highly accurate predictions of dermal toxicity in both chemical and natural compounds. Furthermore, comparative analyses on diverse external data sets demonstrated the superior performance of the BiLSTM model over existing methodologies. Notably, the integration of SMILES tokens and RDKit fingerprints resulted in exceptional sensitivity metrics, underscoring the model's efficacy in identifying toxic compounds (true positives). Enhanced sensitivity in predictive toxicology is critical, as it directly improves the capacity to detect potential toxic agents, reducing the risk of late-stage drug candidate failures. Toxicity represents a significant cost driver in drug development, particularly when toxic effects are identified during advanced clinical trials or postmarket phases.[34] By addressing these challenges, the BiLSTM model provides a transformative tool for identifying skin toxicity agents, including skin sensitizers, reinforcing its potential impact in predictive toxicology and drug discovery.

In our permutation feature importance analysis, carbonyl groups were identified as the most influential substructure affecting the predictive performance of our model. This observation is particularly intriguing in the context of their involvement in the biological mechanism of skin sensitization, which consists of two primary phases: induction and elicitation. During the induction phase, the sensitizing chemical penetrates the *stratum corneum* and reaches the viable epidermis, where it forms covalent bonds with skin proteins or peptides, generating immunogenic complexes. For skin sensitization to occur, the chemical must first be sufficiently bioavailable, allowing it to partition into the relevant intra-, extra-, and subcellular compartments of the epidermis. Additionally, the chemical or its metabolite must possess electrophilic properties to react with nucleophilic groups on skin proteins, facilitating the formation of hapten-protein conjugates that are recognized by the immune system as foreign, triggering sensitization.[35,36] Carbonyl groups, due to their electrophilic nature, readily form covalent bonds with nucleophilic protein residues, contributing to skin sensitization.[37,38] These interactions occur through mechanisms like Schiff-base formation, Michael addition, and nucleophilic substitution.[35,39] Our BiLSTM model effectively captures these chemical features, improving predictive accuracy and mechanistic understanding.

Moreover, the model reclassified three natural compounds, including 6-methylcoumarin (a coumarin), citronellal (a monoterpene), and curcumin (a phenolic compound) from non-sensitizers to sensitizers, leading to several notable insights. The 6-Methylcoumarin (6-MC) faced classification challenges due to conflicting assay results. While multiple *in vitro* and *in vivo* tests reported negative findings, positive responses were observed in KeratinoSens, LuSens assays, and *in silico* predictions.[40,41] Thereby, regulatory agencies, such as the California Safe Cosmetics Program[42] and the European Toys Safety Directive,[43] recognize 6-MC as a fragrance allergen. Our BiLSTM model accurately classified 6-MC as a skin sensitizer, aligning with its established allergenic potential. Similarly, citronellal showed contradictory results in skin sensitization

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

studies. It was sensitizing in guinea pig models but did not induce sensitization in the LLNA test or a human maximization test.[44] Despite this, structural alerts suggest its potential to bind skin proteins and trigger an immune response.[45] Notably, the Hazardous Substances Data Bank (HSDB) identifies citronellal as the primary allergen responsible for lemongrass oil-related hypersensitivity cases.[46] The BiLSTM model's classification of citronellal as a sensitizer is consistent with the HSDB class, further supporting its predictive reliability. Curcumin is widely recognized for its therapeutic potential in dermatology, with clinical trials confirming its efficacy and tolerability.[47,48] Although LLNA studies classify curcumin as non-allergenic,[31] other research indicates it may act as a contact allergen.[49,50] Our BiLSTM model corroborates these concerns by identifying curcumin as a skin sensitizer, highlighting the necessity for cautious application despite its therapeutic benefits. Figure 10 illustrates the chemical structures of 6-methylcoumarin, citronellal, and curcumin, highlighting their major substructures.

Our study introduces several key innovations that distinguish it from previous research in the predictive modeling of skin sensitization end points. First, unlike prior studies that predominantly relied on traditional ML models such as RF and SVM,[28,29,51] our approach leverages the RNN-based approach. This method represents significant advancement by capturing sequential dependencies within molecular representations, which traditional ML models often overlook.[15] Second, we introduce a novel feature engineering strategy by integrating conjoint features derived from SMILES tokens and RDKit fingerprints. To the best of our knowledge, this is the first instance of conjoint feature utilization in this context, demonstrating its potential to enhance model performance through richer molecular representations. Third, our study applies BiLSTM for the identification of skin sensitizers in compounds of natural origin, expanding the scope of DL in toxicology and cheminformatics. These methodological benefits enhance the potential of RNN for improved predictive accuracy and broader applicability in skin sensitization research.

Therefore, this study makes a significant contribution to the field by enhancing predictive performance in toxicity screening compared to existing models. Our model improves overall accuracy and sensitivity (a crucial metric in minimizing false negatives) while providing comprehensive MCC and AUC data. By reducing false negatives, it mitigates the risk of misclassifying harmful compounds as safe, a critical factor in safety assessments. Additionally, our model demonstrates superior predictive capabilities on natural product data sets, achieving over 85% accuracy and 100% sensitivity. Notably, it effectively classifies compounds with conflicting toxicity data (e.g., 6-methyl coumarin, citronellal, and curcumin) by capturing intricate chemical relationships. This enhanced accuracy and robustness mark a significant advancement over existing methodologies, substantially increasing the reliability of toxicity assessment, particularly in the context of natural products.

Advantages and Limitations. Our method utilizing the BiLSTM model with conjoint features of SMILES tokens and RDKit fingerprints has shown remarkable accuracy and reliability in skin sensitization prediction. This method represents a significant contribution to toxicity screening, particularly in pharmaceutical and natural product development. However, certain limitations warrant attention. The complexity of conjoint feature interpretation present challenges. Moreover, the model's predictive capacity is confined to organic compounds, restricting its broader applicability. Furthermore, its specificity in recognizing non-sensitizers is lower than for sensitizers, highlighting the need for model optimization to improve overall predictive balance.

Future Directions. Future research could prioritize expanding and diversifying data sets of natural compounds to strengthen the reliability and robustness of predictive models. Furthermore, this model can be applied in phytochemical screening, allowing researchers to assess compound toxicity directly from GC-MS or LC-MS data using the provided Python script. This approach is particularly useful for natural product and plant extract development.

CONCLUSION

In conclusion, this study highlights the potential of BiLSTM models in predicting skin sensitization, leveraging the conjoint features of SMILES tokens and RDKit fingerprints for enhanced accuracy. The model not only identifies key molecular substructures driving sensitization but also integrates a rigorously defined applicability domain framework, ensuring reliable predictions across diverse chemical landscapes. Notably, the inclusion of natural products expands the chemical space, reinforcing the model's robustness and broadening its utility in toxicity screening. This work exemplifies the power of advanced ML methodologies in predictive toxicology, offering a scientifically robust, scalable solution for safer and more efficient compound evaluation in both pharmaceutical and natural product development applications. The complex interpretability, the need for improved specificity, and its predictions being currently limited to organic compounds represent its current limitations, highlighting opportunities for future development and broader applications.

DATA AND SOFTWARE AVAILABILITY

All data used in this manuscript and script to reproduce this research can be downloaded from the following URL: https://github.com/taraponglab/sensitizationrnn. Questions or comments can be sent to the corresponding author of this work by email.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c00032.

The methodologies for performance analysis, statistical assessment, applicability domain (AD) analysis, permutation feature importance, and Supplementary Tables S1−S5 [Supplementary Table S1: Kruskal−Wallis test results for physicochemical descriptors; Supplementary Table S2: Key substructures associated with skin sensitization (based on SMILES Token12 analysis); Supplementary Table S3: Predictive performance of the BiLSTM model for identifying skin sensitizers from natural compounds; Supplementary Table S4: BiLSTM model performance on an external data set of natural compounds; Supplementary Table S5: The full evaluation model performance of all models] (PDF)

AUTHOR INFORMATION

Corresponding Author

Tarapong Srisongkram − *Division of Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Khon Kaen*

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047

University, Khon Kaen 40002, Thailand; orcid.org/0000-0001-8512-5379; Email: tarasri@kku.ac.th

Author

Huynh Anh Duy − *Graduate School in the Program of Research and Development in Pharmaceuticals, Faculty of Pharmaceutical Sciences, Khon Kaen University, Khon Kaen 40002, Thailand; Department of Health Sciences, College of Natural Sciences, Can Tho University, Can Tho 900000, Vietnam*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c00032

Author Contributions

Conceptualization, T.S.; methodology, T.S.; data curation, H.A.D.; formal analysis, H.A.D.; software, H.A.D. and T.S.; validation, H.A.D. and T.S.; investigation, H.A.D. and T.S.; resources, T.S.; writing original draft, H.A.D. and T.S.; writing review and editing, T.S.; visualization, H.A.D. and T.S.; supervision, T.S.; project administration, T.S.; funding acquisition, T.S.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research on skin sensitization prediction has received funding support from the Fundamental Fund of Khon Kaen University from National Science, Research, and Innovation Fund or NSRF, Thailand (Project no. 68A103000041). H.A.D. would like to thank the KKU Scholarship for ASEAN and GMS Countries' Personnel, Faculty of Pharmaceutical Science, Khon Kaen University, for funding.

■ REFERENCES

(1) Rashid, R. S.; Shim, T. N. Contact dermatitis. *Bmj* **2016**, *353*, i3299.
(2) OECD Test No. 406: Skin Sensitisation; 2022.
(3) OECD Test No. 442B: Skin Sensitization; 2024.
(4) Van Norman, G. A. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is it Time to Rethink Our Current Approach? *JACC Basic Transl Sci.* **2019**, *4*, 845−854.
(5) Husain, A.; Meenakshi, D. U.; Ahmad, A.; Shrivastava, N.; Khan, S. A. A Review on Alternative Methods to Experimental Animals in Biological Testing: Recent Advancement and Current Strategies. *J. Pharm. Bioallied Sci.* **2023**, *15*, 165−171.
(6) Strickland, J.; et al. Skin sensitization testing needs and data uses by US regulatory and research agencies. *Arch. Toxicol.* **2019**, *93*, 273−291.
(7) EPA Interim Science Policy: Use of alternative approaches for skin sensitization as a replacement for laboratory animal testing. 2018.
(8) CPSC (2012). Recommended Procedures Regarding the CPSC's Policy on Animal Testing.
(9) Alves, V. M.; Muratov, E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicol. Appl. Pharmacol.* **2015**, *284*, 262−272.
(10) Cavasotto, C. N.; Scardino, V. Machine Learning Toxicity Prediction: Latest Advances by Toxicity End Point. *ACS Omega* **2022**, *7*, 47536−47546.
(11) Pavlatos, C.; Makris, E.; Fotis, G.; Vita, V.; Mladenov, V. Utilization of Artificial Neural Networks for Precise Electrical Load Prediction. *Technologies* **2023**, *11*, 70.
(12) Pavlatos, C.; Makris, E.; Fotis, G.; Vita, V.; Mladenov, V. Enhancing Electrical Load Prediction Using a Bidirectional LSTM Neural Network. *Electronics* **2023**, *12*, 4652.
(13) Kim, H.; Lee, J.; Moon, S.; Kim, S.; Kim, T.; Jin, S. W.; Kim, J. L.; Shin, J.; Lee, S. U.; Jang, G. Visual field prediction using a deep bidirectional gated recurrent unit network model. *Sci. Rep.* **2023**, *13*, 11154.
(14) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep learning for drug-induced liver injury. *J. Chem. Inf. Model.* **2015**, *55*, 2085−2093.
(15) Peng, Y.; Zhang, Z.; Jiang, Q.; Guan, J.; Zhou, S. TOP: a deep mixture representation learning method for boosting molecular toxicity prediction. *Methods* **2020**, *179*, 55−64.
(16) Duy, H. A.; Srisongkram, T. Comparative Analysis of Recurrent Neural Networks with Conjoint Fingerprints for Skin Corrosion Prediction. *J. Chem. Inf. Model.* **2025**, *65*, 1305.
(17) Borba, J. V. B.; Alves, V. M.; Braga, R. C.; Korn, D. R.; Overdahl, K.; Silva, A. C.; Hall, S. U. S.; Overdahl, E.; Kleinstreuer, N.; Strickland, J.; Allen, D.; Andrade, C. H.; Muratov, E. N.; Tropsha, A. STopTox: An in Silico Alternative to Animal Testing for Acute Systemic and Topical Toxicity. *Environ. Health Perspect* **2022**, *130*, 27012.
(18) GHS. *Globally Harmonized System of Classification and Labelling of Chemicals*; GHS: New York and Geneva, 2011.
(19) Burns, J. W.; Green, W. H. (2023) Better Data Splits for Machine Learning with astartes.
(20) Landrum, G. (2016) RDKit: Open-Source Cheminformatics Software.
(21) Harrison, R. K. Phase II and phase III failures: 2013−2015. *Nat. Rev. Drug Discov* **2016**, *15*, 817−818.
(22) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graph Model* **2008**, *26*, 1315−26.
(23) Boonsom, S.; Chamnansil, P.; Boonseng, S.; Srisongkram, T. ToxSTK: A multi-target toxicity assessment utilizing molecular structure and stacking ensemble learning. *Comput. Biol. Med.* **2025**, *185*, 109480.
(24) Srisongkram, T. DeepRA: A novel deep learning-read-across framework and its application in non-sugar sweeteners mutagenicity prediction. *Comput. Biol. Med.* **2024**, *178*, 108731.
(25) Molnar, C.; König, G.; Bischl, B.; Casalicchio, G. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. *Data Mining and Knowledge Discovery* **2024**, *38*, 2903−2941.
(26) Tian, Y.; Zhang, Y. A comprehensive survey on regularization strategies in machine learning. *Information Fusion* **2022**, *80*, 146−166.
(27) 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*, 258−267.
(28) Wilm, A.; Norinder, U.; Agea, M. I.; de Bruyn Kops, C.; Stork, C.; Kühnl, J.; Kirchmair, J. Skin Doctor CP: Conformal Prediction of the Skin Sensitization Potential of Small Organic Molecules. *Chem. Res. Toxicol.* **2021**, *34*, 330−344.
(29) Hu, Y.; Ren, Q.; Liu, X.; Gao, L.; Xiao, L.; Yu, W. In Silico Prediction of Human Organ Toxicity via Artificial Intelligence Methods. *Chem. Res. Toxicol.* **2023**, *36*, 1044−1054.
(30) Rozas-Muñoz, E.; Lepoittevin, J. P.; Pujol, R. M.; Giménez-Arnau, A. Allergic Contact Dermatitis to Plants: Understanding the Chemistry will Help our Diagnostic Approach. *Actas Dermo-Sifiliográficas* **2012**, *103*, 456−477.
(31) Yamashita, K.; Miyazaki, H.; Shinoda, S.; Hagiwara, S.; Takahashi, H.; Itagaki, H. Assessment of the skin sensitizing potential of chemicals, contained in foods and/or cosmetic ingredients, using a modified local lymph node assay with an elicitation phase (LLNA:DAE) method. *J. Toxicol Sci.* **2018**, *43*, 513−520.
(32) Lin, M.; Gong, W.; Chen, Q.; Sun, L.; Wang, Y.; Fan, X. Evaluation of the potential sensitization of chlorogenic Acid: a meta-analysis. *Evid Based Complement Alternat Med.* **2013**, *2013*, 208467.

(33) Xu, T.; Wu, L.; Xia, M.; Simeonov, A.; Huang, R. Systematic identification of molecular targets and pathways related to human organ level toxicity. *Chemical research in toxicology* **2021**, *34*, 412−421.
(34) Guengerich, F. P. Mechanisms of drug toxicity and relevance to pharmaceutical development. *Drug metabolism and pharmacokinetics* **2011**, *26*, 3−14.
(35) Patlewicz, G.; Aptula, A. O.; Roberts, D. W.; Uriate, E. (2007) Skin sensitisation (Q) SARs/expert systems: From past, present to future. *European Commission report EUR 21866*.
(36) Smith Pease, C.; Basketter, D.; Patlewicz, G. Contact allergy: the role of skin chemistry and metabolism. *Clinical and Experimental Dermatology* **2003**, *28*, 177−183.
(37) Roberts, D. W.; Aptula, A. O.; Patlewicz, G. Mechanistic Applicability Domains for Non-Animal Based Prediction of Toxicological Endpoints. QSAR Analysis of the Schiff Base Applicability Domain for Skin Sensitization. *Chem. Res. Toxicol.* **2006**, *19*, 1228−1233.
(38) Patlewicz, G.; Basketter, D. A.; Smith, C. K.; Hotchkiss, S. A.; Roberts, D. W. Skin-sensitization structure-activity relationships for aldehydes. *Contact Dermatitis* **2001**, *44*, 331−6.
(39) Roberts, D.; Patlewicz, G. Mechanism based structure-activity relationships for skin sensitisation−the carbonyl group domain. *SAR and QSAR in Environmental Research* **2002**, *13*, 145−152.
(40) Kolle, S. N.; Natsch, A.; Gerberick, G. F.; Landsiedel, R. A review of substances found positive in 1 of 3 in vitro tests for skin sensitization. *Regul. Toxicol. Pharmacol.* **2019**, *106*, 352−368.
(41) Urbisch, D.; Mehling, A.; Guth, K.; Ramirez, T.; Honarvar, N.; Kolle, S.; Landsiedel, R.; Jaworska, J.; Kern, P. S.; Gerberick, F.; Natsch, A.; Emter, R.; Ashikaga, T.; Miyazawa, M.; Sakaguchi, H. Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regulatory toxicology and pharmacology: RTP* **2015**, *71*, 337−351.
(42) EU 2009/48/EC of the European Parliament and of the Council as regards the labelling of allergenic fragrances in toys. 2009; https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=oj:JOL_2009_170_R_0001_01.
(43) EU amending Regulation (EC) No 1223/2009 of the European Parliament and of the Council as regards labelling of fragrance allergens in cosmetic products. 2023; https://eur-lex.europa.eu/eli/reg/2023/1545/oj.
(44) Api, A. M.; et al. RIFM fragrance ingredient safety assessment, citronellal, CAS registry number 106−23−0. *Food Chem. Toxicol.* **2021**, *149*, 111991.
(45) Roberts, D. W.; Patlewicz, G.; Kern, P. S.; Gerberick, F.; Kimber, I.; Dearman, R. J.; Ryan, C. A.; Basketter, D. A.; Aptula, A. O. Mechanistic Applicability Domain Classification of a Local Lymph Node Assay Dataset for Skin Sensitization. *Chem. Res. Toxicol.* **2007**, *20*, 1019−1030.
(46) HSDB Hazardous Substances Data Bank (HSDB). 2024; https://pubchem.ncbi.nlm.nih.gov/source/hsdb/594.
(47) Chainani-Wu, N. Safety and anti-inflammatory activity of curcumin: a component of tumeric (Curcuma longa). *J. Altern Complement Med.* **2003**, *9*, 161−8.
(48) Di Lorenzo, R.; Forgione, F.; Bernardi, A.; Sacchi, A.; Laneri, S.; Greco, G. Clinical Studies on Topical Curcumin. *Skin Pharmacol Physiol* **2024**, *36*, 235−248.
(49) Chaudhari, S. P.; Tam, A. Y.; Barr, J. A. Curcumin: A Contact Allergen. *J. Clin Aesthet Dermatol* **2015**, *8*, 43−8.
(50) Babu, V. A. A clinical study on allergic contact dermatitis to turmeric. *Journal of Evolution of Medical and Dental Sciences* **2013**, *2*, 3388−3407.
(51) Guo, W.; Liu, J.; Dong, F.; Song, M.; Li, Z.; Khan, M. K. H.; Patterson, T. A.; Hong, H. Review of machine learning and deep learning models for toxicity prediction. *Exp Biol. Med. (Maywood)* **2023**, *248*, 1952−1973.

https://doi.org/10.1021/acs.jcim.5c00032
J. Chem. Inf. Model. 2025, 65, 3035−3047