Computational Toxicology 38 (2026) 100410

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.sciencedirect.com/journal/computational-toxicology

Full Length Article

Development and evaluation of explainable QSAR models to predict chemical-induced respiratory irritation

Pinyi Lu [a,b], Souvik Dey [a,b], Anders Wallqvist [a,*], Mohamed Diwan M. AbdulHameed [a,b,*]

[a] Department of War Biotechnology High Performance Computing Software Applications Institute, Defense Health Agency Research & Development, Medical Research and Development Command, Fort Detrick, MD 21702, USA
[b] The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD 20817, USA

ARTICLE INFO

Keywords:
Chemical-induced respiratory irritation
Human risk assessment
New approach methodologies
Quantitative structure-activity relationship
Explainable artificial intelligence

ABSTRACT

Chemical-induced respiratory irritation is a critical human health concern because of its potential to cause both acute and chronic injuries. Therefore, assessing the respiratory irritation potential of chemicals is crucial for protecting human health. Due to limited human data, human risk assessment often relies on extrapolating animal data, which can be imprecise and resource intensive. New approach methodologies are being developed to reduce the reliance on animal testing and streamline chemical risk assessment methods. In the present study, we developed a computational workflow to integrate explainable artificial intelligence (XAI) approaches into the quantitative structure–activity relationship (QSAR) modeling pipeline to predict chemical-induced respiratory irritation. We developed and assessed multiple QSAR models by combining different machine learning algorithms and diverse molecular representations. The models achieved an average cross-validated area under the receiver operating characteristic curve of 0.88, accuracy of 0.80, and Matthews correlation coefficient of 0.61, with external test set evaluation indicating good generalizability. We applied different methods, including Shapley additive explanations (SHAP), to explain the predictions of our QSAR models. By providing both global and local explanations of model predictions, the SHAP analyses highlighted key molecular descriptors and fingerprints driving respiratory irritation predictions, revealing physicochemical properties that may provide insights into irritation potential. The explainable models developed in this work have the potential to be an alternative screening tool for traditional animal models to obtain faster and more cost-effective human risk assessment of respiratory irritants.

1. Introduction

Chemical-induced respiratory irritation is a critical human health concern because of its potential to cause both acute and chronic injuries [1], leading to adverse health effects, such as coughing, chest pain, and shortness of breath. Prolonged exposure to respiratory irritants can cause more severe or long-term respiratory diseases, including reactive airway dysfunction syndrome, chemical pneumonitis, and chronic bronchitis. Respiratory irritant exposure can occur in various settings, such as in the workplace, at home, and outdoors, as chemicals in pesticides, cleaning products, and smoke may cause respiratory irritation. As a precautionary measure, international regulatory bodies require companies to classify and label chemicals that cause respiratory irritation. For example, the Globally Harmonized System of Classification and Labelling of Chemicals (GHS) assigns the hazard code H335 to substances that may cause respiratory irritation [2]. Military personnel in combat zones face unique risks of chemical-induced respiratory irritation from exposure to various agents, such as adamsite, diphenylchloroarsine, diphenylcyanoarsine, and chlorine [3]. These agents have previously been referred to as vomiting agents, and they can cause irritation of the respiratory tract as well as other symptoms (e.g., lacrimation, sneezing, nausea, and irritation of the eyes). Thus, the assessment of chemical respiratory irritants is crucial for protecting both civilian and military populations.

Human risk assessment of respiratory irritants is often a multi-step process that relies on toxicological data. The first step, i.e., hazard identification, determines whether a chemical can cause respiratory irritation based on data from human or animal exposure studies as well

* Corresponding authors at: DoW Biotechnology High Performance Computing Software Applications Institute, Defense Health Agency Research & Development, Medical Research and Development Command, ATTN: FCMR-TT, 504 Scott Street, Fort Detrick, MD 21702-5012, USA.
E-mail addresses: sven.a.wallqvist.civ@health.mil (A. Wallqvist), mabdulhameed@bhsai.org (M.D.M. AbdulHameed).

https://doi.org/10.1016/j.comtox.2026.100410
Received 13 January 2026; Received in revised form 11 March 2026; Accepted 12 March 2026
Available online 16 March 2026

as real-world data gathered from poison control center reports, occupational exposure incidents, or human clinical case studies. The second step, i.e., exposure assessment, estimates the degree to which humans are exposed to a respiratory irritant, including the frequency and duration of the exposure as well as the chemical's concentration and physicochemical properties. The third step sets safe human exposure levels. However, due to limited human data, determining safe human guidelines often relies on extrapolating animal data, which can be imprecise and resource intensive because of species-specific differences in genetics, metabolism, and anatomy. Consequently, high failure rates in human trials can occur despite promising animal data. To reduce the reliance on animal testing and streamline chemical risk assessment methods, new approach methodologies (NAMs), such as *in vitro* testing and computational modeling, are being developed [4,5]. For example, quantitative structure–activity relationship (QSAR) models can be applied for predictive toxicology, improving the efficiency of chemical risk assessment.

To date, there have been several QSAR models developed for predicting human respiratory irritation [1,6,7]. Fisher et al. evaluated the published QSAR models for organ-specific predictive toxicology, including respiratory toxicity [8]. They summarized the limitations of the published models and underscored the necessity for modifying existing organ-specific QSAR models or developing novel models to improve chemical risk assessment to meet both industrial and military needs [8]. For example, the RespiraTox model showed a high sensitivity for respiratory irritation predictions but was imbalanced with a relatively low specificity, indicating a high likelihood of producing false positives [1]. Another limitation of the published QSAR models for organ-specific predictive toxicology is the lack of model explainability. While published QSAR models that predict respiratory irritation may have identified influential molecular features, to our knowledge, none has explored SHAP analyses to provide both global and local explanations linking these features to the physicochemical properties that may govern irritation potential. Model explainability focuses on making the complex models more transparent and allows users to understand how the models made their predictions. As reported by Jiménez-Luna et al., the following aspects are desirable in terms of model explainability: *1*) transparency, knowing how the system reached a particular answer; *2*) justification, elucidating why the answer provided by the model is acceptable; *3*) informativeness, providing new information to human decision-makers; and *4*) uncertainty estimation, quantifying how reliable a prediction is [9].

To fill the knowledge gap of model explainability, in the present study we developed a computational workflow to integrate explainable artificial intelligence (XAI) approaches into the QSAR modeling pipeline and used it to create explainable QSAR models to predict chemical-induced respiratory irritation. We developed and assessed multiple QSAR models by combining different machine learning (ML) algorithms and molecular representations and applied one widely used XAI method, i.e., Shapley additive explanations (SHAP), to explain the predictions of the QSAR models. SHAP is an XAI approach used to interpret the predictions of ML models. Based on game theory, SHAP quantifies the contribution of each input feature to a model's prediction, enabling both global feature importance analysis and local, instance-level explanations. In this study, we applied SHAP to the trained ML models to identify the key molecular descriptors and fingerprints driving the prediction of respiratory irritation. Our SHAP analyses revealed the inner working of our respiratory irritation models and provided both global and local explanations of model predictions, which enhanced model transparency and prioritized the key molecular descriptors and fingerprints that contributed most to the models' predictions.

2. Materials and methods

We developed a computational workflow for building explainable QSAR models to predict chemical-induced respiratory irritation and explain the models' predictions. Fig. 1 shows the five key steps in our workflow, including data preprocessing, QSAR model design, QSAR model training, QSAR model evaluation, and QSAR model explanation.

2.1. Dataset collection and preparation

The respiratory irritant data that we used in the present study were previously compiled by Chushak et al. from publicly available databases [7], including three data sources: *1*) the Classification and Labeling Inventory of the European Chemicals Agency, *2*) the GHS database of the Japanese National Institute of Technology and Evaluation, and *3*) the GHS database of the Australian Hazardous Chemicals Information System. Respiratory irritants were searched using the GHS hazard code H335 and the Specific Target Organ Toxicity-Single Exposure class category 3. The nonirritant data were originally retrieved from the Organization for Economic Cooperation and Development eChemPortal. A limitation of this study is that the source databases only provide the regulatory GHS criteria for chemicals and do not provide details on the specific bioassays used to determine the irritancy classification of each compound.

We downloaded the publicly available dataset, which included 623 respiratory irritants and 615 nonirritants, from the Online Chemical Modeling Environment (OCHEM) website (https://ochem.eu/article/114857, accessed on July 18, 2025). We developed a molecular standardization pipeline using RDKit's MolStandardize module and applied it to the compounds in the dataset. Molecular standardization is a crucial step in our workflow because it formats compounds based on predefined rules, removes salts and solvents from compounds to create their parent molecules, and ensures that compounds are represented in a uniform way. Following the molecular standardization process, we removed duplicates based on the standardized molecular structures in the simplified molecular input line entry system (SMILES) format, generating a curated dataset that consisted of 617 respiratory irritants and 609 nonirritants.

2.2. Molecular representation

A key prerequisite for developing QSAR models is to convert molecules into a computer-readable format. To represent molecules numerically, we employed four different methods [molecular access system keys (MACCSKeys), Morgan, Mordred, and RDKit], which can be classified as molecular fingerprints and molecular descriptors (Table 1). Molecular descriptors represent physicochemical properties of molecules, which can be calculated using various tools. Moriwaki et al. developed Mordred, a descriptor-calculation software application, to calculate a large number of two- and three-dimensional molecular descriptors [10]. We applied Mordred (community-maintained version 2.0.6) to calculate 1,613 two-dimensional molecular descriptors and removed any descriptors that yielded missing values. In addition, using RDKit (version 2023.9.5), we obtained a set of 210 commonly used molecular descriptors, such as molecular weight, logarithm of octanol–water partition coefficient (logP), and various topological indices. We standardized the computed Mordred and RDKit descriptors by removing the mean and scaling to unit variance using the StandardScaler function of scikit-learn (version 1.3.0). In addition, we calculated the correlation matrices of the Mordred and the RDKit descriptors, and we removed the descriptors with a maximum absolute correlation of one within each matrix to reduce redundant information in the training set.

Fingerprints are another type of molecular representation that focus on structural features (e.g., substructures and topological paths) and represent molecular structures as bit strings. Morgan fingerprints, also known as extended-connectivity fingerprints, are binary vectors where each bit represents the presence or absence of a substructure in a molecule [11]. The number of bits used determines the fingerprint length, while the defined radius around each atom determines the size of

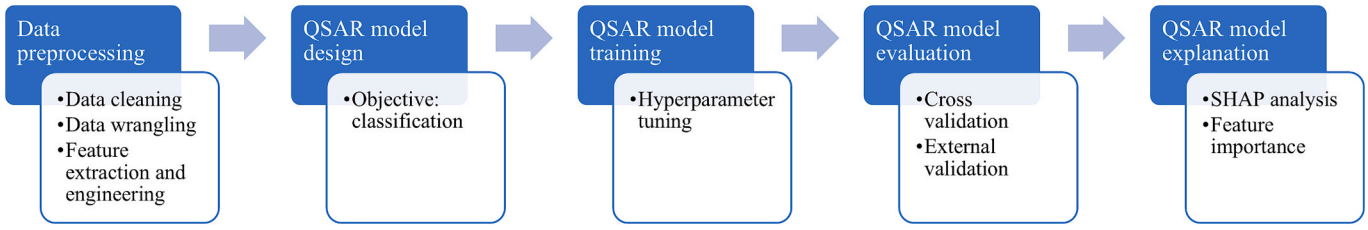


Fig. 1. Computational workflow for building explainable quantitative structure–activity relationship (QSAR) models.

Table 1
Comparison of molecular representation methods.

	Fingerprint	Descriptor
Characteristic Information	Binary representations Structure (e.g., ring)	Numerical values Property (e.g., MW, logP)
Type (no. of features)	MACCSKeys (166), Morgan (2,048)	Mordred (1,613), RDKit (210)

LogP: logarithm of octanol–water partition coefficient; MACCSKeys: molecular access system keys; MW: molecular weight.

the substructures considered. We chose 2,048 bits and a radius of two to generate Morgan fingerprints for each compound. MACCSKeys are a predefined set of 166 structural keys, and each bit of MACCSKeys fingerprints represents the presence or absence of one substructure defined as a specific SMILES arbitrary target specification pattern [12]. We created Morgan and MACCSKeys fingerprints using RDKit (version 2023.9.5).

2.3. Data exploration

t-Distributed stochastic neighbor embedding (t-SNE) is a widely applied dimensionality reduction method specifically used to visualize high-dimensional data in a lower dimensional space while preserving local relationships between data points. To visualize the chemical space of the preprocessed dataset represented in Morgan fingerprints, a high-dimensional vector of 2,048 bits, we first used another dimensionality reduction method, i.e., principal component analysis (PCA), to reduce the number of dimensions to 50 as recommended [13], which could speed up the computation of pairwise similarities between data points and suppress noise. We calculated the t-SNE of Morgan fingerprints using the scikit-learn implementation of t-SNE (version 1.3.0) and created three-dimensional scatter plots to explore patterns and relationships within the Morgan fingerprints of respiratory irritants and nonirritants. In addition, we compared respiratory irritants and nonirritants based on individual molecular properties, such as molecular weight, number of acceptors, and number of rotatable bonds. We performed Mann-Whitney U tests ($\alpha = 0.05$) using the stats module of SciPy (version 1.15.2) to determine if there were any statistically significant differences between the two classes of compounds.

2.4. Model development and optimization

To assess the models' generalizability, we reserved 20% of the curated dataset as a test set and only trained the models on the remaining 80%. We applied two common data-splitting strategies to divide the preprocessed dataset into training and test sets. The random split approach randomly distributes molecules, while the scaffold split method groups molecules based on their scaffolds, i.e., their core chemical structures. To ensure a model's test set contains structurally distinct molecules, we grouped the molecules sharing the same scaffold into either the training or the test set. The scaffold split method is more challenging for model fitting than the random split approach, which forces the model to generalize to new chemical structures and gives a better indication of how the model will perform on compounds with entirely new core structures [14]. To visualize and compare the two splitting methods, we computed Tanimoto distances (TDs) to measure the structural dissimilarity between the compounds in the test set and the training set using Morgan fingerprints (2,048 bits) with a radius of two. We developed 20 QSAR models (binary classifiers) using random forest (RF) [15], XGBoost [16], support vector machine (SVM) [17], artificial neural network (ANN) [18], and logistic regression (LR) [19] algorithms for four types of molecular representations. We optimized the hyperparameters of the QSAR models using five-fold cross-validation and a grid search. We split the training set into five equally sized folds, trained the models on four folds, and validated the trained models with the remaining fold. We repeated this process five times for each model so that each fold served as the validation set once. We evaluated the average model performance to select the optimal hyperparameters. **Supplementary Table S1** lists the optimized hyperparameters and their search space. After selecting the optimal hyperparameters, we fitted each model using the entire training set. We also created three consensus models by averaging the individual QSAR model outputs (predicted probabilities), including a consensus model combining the descriptor models, a consensus model combining the fingerprint models, and a consensus model combining both the descriptor and fingerprint models. We built all of the models using scikit-learn (version 1.3.0).

2.5. Model evaluation

We evaluated the predictive performance of the generated models on both the training and test sets. For the training set, we evaluated each model with the optimal hyperparameters using the same five-fold cross-validation process as applied for hyperparameter optimization. We selected the models with the best performance during cross-validation for each molecular representation and fitted selected models using the entire training set. We further evaluated the trained models and consensus models on the separate test set. We calculated three performance metrics based on the confusion matrix consisting of true positive (TP), true negative (TN), false positive (FP), and false negative (FN): area under the receiver operating characteristic curve (AUROC), accuracy, and Matthews correlation coefficient (MCC). Accuracy and MCC are defined by Equations (1) and (2), respectively. We computed analytical confidence intervals (95% confidence level) of these metrics using the Python package confidenceinterval (version 1.0.5).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

2.6. Model explainability

Shapley values, a widely used approach from cooperative game theory, determine a way to fairly distribute a total payout among the players of a game based on their individual contributions [20]. SHAP is a unified framework designed to efficiently estimate Shapley values for ML models by treating the features as players and the model's prediction as the payout [21]. The Shapley value of a feature is the average of its

marginal contributions across all possible subsets of feature coalitions, weighted by the size of the coalitions. The Python implementation of SHAP (version 0.46.0) provides various functions to estimate Shapley values, which can be grouped into model-specific and model-agnostic approaches. Shap.Explainer, a function in SHAP, can automatically choose the best SHAP algorithm given the model and masker. We selected TreeExplainer, a model-specific method, for the tree-based QSAR models, such as RF and XGBoost, but applied PermutationExplainer, a model-agnostic approach, for SVM, ANN and LR. We performed both global and local SHAP analyses. The global analysis focused on understanding overall model behavior by calculating feature importance across the entire dataset, whereas the local analysis explained individual predictions by understanding how each feature contributed to the specific prediction. In addition to SHAP analyses, we also determined feature importance of the LR model trained with Morgan fingerprints by analyzing the magnitude of the model's coefficients, where higher absolute values indicate stronger predictors for the target property.

2.7. Model applicability domains

QSAR models are often used to predict the properties of uncharacterized compounds, so it is crucial to define their applicability domains (ADs), i.e., the chemical space where the models are expected to provide reliable predictions. To define the ADs for our QSAR models, we applied the sum of distance-weighted contributions (SDC), which considers the weighted distances between the query compound and all the molecules used to train the model [22]. SDC is defined in Equation (3) as follows:

$$\mathrm{SDC} = \sum_{i=1}^{n} e^{\frac{-3TD_i}{1-TD_i}} \quad (3)$$

where TD_i represents the Tanimoto distance between a query compound and the ith training molecule and n represents the total number of training molecules. We computed the TD of two molecules using their Morgan fingerprints (2,048 bits) with a radius of two. TDs range from zero to one, where zero indicates that two molecules share all Morgan fingerprints and one indicates that they share no fingerprints.

3. Results

3.1. Data processing and exploration

We retrieved the respiratory irritant data from the OCHEM website.

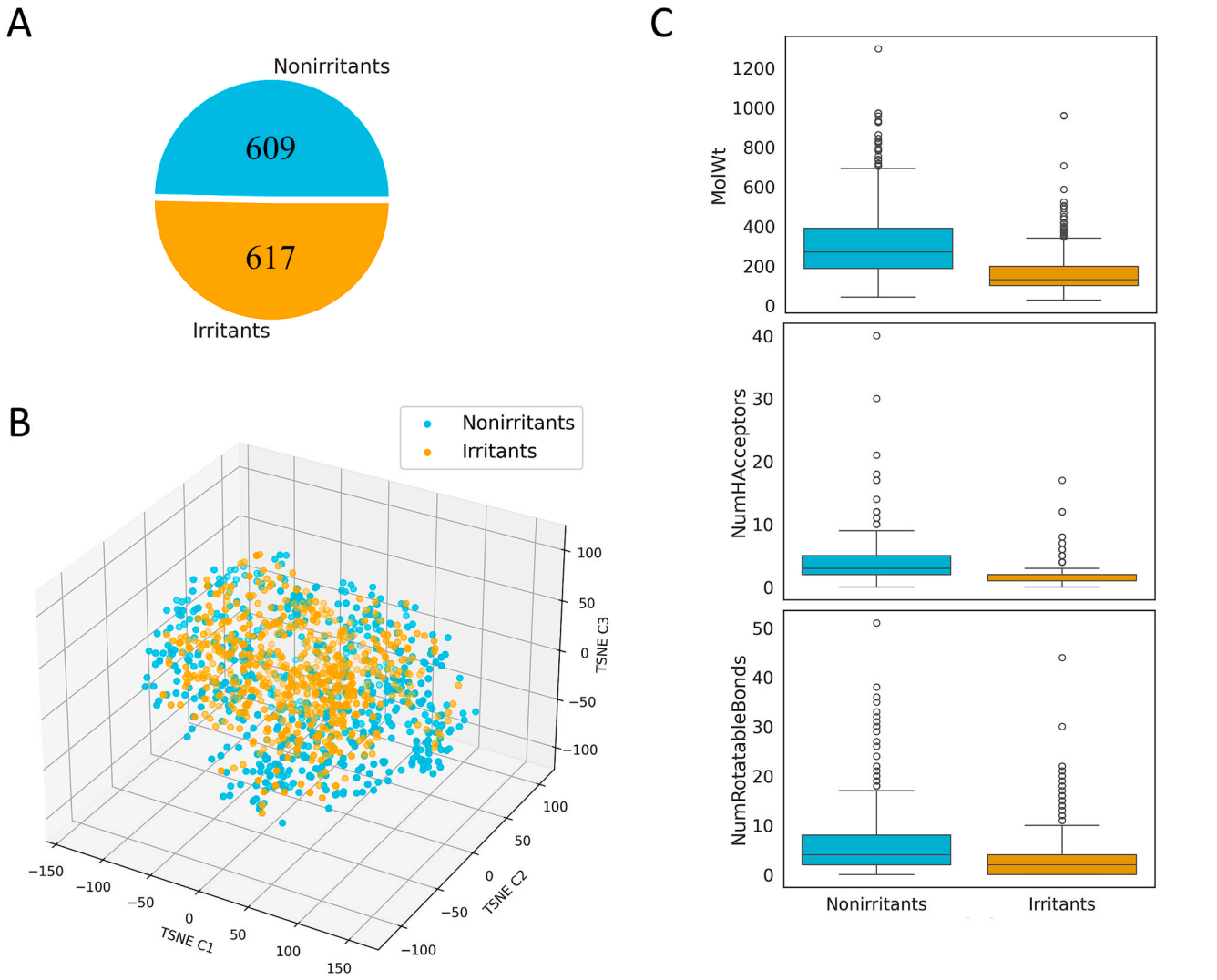


Fig. 2. Chemical space analysis. (A) Pie plot shows the number of respiratory irritants and nonirritants. (B) Three-dimensional t-distributed stochastic neighbor embedding (t-SNE) plot shows the distribution of respiratory irritants and nonirritants in the chemical space. (C) Comparison between respiratory irritants and nonirritants based on selected physicochemical properties, including molecular weight (MolWt), number of hydrogen acceptors (NumHAcceptors), and number of rotatable bonds (NumRotatableBonds).

After applying our molecular standardization pipeline to the downloaded data, we obtained a curated dataset of 1,226 chemicals, including 617 respiratory irritants and 609 nonirritants (Fig. 2**A**). We presented example respiratory irritants and nonirritants in **Supplementary Figs. S1 and S2**. We explored the chemical space of all 1,226 chemicals using Morgan fingerprints and PCA and found that the first 50 principal components accounted for 54% of the total variation in the whole dataset. To obtain a more accurate representation of the chemical space, we employed t-SNE to create a three-dimensional scatter plot to explore patterns and relationships between respiratory irritants and nonirritants. Fig. 2**B** shows that both respiratory irritants and nonirritants are widely distributed in the chemical space, with some overlap between them. In addition, we compared respiratory irritants and nonirritants based on individual molecular properties and observed statistically significant differences in molecular weight, number of acceptors, and number of rotatable bonds (Fig. 2**C**). The molecular weight distribution of the respiratory irritants was between 100 and 400, whereas the majority of the nonirritants had higher molecular weights, ranging from 100 to 700. The nonirritants also had a larger number of hydrogen acceptors and rotatable bonds compared to the respiratory irritants.

3.2. *Model building and evaluation*

We split the curated dataset into training (80%) and test (20%) sets and trained the QSAR models using the training set only, reserving the test set as external data to assess the models' generalizability. We adopted two different data-splitting strategies, random split and scaffold split, and compared the approaches. Because TDs measure structural dissimilarity between chemicals represented using Morgan fingerprints, we computed the TDs between each test set chemical and its nearest neighbor in the training set and plotted their distribution (**Supplementary Fig. S3**). In the test set generated using the scaffold split, the majority of chemicals had TDs between 0.5 and 0.8, indicating that they were structurally different from the chemicals in the training set. In contrast to the scaffold split, the random split generated test set chemicals more similar to the training set compounds. Thus, we chose

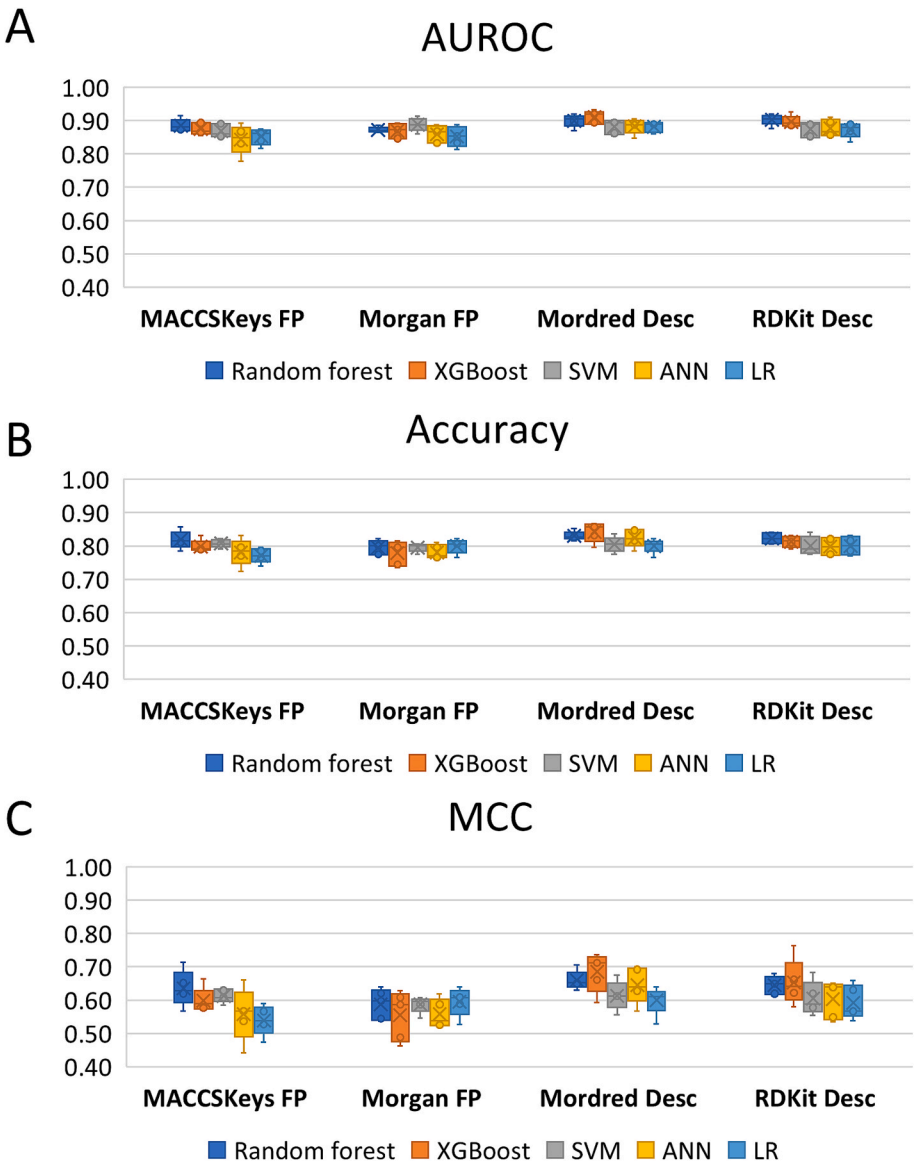


Fig. 3. Model performance assessed using five-fold cross-validation on the training data – scaffold split. Box plots show the models' performance based on the following metrics: (A) area under the receiver operating characteristic curve (AUROC), (B) accuracy, and (C) Matthews correlation coefficient (MCC). We generated the models using each of the four molecular representation methods [molecular access system keys (MACCSKeys), Morgan, Mordred, and RDKit] with each of the five machine learning algorithms [random forest, XGBoost, support vector machine (SVM), artificial neural network (ANN), and logistic regression (LR)].

the scaffold split to obtain a more realistic assessment of generalizability.

We created 20 QSAR models to classify respiratory irritants and nonirritants. For each of the four types of molecular representations, we employed five ML algorithms (RF, XGBoost, SVM, ANN, and LR) to train the models using the training set only. Then, we applied five-fold cross-validations to optimize hyperparameters and evaluate the average model performance based on AUROC, accuracy, and MCC. Fig. 3 shows the distributions of the three performance metrics for all 20 ML-molecular representation combinations. All of the QSAR models' predictions were consistently accurate for the implemented ML and molecular representation methods. On average, the models achieved an AUROC of 0.88, accuracy of 0.80, and MCC of 0.61. In addition, the QSAR models trained on molecular descriptors, i.e., Mordred and RDKit descriptors, generally outperformed the models trained on MACCSKeys and Morgan fingerprints. The tree-based algorithms, i.e., RF and XGBoost, performed better than the other ML algorithms in most cases, except for the model trained with Morgan fingerprints, where the LR model outperformed the others in terms of accuracy and MCC. We selected the best models for each molecular representation based on their average performance in the cross-validation process, including one RF model (MACCSKeys), one LR model (Morgan), and two XGBoost models (Mordred and RDKit). To further validate the models' generalizability, we evaluated their performance against the test set. In addition, we developed and evaluated three consensus models by averaging the individual QSAR model outputs, where we combined two descriptor models, two fingerprint models, and all four individual models. Table 2 summarizes the performance of the four individual models and three consensus models on the test set, including the 95% confidence intervals for all three performance metrics. In general, compared to their performance on the training set, all QSAR models exhibited comparable performance on the test set, indicating that the models had good generalizability and were not overfitted to the training set. The consensus model combining the two descriptor models or the two fingerprint models slightly outperformed the individual fingerprint models. The consensus model combining all four individual models yielded an AUROC of 0.91, accuracy of 0.87, and MCC of 0.70. In addition, we checked the distribution of SDC density for the prediction results of the Mordred model on the test set. **Supplementary Fig. S4** shows that correctly predicted chemicals exhibited a similar SDC distribution as incorrectly predicted chemicals, suggesting that model performance on the test set was not correlated with the similarity between the test chemicals and the chemicals in the training set. We evaluated model performance on the test chemicals with low SDC values (<0.05), which are potentially outside of the model's AD. Notably, the model still exhibited comparable performance (e.g., accuracy of 0.82) on these chemicals, indicating it had a broad AD and good generalizability.

3.3. Model explainability

To better understand the QSAR models' predictions and trace their decision-making process, we performed both global and local SHAP analyses on the best-performing models for each molecular representation. A global SHAP analysis identifies the significant molecular descriptors and structural fingerprints that influence the QSAR models to make accurate predictions of chemical-induced respiratory irritation. Fig. 4 shows SHAP summary plots for the two QSAR models trained on the Mordred and RDKit descriptors, providing a global overview of the SHAP values for the 10 most important molecular descriptors of each model and their relationships with the predicted chemical classes. The vertical axis lists the molecular descriptors ranked by importance, based on their mean absolute SHAP values, with the most impactful at the top. The horizontal axis shows the SHAP values, representing the molecular descriptors' impact on the models' predictions. Each dot corresponds to the SHAP value of one chemical in the test set. A molecular descriptor's influence on predictions is more variable when its SHAP values are widely spread. Consistent SHAP values for a molecular descriptor lead to a consistent effect on predictions, whereas scattered values result in a context-dependent impact. The color of the dot represents the descriptor value, with red signifying a high descriptor value and blue a low descriptor value, indicating how descriptor values correlate with impact direction. Table 3 lists the 10 most important molecular descriptors and their corresponding classes for the Mordred and the RDKit models, respectively.

Fig. 4**A** shows that the majority of the most important molecular descriptors in the Mordred model belong to the autocorrelation and burden chemical abstract service University of Texas (BCUT) classes. The autocorrelation descriptors, such as ATS6s, have a negative impact on the respiratory irritant predictions, whereas the BCUT descriptors are positively correlated with chemical-induced respiratory irritation. Autocorrelation descriptors encode both structural and physicochemical properties of a molecule by considering pairs of atoms and their topological distance within a molecule, known as lag [23]. In the Mordred application, autocorrelation descriptors are calculated using Moreau-Broto (ATS), Moran (MATS), and Geary (GATS) algorithms, each of which can be applied with various weighting schemes (atomic properties) and different lags to generate a diverse set of descriptors. For example, ATS6s is a Moreau-Broto autocorrelation descriptor of lag 6, which is weighted by the intrinsic state. In contrast to autocorrelation descriptors, BCUT descriptors characterize the electronic and topological features of a molecule [24]. They are calculated based on the eigenvalues of a modified adjacency matrix that incorporates atomic properties, such as Gasteiger charge, sigma electrons, and Allred-Rochow electronegativity. For example, BCUTc-1l is a BCUT descriptor calculated using the first lowest eigenvalue of a Burden matrix weighted by Gasteiger charge. **Supplementary Fig. S5** presents two example test compounds with high BCUTc-1l and low ATS6s values that were correctly predicted as respiratory irritants. Fig. 4**B** shows that the majority of the most important molecular descriptors in the RDKit model belong to the topological and connectivity classes. Topological descriptors represent a broader category of two-dimensional descriptors that characterize the molecule's overall shape, size, and branching based on its graph representation, whereas connectivity descriptors are a well-defined and classic subset of topological descriptors that quantify the degree of branching and cyclicity in molecules. In addition, we found that the descriptors in the BCUT class appeared in the top 10 descriptor lists for both models (Fig. 4).

Furthermore, we determined the significant structural fingerprints by performing global SHAP analyses for the QSAR models trained on

Table 2
Model performance on the test data – scaffold split.

Performance Metrics	MACCSKeys FP-RF	Morgan FP-LR	Mordred Desc-XGBoost	RDKit Desc-XGBoost	Consensus FP	Consensus Desc	Consensus FP + Desc
AUROC(95% CI)	0.88(0.84,0.93)	0.83(0.78,0.89)	0.92(0.89,0.96)	0.91(0.88,0.95)	0.88(0.84,0.93)	0.92(0.89,0.96)	0.91(0.88,0.95)
Accuracy(95% CI)	0.80(0.75,0.85)	0.78(0.73,0.83)	0.84(0.79,0.88)	0.85(0.80,0.89)	0.81(0.76,0.86)	0.86(0.81,0.90)	0.87(0.82,0.91)
MCC(95% CI)	0.54(0.42,0.65)	0.53(0.42,0.64)	0.63(0.52,0.73)	0.64(0.53,0.73)	0.59(0.48,0.69)	0.67(0.56,0.76)	0.70(0.59,0.79)

AUROC: area under the receiver operating characteristic curve; CI: confidence interval; Desc: descriptor; FP: fingerprint; LR: logistic regression; MACCSKeys: molecular access system keys; MCC: Matthews correlation coefficient; RF: random forest; SVM: support vector machine.

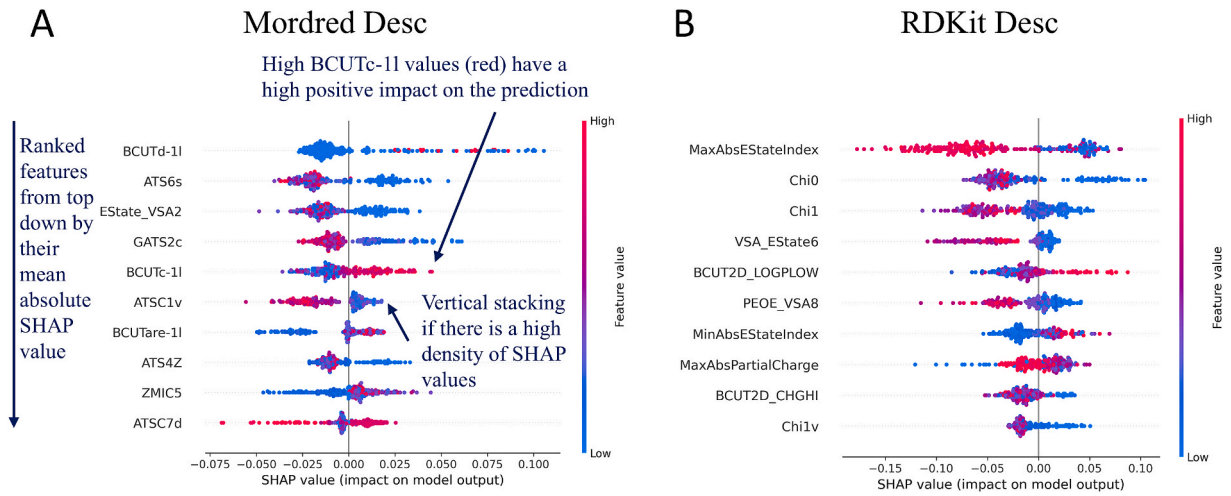


Fig. 4. Shapley additive explanations (SHAP) global analysis – molecular descriptors. Summary plots show the global impact of molecular descriptors on a machine learning model's prediction. We ranked the (A) Mordred descriptors and (B) RDKit descriptors by their overall importance, with the most impactful features at the top.

Table 3
Top 10 Mordred and RDKit descriptors ranked by their mean absolute SHAP values.

Mordred Descriptor	Class	RDKit Descriptor	Class
BCUTd-1l	BCUT	MaxAbsEStateIndex	Topological
ATS6s	Autocorrelation	Chi0	Connectivity
EState_VSA2	MoeType	Chi1	Connectivity
GATS2c	Autocorrelation	VSA_Estate6	MoeType
BCUTc-1l	BCUT	BCUT2D_LOGPLOW	BCUT2D
ATSC1v	Autocorrelation	PEOE_VSA8	Topological
BCUTare-1l	BCUT	MinAbsEStateIndex	Topological
ATS4Z	Autocorrelation	MaxAbsPartialCharge	Topological
ZMIC5	InformationContent	BCUT2D_CHGHI	BCUT2D
ATSC7d	Autocorrelation	Chi1v	Connectivity

Descriptors ranked in descending order. 2D: two-dimensional; BCUT: Burden chemical abstract service University of Texas; SHAP: Shapley additive explanations.

Morgan and MACCSKeys fingerprints. Fig. 5 shows that the most important structural fingerprints had a negative impact on the respiratory irritant predictions, indicating the absence of these substructures in respiratory irritants. Tables 4 and 5 list the 10 most important structural fingerprints for the Morgan and MACCSKeys models, respectively. In addition, we determined feature importance of the model trained with Morgan fingerprints by analyzing the magnitude of the LR model's coefficients (**Supplementary Fig. S6**). A large positive coefficient means a higher likelihood of the positive class (respiratory irritants); a negative coefficient indicates a lower likelihood. In Table 4, we compared the most important Morgan fingerprints identified from the global SHAP analysis on the test data with those determined by the LR model's coefficients on the training data. There was only one overlapping feature, and the most important Morgan fingerprints determined by the LR model's coefficients have much lower counts on the test data, indicating they are biased toward the training data.

To identify which molecular descriptors and structural fingerprints contribute the most to the individual predictions of the chemical classes, we also performed local SHAP analyses. Fig. 6 shows SHAP waterfall plots that illustrate the impact of each Mordred descriptor and Morgan fingerprint on the model output to correctly predict two example test chemicals (Compound CIDs: 7570 and 161693) as respiratory irritants.

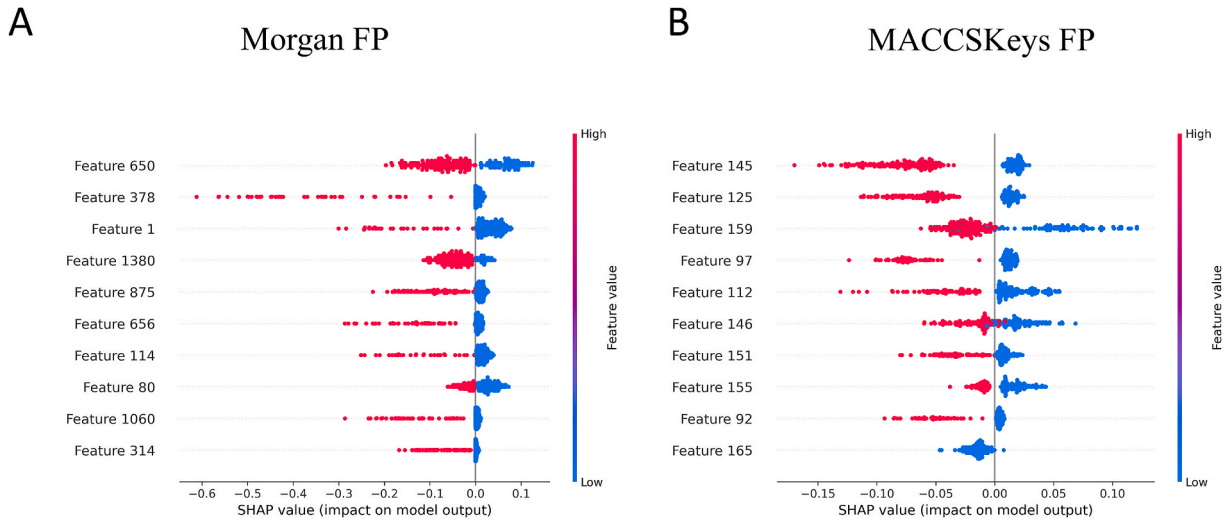
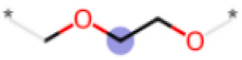
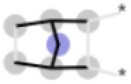
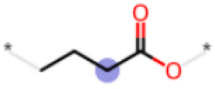
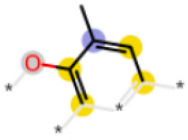
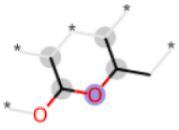
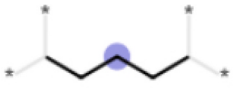
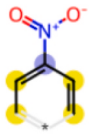
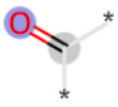
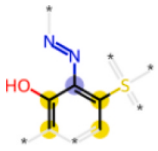


Fig. 5. Shapley additive explanations (SHAP) global analysis – structural fingerprints. Summary plots show the global impact of structural fingerprints on a machine learning model's prediction. We ranked the (A) Morgan fingerprints and (B) molecular access system keys (MACCSKeys) fingerprints by their overall importance, with the most impactful features at the top.

Table 4
Top 10 Morgan fingerprints ranked by their mean absolute SHAP values and their absolute logistic regression coefficients, and their counts on the test data.

Morgan Fingerprint Ranked by SHAP	Structure	Count	Morgan Fingerprint Ranked by Coefficients	Structure	Count
Feature 650		159	Feature 1649		1
Feature 378		41	Feature 755		4
Feature 1		28	**Feature 378**		41
Feature 1380		206	Feature 1399		0
Feature 875		94	Feature 929		1
Feature 656		47	Feature 59		0
Feature 114		29	Feature 1977		3
Feature 80		84	Feature 146		1
Feature 1060		43	Feature 769		4
Feature 314		68	Feature 1549		3

Morgan fingerprints ranked in descending order. SHAP: Shapley additive explanations.

The predicted probability of each example test chemical to be a respiratory irritant is shown at the top right of each plot, such as $f(x) = 0.994$. The vertical axis lists the molecular features (descriptors/fingerprints) ranked by their mean absolute SHAP values, with the feature's value for the specific example test chemical denoted in gray. The SHAP value corresponding to each feature is marked with an arrow, which shows how each feature's contribution moves the model's prediction from a baseline to the final output. The impact of each feature is proportionate to its arrow length. We used the average of all of the features' SHAP values as the base value/probability, denoted as $E[f(X)]$. Local SHAP analyses explain how molecular features work together to push the predicted probability from the base probability for each test chemical. Fig. 6 shows that the top Mordred descriptor was identical for the two example test chemicals. GATS2c is an autocorrelation descriptor calculated using the GATS algorithm, and its low values positively contributed to the predicted probability of the chemicals to be a respiratory irritant. In addition, we found that the two example test chemicals shared several important Morgan fingerprints, such as feature/

Table 5
Top 10 MACCSKeys fingerprints ranked by their mean absolute SHAP values.

MACCSKeys Fingerprint	Description	SMART Patterns
Feature 145	6 M ring > 1	('*1~*~*~*~*~*~1', 1)
Feature 125	Aromatic ring > 1	('?', 0)
Feature 159	O > 1	('[#8]', 1)
Feature 97	NAAAO	('[#7]~*~*~*~[#8]', 0)
Feature 112	AA(A)(A)A	('*~*(~*)(~*)~*', 0)
Feature 146	O > 2	('[#8]', 2)
Feature 151	NH	('[#7;!H0]', 0)
Feature 155	A!CH2!A	('*!@[CH2]!@*', 0)
Feature 92	OC(N)C	('[#8]~[#6](~[#7])~[#6]', 0)
Feature 165	Ring	('[R]', 0)

MACCSKeys fingerprints ranked in descending order. MACCSKeys: molecular access system keys; SHAP: Shapley additive explanations; SMART: simplified molecular input line entry system (SMILES) arbitrary target specification.

fingerprint 507 and 1004, and each feature/fingerprint corresponded to a substructure shown in Fig. 6. The predicted respiratory irritants with similar important structural features indicate that the QSAR model focused on the same substructures to make their predictions. Similarly, we performed local SHAP analyses to explain how the QSAR models predicted respiratory nonirritants. Fig. 7 shows waterfall plots that indicate the local impact of features on two example test compounds that were correctly predicted as respiratory nonirritants (Compound CIDs: 158619 and 8437).

4. Discussion and conclusion

Chemical-induced respiratory irritation is an important human health endpoint in chemical risk assessment. Currently, assessing whether a chemical will cause respiratory irritation in humans includes animal testing, making it infeasible to experimentally evaluate respiratory toxicity for all the chemicals in large two-dimensional chemical libraries. The U.S. Department of War as well as regulatory authorities, such as the U.S. Environmental Protection Agency, the U.S. Food and Drug Administration, and the European Chemicals Agency, recommend NAMs as the preferred path forward to improve the efficiency of chemical risk assessment and reduce the reliance on animal testing [4,5]. In the present work, we created explainable QSAR models as an alternative screening tool for traditional animal models to obtain faster and more cost-effective human risk assessment of respiratory irritants.

To fill the knowledge gap of model explainability, we performed global SHAP analyses to prioritize the significant key molecular descriptors that contributed the most to the models' predictions and identify critical determinants of respiratory irritation (Fig. 4). For instance, we identified the autocorrelation and BCUT descriptors as the most important molecular descriptors from the best-performing Mordred model. Autocorrelation descriptors have been reported to be

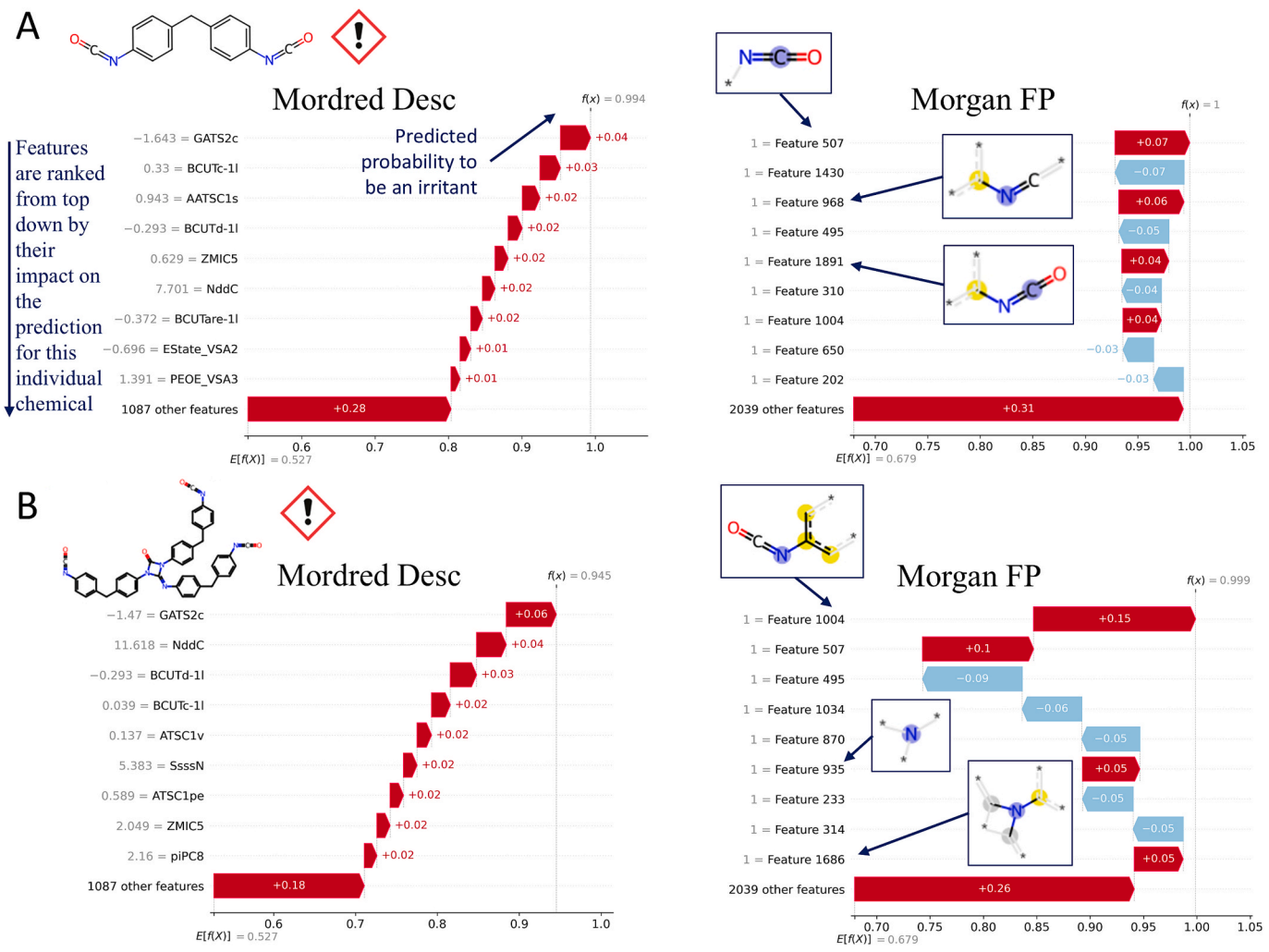


Fig. 6. Shapley additive explanations (SHAP) local analysis – respiratory irritants. Waterfall plots show the local impact of features on two example test compounds that were correctly predicted as respiratory irritants: (A) Compound CID: 7570 and (B) Compound CID: 161693. We ranked the Mordred descriptors and Morgan fingerprints by their local importance, with the most impactful features at the top.

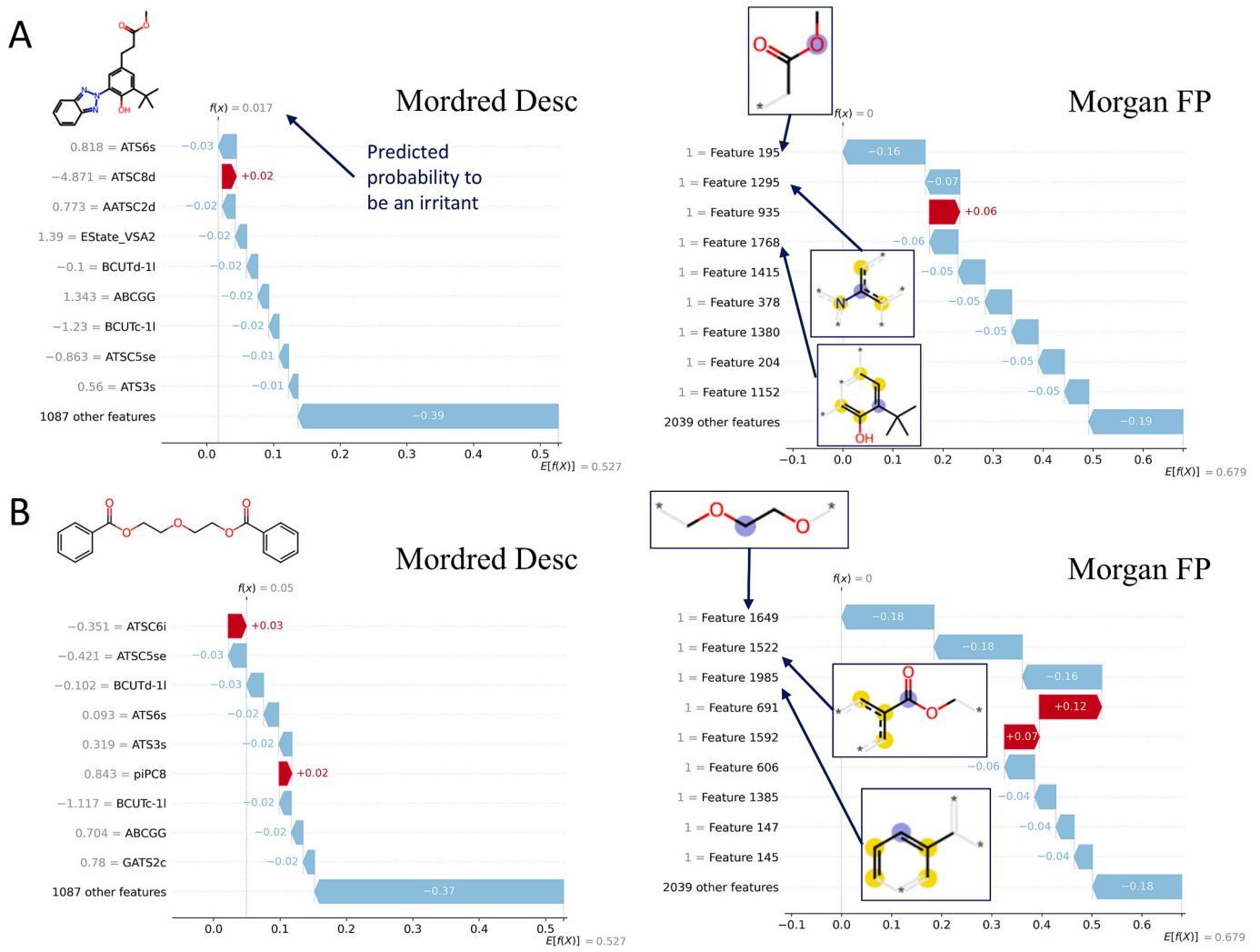


Fig. 7. Shapley additive explanations (SHAP) local analysis – respiratory nonirritants. Waterfall plots show the local impact of features on two example test compounds that were correctly predicted as respiratory nonirritants: (A) Compound CID: 158619 and (B) Compound CID: 8437. We ranked the Mordred descriptors and Morgan fingerprints by their local importance, with the most impactful features at the top.

predictive of respiratory toxicity [25] and chemical-induced mitochondrial toxicity [26], indicating that our findings agree with current knowledge. Furthermore, previous studies discovered BCUT as important molecular descriptors for ecological risk assessment of chemical compounds [27], but to our knowledge, this work is the first study reporting BCUT as critical molecular descriptors for predicting respiratory irritation. However, we note that some molecular descriptors have limitations. For example, some Mordred descriptors are purely mathematical (e.g., specific topological indices) and lack direct, intuitive physicochemical interpretation.

Thus, we also identified important structural features predictive of respiratory irritation, such as Morgan fingerprints 507 (isocyanate) and 1004, by performing local SHAP analyses. Isocyanates and isocyanate-containing structures are known to cause respiratory irritation in occupational settings (e.g., toluene diisocyanate). As illustrated in Fig. 6, we were able to correctly classify known respiratory irritants and nonirritants in the test set using these fingerprints, which could also be applied to derive structural alerts for flagging chemicals for potential respiratory toxicity. Furthermore, the global SHAP analyses showed that the most predictive structural fingerprints were present in respiratory nonirritants and absent in respiratory irritants, indicating that the classification models' performance heavily relied on the structural patterns recognized from respiratory nonirritants (Fig. 5). Thus, the global SHAP analyses of the models trained on structural fingerprints provided limited information on important structural patterns of respiratory irritants. In contrast, the global SHAP analyses of the models trained on molecular descriptors were able to quantitatively capture molecular attributes unique to respiratory irritants. One of the goals of the present study was to compare different molecular representations. We note that structural fingerprints are better at identifying functional groups or substructures and calculating similarity between molecules, whereas molecular descriptors are better at representing overall physicochemical properties (e.g., molecular weight). The choice of molecular representation depends on the specific intended tasks. In addition to SHAP analyses, we also determined the feature importance of Morgan fingerprints using LR model's coefficients. These coefficients are simple, highly interpretable, and fast to compute, but they are limited to linear relationships. In contrast, SHAP analyses can capture nonlinear relationships and feature interactions and provide consistent importance.

SHAP is a model-agnostic XAI method that can be applied to any ML model, which is a significant advantage as it allows for consistent explainability across diverse models without requiring model-specific implementation. However, the model-agnostic nature of SHAP also presents a computational challenge. Calculating the exact Shapley value for a feature in a model requires considering all the possible subsets of features to determine the marginal contribution of that feature. It requires evaluating 2^n possible subsets for a model with n features, leading to an exponential time complexity, and becomes computationally

prohibitive for respiratory irritation models with a large number of features. To address this computational challenge, we employed approximation methods, such as TreeExplainer, to estimate the Shapley values. TreeExplainer was designed for tree-based models, leveraging their structures to calculate Shapley values faster. While approximation methods could mitigate the computational burden, they may introduce a trade-off in terms of accuracy. For example, a recent study compared explanations of molecular ML models generated with different methods for calculating Shapley values [28] and found that methodological variants led to distinct feature importance distributions for highly accurate predictions, with significant agreement only observed between model-agnostic explanation methods and TreeExplainer. Thus, a next step in our modeling workflow refinement would be the integration of different XAI approaches, such as instance-based methods, to evaluate the consistency of model explanations. We also plan to further explore how to translate the quantitative descriptor explanations into human-understandable interpretations using natural language processing models, which can enable human reasoning and facilitate better communication between stake-holders [29].

In conclusion, we developed explainable respiratory irritation models that can be integrated into a decision support system to rapidly screen chemicals and prioritize potential respiratory irritants for more detailed experimental evaluation. The explanations generated by the SHAP analyses increase the transparency and utility of our models, making them a useful tool for compound design as well.

5. Disclaimer

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Defense Health Agency, the U.S. Department of War, the U. S. Government, or The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. Distribution Statement A. Approved for public release: distribution is unlimited.

Ethics statement

The modeling results presented herein were independently assessed for reproducibility.

CRediT authorship contribution statement

Pinyi Lu: Writing – review & editing, Writing – original draft, Formal analysis, Data curation, Conceptualization. **Souvik Dey:** Writing – review & editing, Conceptualization. **Anders Wallqvist:** Writing – review & editing, Funding acquisition, Conceptualization. **Mohamed Diwan M. AbdulHameed:** Writing – review & editing, Conceptualization.

Funding

This research was funded by the U.S. Army Medical Research and Development Command under Contract Nos. W81XWH20C0031 and HT942524F0189 and by Defense Threat Reduction Agency Grant CBCall14-CBS-05-2-0007.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.comtox.2026.100410.

Data availability

The data used for model training and testing as well as the scripts used in the present study are provided on GitHub (https://github.com/BHSAI/explainableQSAR/tree/main/Respiratory_irritation).

References

[1] M.M. Wehr, S.S. Sarang, M. Rooseboom, P.J. Boogaard, A. Karwath, S.E. Escher, RespiraTox - development of a QSAR model to predict human respiratory irritants, Regul. Toxicol. Pharmacol. 128 (2022) 105089, https://doi.org/10.1016/j.yrtph.2021.105089.
[2] United Nations. Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Available at https://unece.org/sites/default/files/2023-07/GHS%20Rev10e.pdf, accessed on November 20, 2025.
[3] United States Army Chemical School. Potential Military Chemical/Biological Agents and Compounds. Available at https://irp.fas.org/doddir/army/fm3-11-9.pdf, accessed on November 20, 2025.
[4] J. Strickland, E. Haugabrooks, D.G. Allen, L.B. Balottin, Y. Hirabayashi, N. C. Kleinstreuer, et al., International regulatory uses of acute systemic toxicity data and integration of new approach methodologies, Crit. Rev. Toxicol. 53 (2023) 385–411, https://doi.org/10.1080/10408444.2023.2240852.
[5] J.L. Fisher, K. Yamada, A.J. Keebaugh, K.T. Williams, C.L. German, A.M. Hott, et al., Evaluating applicability domain of acute toxicity QSAR models for military and industrial chemical risk assessment, Toxicol. Lett. 403 (2025) 1–8, https://doi.org/10.1016/j.toxlet.2024.11.006.
[6] Y.N. Fuadah, M.A. Pramudito, L. Firdaus, F.J. Vanheusden, K.M. Lim, QSAR classification modeling using machine learning with a consensus-based approach for multivariate chemical hazard end points, ACS Omega 9 (2024) 50796–50808, https://doi.org/10.1021/acsomega.4c09356.
[7] Y. Chushak, A. Keebaugh, R.A. Clewell, Prediction of respiratory irritation and respiratory sensitization of chemicals using structural alerts and machine learning modeling, Toxics 13 (2025) 243, https://doi.org/10.3390/toxics13040243.
[8] J.L. Fisher, K.T. Williams, L.J. Schneider, A.J. Keebaugh, C.L. German, A.M. Hott, et al., Evaluation of QSAR models for tissue-specific predictive toxicology and risk assessment of military-relevant chemical exposures: a systematic review, Computat Toxicol. 32 (2024) 100329, https://doi.org/10.1016/j.comtox.2024.100329.
[9] J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence, Nat. Mach. Intell. 2 (2020) 573–584, https://doi.org/10.1038/s42256-020-00236-4.
[10] H. Moriwaki, Y.S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, J Cheminform. 10 (2018) 4, https://doi.org/10.1186/s13321-018-0258-y.
[11] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754, https://doi.org/10.1021/ci100050t.
[12] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL keys for use in drug discovery, J. Chem. Inf. Comput. Sci. 42 (2002) 1273–1280, https://doi.org/10.1021/ci010132r.
[13] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, JMLR 9 (2008) 2579–2605.
[14] G.W. Bemis, M.A. Murcko, The properties of known drugs. 1. Molecular frameworks, J. Med. Chem. 39 (1996) 2887–2893, https://doi.org/10.1021/jm9602928.
[15] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[16] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco, California, USA, 2016, p. 785–794.
[17] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297, https://doi.org/10.1007/BF00994018.
[18] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133, https://doi.org/10.1007/BF02478259.
[19] D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, Applied logistic regression, Wiley, New York, USA (2000), https://doi.org/10.1002/0471722146.
[20] L.S. Shapley, A value for n-person games, in: Contributions to the Theory of Games, Princeton University Press, Princeton, NJ, USA, 1953, pp. 307–317.
[21] S.M. Lundberg, S.-I. Lee, in: A Unified Approach to Interpreting Model Predictions, Curran Associates Inc., Long Beach, CA, USA, 2017, pp. 4768–4777.
[22] R. Liu, A. Wallqvist, Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds, J. Chem. Inf. Model. 59 (2019) 181–189, https://doi.org/10.1021/acs.jcim.8b00597.
[23] B. Hollas, An analysis of the autocorrelation descriptor for molecules, J. Math. Chem. 33 (2003) 91–101, https://doi.org/10.1023/A:1023247831238.
[24] R.S. Pearlman, K.M. Smith, Metric validation and the receptor-relevant subspace concept, J. Chem. Inf. Comput. Sci. 39 (1999) 28–35, https://doi.org/10.1021/ci980137x.
[25] K. Jaganathan, H. Tayara, K.T. Chong, An explainable supervised machine learning model for predicting respiratory toxicity of chemicals using optimal molecular descriptors, Pharmaceutics 14 (2022) 832, https://doi.org/10.3390/pharmaceutics14040832.
[26] K. Jaganathan, M.U. Rehman, H. Tayara, K.T. Chong, XML-CIMT: explainable machine learning (XML) model for predicting chemical-induced mitochondrial

toxicity, Int. J. Mol. Sci. 23 (2022) 15655, https://doi.org/10.3390/ijms232415655.
[27] J. Wei, L. Tian, F. Nie, Z. Shao, Z. Wang, Y. Xu, et al., Quantitative structure-activity relationship model development for estimating the predicted no-effect concentration of petroleum hydrocarbon and derivatives in the ecological risk assessment, Heliyon 10 (2024) e26808, https://doi.org/10.1016/j.heliyon.2024.e26808.
[28] A. Lamens, J. Bajorath, Comparing explanations of molecular machine learning models generated with different methods for the calculation of Shapley values, Mol Inform. 44 (2025) e202500067, https://doi.org/10.1002/minf.202500067.
[29] H.A. Gandhi, A.D. White, Explaining molecular properties with natural language, ChemRxiv (2022). https://doi.org/10.26434/chemrxiv-2022-v5p6m-v3.