

Toward Explainable Carcinogenicity Prediction: An Integrated Cheminformatics Approach and Consensus Framework for Possibly Carcinogenic Chemicals

Huynh Anh Duy and Tarapong Srisongkram*



Cite This: *J. Chem. Inf. Model.* 2025, 65, 10194–10220



Read Online

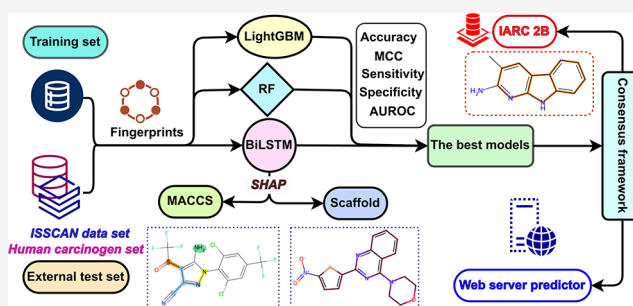
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: A carcinogenicity assessment of possibly carcinogenic chemicals (International Agency for Research on Cancer: IARC class 2B) was conducted using a consensus framework constructed from three complementary machine learning models: BiLSTM with MACCS fingerprints, LightGBM with RDKit descriptors, and Random Forest (RF) with E-state features. These models were developed and rigorously evaluated on benchmark carcinogenicity data sets, with LightGBM emerging as the top performer (accuracy = 0.800, MCC = 0.615, AUROC = 0.882, sensitivity = 0.739, specificity = 0.857). Consistent and competitive performance was also observed for RF and BiLSTM, affirming the reliability of individual predictions. Notably, LightGBM maintained strong generalization ability on independent human carcinogen test sets from IARC and IRIS (accuracy = 0.753, MCC = 0.535, AUROC = 0.842). For the ISSCAN internal test set, the top three models achieved MCC values ranging from 0.564 to 0.615, with AUROC scores between 0.858 and 0.882. For the human carcinogen test set, the top three models attained MCC values from 0.335 to 0.535 and AUROC scores ranging from 0.785 to 0.842. The consensus model was subsequently applied to 47 within-domain compounds from the 2B category, classifying them into 16 potential carcinogens, 8 presumed noncarcinogens, and 23 cases with inconclusive results. To uncover structural correlates, a SHAP-based interpretation of the BiLSTM model was performed, revealing discriminative molecular features including MACCS fingerprint keys and core Bemis–Murcko scaffolds associated with predicted carcinogenicity. To support practical applications, a freely accessible web server for carcinogenicity assessment has been developed and is available at <https://carcinogenicity-predictor.streamlit.app>.



1. INTRODUCTION

Cancer has escalated into a significant public health challenge, currently recognized as the second major cause of death and disability worldwide. In 2024, there were 20 million new cases and more than 9.7 million deaths due to cancer, with forecasts suggesting an ongoing rise in cases worldwide.¹ The increase in cancer incidence and mortality reflects a variety of factors linked to socioeconomic conditions, lifestyle habits, hygiene practices, and exposure to both occupational settings and everyday life.² Carcinogens are one of the factors that induce tumors by indirectly interacting with DNA or disrupting cellular metabolic activities. In the pharmaceutical industry, many synthetic chemicals are manufactured to meet demand. During this process, the chemical properties of the molecules can be altered, leading to the formation of carcinogens. As a result, carcinogenicity assessment of these residues or novel compounds is critical. Regulatory agencies such as the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), Organisation for Economic Cooperation and Development (OECD), and European Chemicals Agency (ECHA) also mandate the

assessment of mutagenicity and carcinogenicity for manufactured products or byproducts. However, traditional carcinogenicity assessments require long-term animal testing, typically spanning at least two years, which is both time-consuming and costly. Consequently, there is a need for alternative methods that are faster, more cost-effective, and involve fewer animals in carcinogenicity assessment protocols.

Quantitative structure–activity/toxicity relationship (QSAR/QSTR) is a gold standard new alternative method (NAM) used to evaluate data gaps in the chemical risk assessment.^{3,4} This method relies on a statistically validated relationship between chemical structure attributes and their activities.⁴ The chemical attributes can be varied from a chemical representation, such as a Simplified Molecular-Input Line Entry System (SMILES), to

Received: August 6, 2025
Revised: September 1, 2025
Accepted: September 3, 2025
Published: September 12, 2025



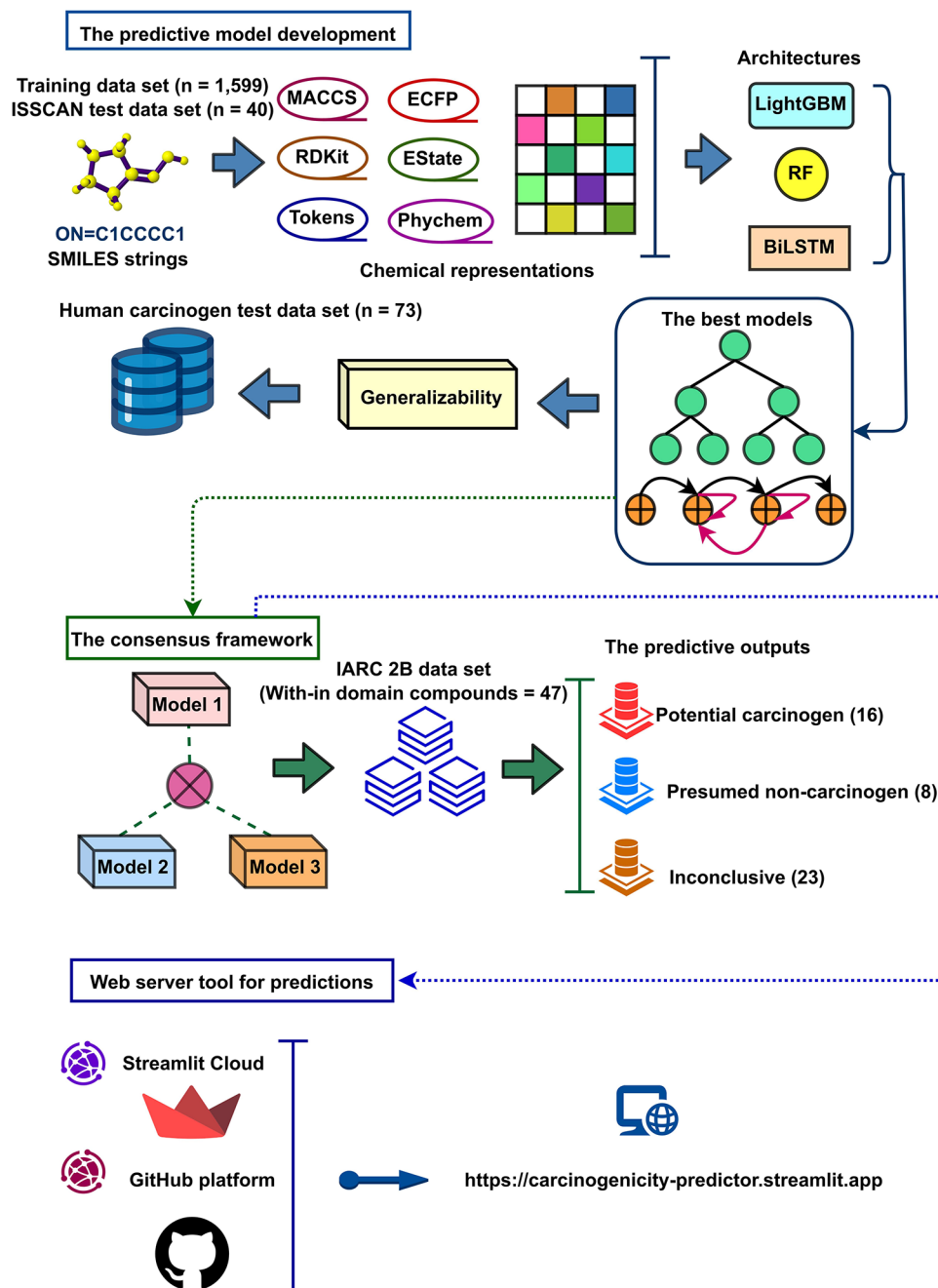


Figure 1. Overall workflow of carcinogenicity prediction based on the consensus framework. (A) The predictive model development, (B) the consensus framework for predicting possibly carcinogenic chemicals from IARC class 2B, and (C) deployment of carcinogenicity predictor using a web server tool.

molecular descriptors or molecular fingerprints.⁵ The mathematical algorithm of the QSAR method can also be varied from a simple multiple linear regression⁶ to sophisticated machine learning (ML)^{7,8} and deep learning (DL) techniques.^{9,10} Previous studies have constructed deep learning models (i.e., capsule networks and attention mechanisms) for carcinogenic assessments, which have received excellent performance on external validation sets.^{11–13} This prompted us to further experiment with the potential of other ML and DL models to improve the carcinogenicity assessments.

Additionally, the “possibly carcinogenic to humans” substances classified by the International Agency for Research on Cancer (IARC) under Group 2B are particularly concerning.

These substances are not conclusively proven as carcinogens in humans due to limited evidence from human studies and insufficient evidence in animal experiments.^{14,15} The carcinogenicity assessment focusing on IARC class 2B is critical as it represents a gap area where the potential carcinogens may exist and are not yet fully understood. Examples of IARC class 2B substances are aspartame,¹⁶ griseofulvin,¹⁷ or digoxin,¹⁸ which are being used in daily life. Therefore, assessing their safety potential can reduce risks associated with exposure to these agents in both medical and industrial environments.

Predictive computational models for chemical toxicity are increasingly recognized as essential for regulatory decision-making and reducing animal testing. Their adoption, however,

requires transparent, comprehensive, and methodologically rigorous documentation to ensure both scientific credibility and practical applicability.¹⁹ To this end, the OECD has established five fundamental principles that predictive models must meet to achieve scientific validity and regulatory acceptance.²⁰ Specifically, (i) the model should have a clearly defined end point directly relevant to human health or environmental safety; (ii) there must be a scientifically sound rationale underlying the relationship between chemical structure and the predicted activity; (iii) the applicability domain should be explicitly characterized, thereby delineating the chemical space within which reliable predictions can be made; (iv) predictive performance must be demonstrated through appropriate statistical evaluations; and (v) the model should be transparent and interpretable, ensuring that predictions can be critically assessed.

In this study, a bidirectional long short-term memory (BiLSTM), a light gradient-boosting machine (LightGBM), and a Random Forest (RF) were experimented with the carcinogenicity data for constructing effective QSAR models. These three models were selected based on the distinct learning mechanisms, which effectively capture structure–toxicity relationships, as demonstrated in our previous studies.^{21–23} The carcinogenic features tested include molecular fingerprints, physicochemical properties, and SMILES notation. Only the molecular fingerprints and physicochemical properties are human-crafted features, whereas SMILES notation is a chemical notation language, an end-to-end feature that requires no human intervention for feature selection and is suitable for both individual, mixtures, and salt forms of chemicals. Following the development of multiple QSAR models, the most robust performers were integrated into a novel consensus framework designed to reliably identify carcinogenic compounds. This framework was subsequently applied to external validation data, with a particular focus on assessing the carcinogenic potential of compounds classified as possibly carcinogenic to humans (IARC class 2B). To gain structural correlates into the molecular basis of these predictions, an in-depth cheminformatics analysis was performed, focusing on the identification of key structural scaffolds and substructural fingerprints that may drive carcinogenic potential. An overview of the integrated carcinogenicity prediction pipeline is illustrated in Figure 1. The key contributions of this study are summarized as follows:

1. We systematically developed a total of 18 predictive models using a well-curated carcinogenicity data set comprising 1599 compounds, including 872 noncarcinogens and 727 carcinogens for the training set and rigorously evaluated their generalization capabilities. The best-performing models demonstrated strong predictive performance across multiple evaluation metrics, highlighting their applicability for subsequent application in carcinogenicity assessment.
2. Among all features, MACCS key 19 exhibited the greatest influence, with a mean SHAP value of 0.0096, contributing 2.65% to the overall prediction. Among the features with a positive directional impact, MACCS125 demonstrated the highest importance, with a mean SHAP value of 0.0085, accounting for 2.26% of the contribution.
3. A novel consensus framework was established by integrating three complementary base learners, specially BiLSTM_MACCS, LightGBM_RDKit, and RF_E-state, to predict potentially carcinogenic compounds from the

IARC class 2B data set. This integrative strategy enabled the stratification of chemicals into relevant risk-based subsets, offering a valuable foundation for further toxicological assessments.

4. We conducted comprehensive cheminformatics profiling to elucidate key structural determinants associated with carcinogenicity. By leveraging BiLSTM-based SHAP analysis, we identified key molecular substructures and core scaffolds derived from MACCS fingerprints and Bemis–Murcko frameworks, which substantially contribute to the predictive outcomes of the models.
5. Finally, we successfully deployed a web-based tool for carcinogenicity prediction, which accepts SMILES strings as input and utilizes the consensus framework to generate high-confidence outputs. This tool serves as a practical resource to support *in silico* toxicity screening and early stage compound prioritization in chemical safety evaluation.

2. EXPERIMENTAL METHODS

2.1. Carcinogenic Data Preparation. The curated carcinogenic data set was primarily acquired from the previous study, before data preprocessing. The detailed descriptions are as follows:

1. The Carcinogenic Potency Database (CPDB) is a standardized resource that contains the results of chronic, long-term animal cancer tests conducted since the 1950s. These compounds are primarily based on experiments in mice, rats, and hamsters.²⁴
2. The Chemical Carcinogenesis Research Information System (CCRIS) database is a chemical record with carcinogenicity, mutagenicity, tumor promotion, and tumor inhibition test results (https://datadiscovery.nlm.nih.gov/Chemicals-and-Drugs/Chemical-Carcinogenesis-Research-Information-System/uvv3-fqwr/about_data).
3. The Istituto Superiore di Sanità Carcinogenicity and Mutation (ISSCAN) database is a long-term carcinogenicity bioassay results on rodents (rat and mouse) database provided by the Istituto Superiore di Sanità Institute.²⁵

After retrieving the curated combination version of these three databases from the previous study, the missing data of canonical SMILES and toxicity end points were then checked. The canonical SMILES were checked using the RDKit Python package (<https://rdkit.org/>). The duplicate entries were screened. After that, the redundant molecules compared to the external test sets described in Section 2.2 were removed from the training data, as they were used for external validation. The final data was 1599 compounds (see Supporting Information), comprising 872 noncarcinogens and 727 carcinogens, and used as a training data set.

2.2. External Test Sets Data Preparation. This study consists of two external test sets, which were used to assess the model's generalization and translation of the prediction to identify human carcinogens. The two external test sets were described as follows:

1. ISSCAN: Forty molecules from the ISSCAN database were selected as the primary test set to evaluate the model's performance. This test set was obtained from the previous studies^{11,13} and was curated by checking the inorganics, mixtures, and duplicate molecules. The

redundant molecules between this data set and the training set were removed from the training set. Overall, there are 17 noncarcinogens and 23 carcinogens, which sum up to 40 molecules in this test set (see [Supporting Information](#)).

2. IRIS and IARC: The human carcinogens data were compiled from the Integrated Risk Information System (IRIS) database (<https://www.epa.gov/iris>) and the International Agency for Research on Cancer (IARC) (<https://monographs.iarc.who.int/>). The IRIS data set was critically evaluated and reviewed by the United States Environmental Protection Agency (EPA). The molecules in this database were categorized by an evidence-based risk assessment approach, which makes it vulnerable to translational performance evaluation of the developed model. Moreover, the IARC test set was used in the test set, as it provides translational evidence for human carcinogens. This test set was critically reviewed and assessed the carcinogenic potential of substances and chemicals to humans under the working group of independent researchers, which makes it vulnerable to be used for translational and generalized human-based carcinogen identification.

In the IRIS test set, only category A (carcinogenic to humans) and category E (evidence of noncarcinogenicity for humans) were used as human carcinogens and nonhuman carcinogens, respectively. For the IARC test set, only class 1 (carcinogenic to humans) was used as human carcinogens. These two sets were combined into one test set and further retrieved their canonical SMILES from the PubChem database.²⁶ The missing data and duplicates were removed from this test set. The redundant molecules of this test set, compared to the training set, were retained; only the redundant molecules within the training set were removed. To ensure a more representative distribution of compounds, we expanded the test set by including additional noncarcinogens from the first test set. This was done to increase the number of negative samples for evaluation and not as a balancing procedure applied to the test set. Finally, the second test data comprised 73 chemicals, contributing to 53 human carcinogens (see [Supporting Information](#)).

2.3. Possibly Carcinogenic Chemicals. The possibly carcinogenic chemicals were obtained from group 2B (possibly carcinogenic as they are insufficient evidence of human carcinogenicity) of the IARC database. The canonical SMILES were retrieved from the PubChem database.²⁶ The IARC class 2B set was further preprocessed by removing missing data and duplicating entries. The redundant molecules of this carcinogenic assessment set compared to the training set were removed as they have their animal testing results from the training set.

We provide a detailed clarification regarding the entry “Kava extract”. In the original data set, this entry was annotated as “Kava extract” but was accompanied by complete SMILES information. Further verification using the PubChem database confirmed that this entry corresponds to a well-defined chemical entity, annotated either as “Kava” or by its systematic name, 2-[(E)-2-cyclohexa-1,5-dien-1-ylethenyl]-4-methoxy-2,3-dihydropyran-6-one. This compound belongs to the kavalactone family, a group of phytochemicals found in the Kava plant (*Piper methysticum*).²⁷ Importantly, this record does not represent an undefined extract or crude mixture; rather, the apparent ambiguity arises solely from the data set annotation. Therefore, we retained this record for subsequent analyses.

Furthermore, the record “Carbon black”, represented by the SMILES notation “[C]”, constitutes an ambiguous and structurally undefined input. Owing to the lack of meaningful molecular descriptors, it was excluded from the data set to ensure the reliability of ML-based predictions. As a result, only 104 candidate compounds remain to assess their carcinogenic assessment.

2.4. Molecular Representations. In this study, we utilized six molecular representations, including four chemical fingerprints (ECFP, MACCS, RDKit, E-state), SMILES tokenization (abbreviated as Tokens) and physicochemical descriptors (abbreviated as Phychem). These features were used as inputs for predictive modeling.

ECFP, also known as circular fingerprints, were calculated using the Morgan algorithm with a radius of 10 and a bit vector length of 4096. MACCS fingerprints, consisting of 167 predefined structural keys, were generated to represent the presence or absence of specific substructures within molecules. RDKit fingerprints were computed as 2048-bit path-based binary vectors that encode linear molecular fragments and capture overall topological connectivity. E-state fingerprints, computed using RDKit package, encode atom-level electronic and topological features into a 79-dimensional continuous vector that reflects molecular reactivity and local atomic environments.

For the SMILES tokenization, a unique vocabulary is contrasted from all characters present in the SMILES strings. Each character is assigned a unique integer index, with a special index reserved for padding purposes. After that, each SMILES string is converted into a sequence of integers based on the constructed vocabulary. Sequences are either truncated or padded to a fixed 100 maximum length characters to ensure uniformity across inputs.

A set of 12 physicochemical descriptors was computed using RDKit package to capture fundamental molecular properties. These included molecular weight (MolWt), Log P, number of hydrogen bond donors (NumHDonors), number of hydrogen bond acceptors (NumHAcceptors), topological polar surface area (TPSA), number of rotatable bonds (NumRotatableBonds), number of aromatic rings (NumAromaticRings), number of saturated rings (NumSaturatedRings), number of hetero atoms (NumHeteroatoms), number of rings for a molecule (RingCount), number of heavy atom (HeavyAtomCount), and number of aliphatic rings (NumAliphaticRings).

2.5. BiLSTM Classification of Carcinogenic Potential. A bidirectional long short-term memory (BiLSTM) neural network was implemented in Python using Keras 3.4.1, and Tensorflow 2.17.0 as a backend. The model architecture consisted of two stacked bidirectional LSTM layers, with 64 and 32 hidden units, respectively. To mitigate overfitting, dropout layers with a rate of 0.3 were applied after each BiLSTM block. The BiLSTM outputs were followed by a dense layer of 100 neurons with rectified linear unit (ReLU) activation, and a final output layer with a single neuron using the sigmoid activation function for binary classification.

The input features were reshaped into a three-dimensional format suitable for recurrent layers, with each compound represented as a one-step time series of molecular descriptors. The model was compiled with a binary cross-entropy loss function. A method for stochastic optimization (ADAM) was used as an optimizer,²⁸ with a fixed learning rate of 0.001. The training data were further split into training and validation sets in an 80:20 ratio. Model training was performed for 20 epochs with

a batch size of 32. To ensure reproducibility, all experiments were conducted using a fixed random seed (42) across NumPy, Python, and TensorFlow environments. After training, the model was evaluated on the external test set.

2.6. RF Classification of Carcinogenic Potential. A RF classifier was employed to predict the carcinogenic potential of chemical compounds based on molecular descriptors. The model was implemented using the `RandomForestClassifier` from the `scikit-learn` library, with 100 decision trees (`n_estimators` = 100) and default hyperparameters unless otherwise specified. The data set was randomly partitioned into training and validation subsets using an 80:20 split. A fixed random seed (`random_state` = 42) was applied to ensure reproducibility. Model performance was evaluated on the external test set.

2.7. LightGBM Classification of Carcinogenic Potential. In parallel, a gradient boosting classifier based on the LightGBM framework was used to assess the carcinogenic potential of compounds. The `LGBMClassifier` from the `lightgbm` library was initialized with 100 estimators and default hyperparameters. Input features were preprocessed as NumPy arrays to ensure compatibility with the model. The same 80:20 training–validation split and random seed (`random_state` = 42) were applied for consistency. Performance evaluation was conducted on the external test set following the same protocol as for the RF classifier.

2.8. Performance Evaluation Metrics. In this study, we employed a suite of complementary performance metrics that capture diverse aspects of classification behavior. Accuracy (eq 1) was used as a baseline metric, representing the overall proportion of correctly classified instances across both classes. We also included the Matthews correlation coefficient (MCC) (eq 2), which provides a balanced assessment by incorporating all four components of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).²⁹

To further dissect model behavior, we reported both sensitivity (eq 3), which reflects the true positive rate, and specificity (eq 4), which indicates the true negative rate. These metrics offer insights into the model's ability to correctly identify active and inactive compounds, respectively.

Finally, we included the area under the receiver operating characteristic curve (AUROC) as a threshold-independent metric.³⁰ Rather than measuring classification accuracy at a fixed cutoff, AUROC reflects the model's ability to rank positive instances above negative ones across all possible decision thresholds, effectively summarizing the trade-off between sensitivity and 1-specificity.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

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

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

2.9. Model Evaluation Via Train, Test Split and Cross-Validation. Model validation was performed on the full original training data set using both a fixed training/test split and stratified 5-fold cross-validation (5-fold CV), focusing on the three top-performing models identified in our study: LightGBM, RDKit, RF_E-state, and BiLSTM_MACCS.

First, a single stratified hold-out split was applied to the full original training data set, with 80% used for training and 20% for test. Stratification ensured that the class distribution was preserved across both partitions. This split was kept fixed for all experiments to guarantee fair comparison among models. Each model was trained with the same set of hyperparameters as in prior experiments to ensure consistency. To account for stochasticity in training, each model was trained independently three times with different random seeds. The seeds affected model initialization and batch shuffling, but the data split remained unchanged. Reported metrics of accuracy, MCC, AUROC, sensitivity, and specificity were recorded as the mean $\pm$ SD for the training and test partitions, while prediction probabilities ($y_{\text{predicted}}$) were averaged over three runs and retained to construct the corresponding confusion matrices for each partition.

In addition, a 5-fold CV was conducted on the full original training data set. Stratification again preserved class ratios within each fold. We employed `StratifiedKFold` (`n_splits` = 5, `shuffle` = `True`) to ensure reproducibility. In each round, a model was trained on four folds (training portion) and evaluated on the remaining fold (test portion). To enhance model diversity, a distinct random seed was assigned to each fold. This ensured variability in model initialization and optimization across folds while preserving identical data partitions. Each sample served as the test instance exactly once. Performance metrics were recorded as the mean $\pm$ SD across all five test portions. For reporting prediction probabilities ($y_{\text{predicted}}$) in cross-validation, we highlight the best-performing fold, selected based on the highest MCC, with AUROC and accuracy serving as secondary criteria in the event of a tie. The corresponding $y_{\text{predicted}}$ outputs were subsequently used to construct the associated confusion matrices.

2.10. Applicability Domain Analysis. To ensure the reliability of model predictions and to identify regions of trustworthy extrapolation, we employed an Applicability Domain (AD) analysis based on a k -nearest neighbors (k -NN) distance-based approach in the chemical descriptor space.^{10,21} The Euclidean distance between each test sample and the training data was computed using molecular fingerprints, followed by classification of test samples into within-domain and out-of-domain categories. The AD of the model can be expressed using the following mathematical conditions (eqs 5 and 6).

For compounds considered within-domain:

$$D_i \leq D_k + Z \cdot \sigma \quad (5)$$

For compounds judged as out-of-domain:

$$D_i > D_k + Z \cdot \sigma \quad (6)$$

Here, D_i denotes the Euclidean distance between the candidate compound and its nearest k training samples. The term D_k is the mean of such distances computed across the training set, while σ is their standard deviation. The parameter Z is an empirically selected scaling factor to control the sensitivity of the AD boundary; in this study, Z was fixed at 0.5.

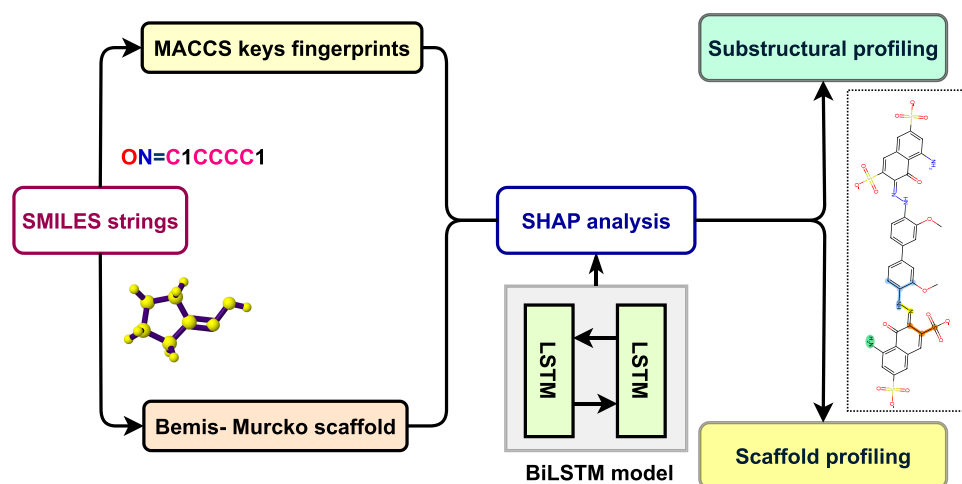


Figure 2. Identification of cheminformatics profiling, including substructure and scaffold profilings, associated with carcinogenicity of chemical compounds using BiLSTM-based SHAP analysis.

To rigorously assess domain applicability, we independently conducted AD analysis for each model using distinct molecular descriptor sets, specially RDKit for LightGBM, E-state for RF, and MACCS keys for BiLSTM. The evaluation was performed across a range of k -values from 2 to 10 to define the optimal neighborhood size in the k -nearest neighbors (k -NN) space. For each model, only the test compounds classified as within domain were retained for further analysis. The optimal k -value was selected based on a dual criterion: maximizing performance improvement across key metrics (accuracy, MCC, AUROC, sensitivity, and specificity) while simultaneously minimizing the proportion of out-of-domain samples, thereby balancing predictive robustness with domain coverage.

2.11. Activity Cliff Analysis. Activity cliff networks for the ISSCAN data set and the Human carcinogen data set were constructed using three molecular fingerprint representations: RDKit, E-state, and MACCS. The pairwise structural similarity between molecules was quantified using the Tanimoto coefficient,³¹ defined as

$$T(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where $|A \cap B|$ represents the number of common features (bits) present in both fingerprints A and B , and $|A \cup B|$ represents the total number of features present in either A or B . Molecular pairs with Tanimoto similarity greater than or equal to a threshold ($\tau = 0.85$) were considered high-similarity pairs. A pair was classified as an activity cliff if the two molecules had different activity labels (carcinogen vs noncarcinogen).

2.12. Deployment of Carcinogenicity Predictor Based on Web Server Tool. To enable real-time and user-friendly access to our carcinogenicity prediction models, we developed an interactive web platform based on the Streamlit framework (v1.46.0)³² in Python 3.10. The application supports both single-molecule and batch-mode predictions, accepting input as either individual SMILES strings or CSV files. All molecular structures were first standardized via canonicalization. Subsequently, a comprehensive set of molecular descriptors was computed using the RDKit cheminformatics toolkit. Predictions were obtained through a consensus-based approach that aggregated the outputs of three top-performing models. These model pipelines were serialized in *.keras* or *.pkl* formats.³²

The web server was hosted on Streamlit Cloud and synchronized with a public GitHub repository for transparent and reproducible deployment. Prediction results included both binary classifications (Possible carcinogen/Noncarcinogen) and corresponding confidence probabilities. All computations were executed within a containerized environment, ensuring session-specific processing without external data transmission or storage. This architecture ensured secure handling of user-submitted chemical data while promoting broad accessibility and scientific reproducibility.

2.13. Bemis–Murcko Scaffold Decomposition. To systematically characterize the core structural frameworks of the compounds, Bemis–Murcko scaffolds were extracted from their SMILES representations using the RDKit cheminformatics toolkit (version 2023.9.6). For each molecule, the scaffold was generated via the `GetScaffoldForMol` function from the `MurckoScaffold` module, which retains all ring systems and the linkers connecting them, while removing side chains and terminal substituents.^{33,34} These resulting scaffolds reflect the central molecular architectures commonly associated with biological activity. Scaffold extraction enabled consistent comparisons across structurally diverse compounds and facilitated the identification of recurring cores potentially linked to carcinogenicity. The chemical structures and corresponding names of the extracted scaffolds were retrieved from the PubChem database (<https://pubchem.ncbi.nlm.nih.gov/>).

2.14. Cheminformatics Profiling Employing BiLSTM-Based SHAP Analysis. To elucidate the influence of specific molecular features on the BiLSTM model's carcinogenicity predictions, we employed SHapley Additive exPlanations (SHAP) analysis, a model-agnostic, game-theoretic approach that delivers consistent and locally faithful estimates of feature importance.³⁵ In this study, SHAP values were derived using the `KernelExplainer` on the training data, with MACCS structural keys and Bemis–Murcko scaffolds serving as input descriptors to capture both substructural elements and core molecular frameworks. This integrative strategy enabled a detailed interpretation of the chemical patterns most strongly driving the model's classification decisions between carcinogenic and noncarcinogenic compounds. A comprehensive visualization of these SHAP-based attributions is provided in Figure 2.

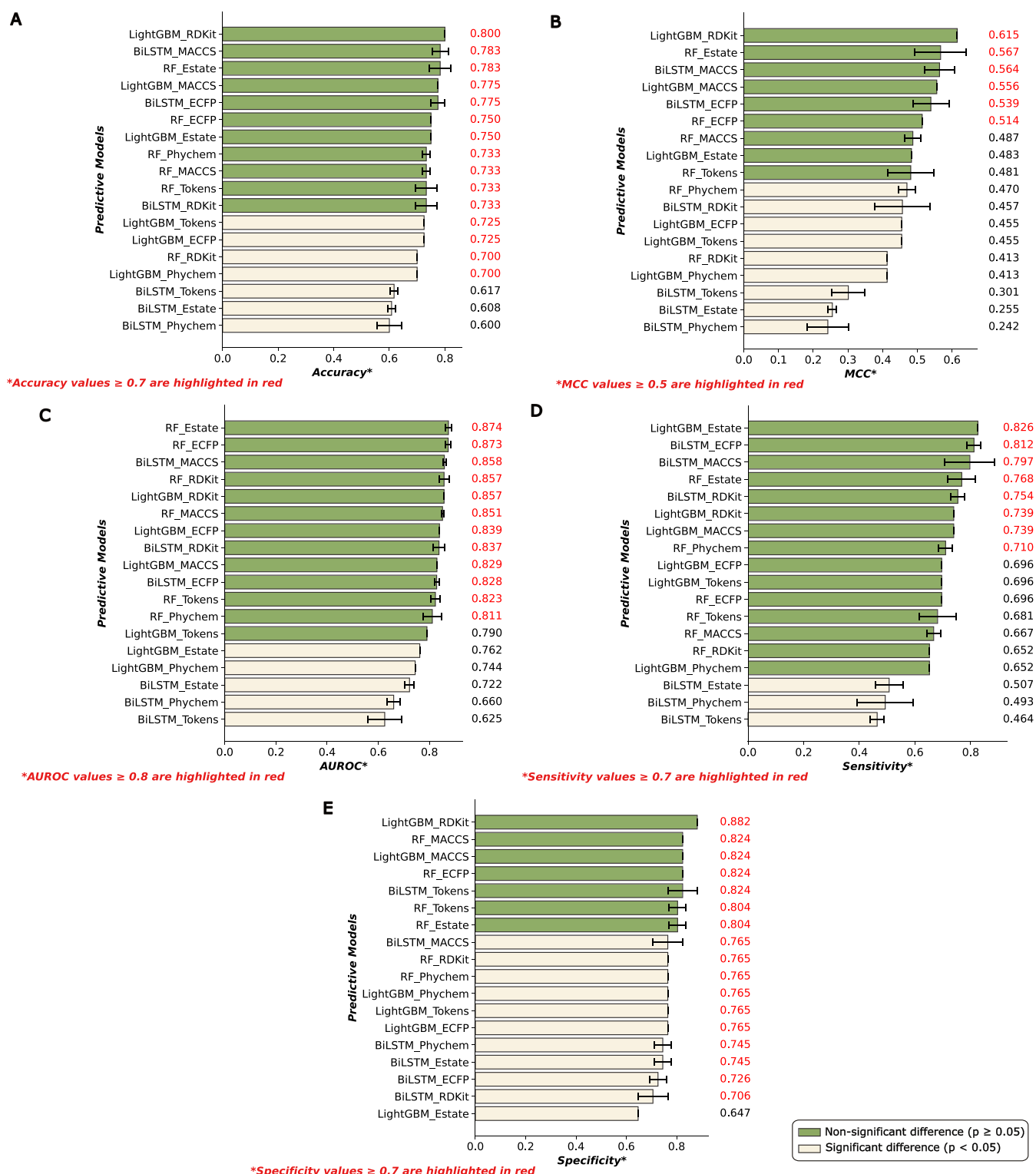


Figure 3. Evaluation metrics of predictive models on the ISSCAN external test set ($n = 40$). (A) Accuracy; (B) MCC; (C) AUROC; (D) Sensitivity; (E) Specificity. All values were obtained from three independent runs.

3. RESULTS

3.1. Performance of Predictive Models on the ISSCAN External Test Set. In this study, six chemical representations (ECFP, E-state, MACCS, Phychem, RDKit, and Tokens) and three ML architectures (BiLSTM, LightGBM, and RF) were employed to construct a total of 18 predictive models. Their performance was evaluated using five metrics, as illustrated in

Figure 3. The top three models with the highest accuracy were: LightGBM using RDKit fingerprints (0.800), BiLSTM with MACCS fingerprints (0.783), and RF with E-state descriptors (0.783). Notably, the LightGBM-RDKit model also achieved the highest MCC (0.615) and specificity (0.882). The RF model with E-state descriptors showed the highest AUROC value

Table 1. Comparative Analysis between Three Best-Performing Models and State-of-the-Art Models

predictive models	accuracy	MCC	AUROC	sensitivity	specificity
LightGBM_RDKit	0.800 ± 0.000	0.615 ± 0.000	0.882 ± 0.000	0.739 ± 0.000	0.857 ± 0.000
RF_E-state	0.783 ± 0.031	0.567 ± 0.060	0.874 ± 0.010	0.768 ± 0.041	0.804 ± 0.028
BiLSTM_MACCS	0.783 ± 0.024	0.564 ± 0.035	0.858 ± 0.005	0.797 ± 0.074	0.765 ± 0.048
CarcinoPred-EL ¹¹	0.700	-	0.803	0.652	0.765
CapsCarcino ¹³	0.818	0.640	0.819	0.789	0.857

(0.874), while BiLSTM-MACCS ranked third in terms of sensitivity among all models.

Subsequently, a comprehensive statistical analysis was conducted. The Shapiro-Wilk normality test revealed that all evaluation metrics deviated from a normal distribution ($p < 0.05$). Therefore, nonparametric statistical comparisons were performed using the Kruskal–Wallis test followed by Dunn's post hoc test to ensure the robustness and validity of the analysis. The statistical results showed that the top three models, specially LightGBM with RDKit, BiLSTM with MACCS, and RF with E-state, did not exhibit statistically significant differences across all evaluation metrics, except for a single significant difference in specificity observed for the BiLSTM–MACCS model. Detailed results of the statistical comparisons are illustrated in Figure 3, where statistically significant and nonsignificant differences are represented by distinct colors. The comprehensive performance of all predictive models can be found at Supporting Information.

Table 1 provides a comparative analysis of the predictive performance of our top three models against CarcinoPred-EL¹¹ and CapsCarcino,¹³ two state-of-the-art tools for carcinogen prediction evaluated on a similar test data set. Notably, all of our models consistently achieved higher accuracy, AUROC, sensitivity, and specificity scores than those reported for CarcinoPred-EL, suggesting superior classification performance across all evaluated metrics. The CarcinoPred-EL is an ensemble learning that trained on conventional ML models.¹¹ This data underscores that our frameworks outperformed the previous ensemble learning framework in term of overall correctness, ranking capability and the ability to identify TP and TN. Moreover, we also found that all of our models consistently achieved higher AUROC scores than those reported for CapsCarcino, suggesting a superior ability to discriminate between carcinogenic and noncarcinogenic compounds. Given that AUROC captures the overall ranking capability across all thresholds,³⁶ this superiority underscores the potential and competitive edge of our models in carcinogen prediction. Additionally, the sensitivity and specificity of our BiLSTM_MACCS and LightGBM_RDKit models achieved a performance on par with the CapsCarcino model, meaning that our approach contain similar performance for identifying TP and TN molecules. Nonetheless, the accuracy and MCC of our best model slightly less than the CapsCarcino model, indicating that CapsCarcino model achieved marginally better performance in overall correctness and balanced classification.

3.2. Performance of Predictive Models on the Human Carcinogen Test Data Set. Figure 4 presents the predictive performance of the developed models on the human carcinogen test data set. Notably, the LightGBM_RDKit model consistently achieved superior results across all evaluation metrics, attaining an accuracy of 0.753, a MCC of 0.535, an AUROC of 0.842, a specificity of 0.900, and a sensitivity of 0.698. In comparison, the BiLSTM_MACCS model ranked second, with performance metrics of accuracy, MCC, AUROC, specificity, and sensitivity being 0.703, 0.410, 0.803, 0.783, and 0.673, respectively. These

results highlight the effectiveness of the LightGBM_RDKit architecture in capturing discriminative features relevant to carcinogenicity prediction. Notably, only these two models achieved an accuracy exceeding 0.70, indicating that they correctly predicted over 70% of instances in the human carcinogen test data set, which comprises both carcinogenic and noncarcinogenic compounds. Furthermore, the AUROC quantifies the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative one.^{37,38} Both LightGBM_RDKit and BiLSTM_MACCS models ranked among the top performers with AUROC values exceeding 0.80, reflecting excellent discriminative performance and a robust ability to distinguish carcinogens from noncarcinogens.³⁹ Nevertheless, the RF_E-state model exhibited inferior performance metrics compared to the other two models, LightGBM_RDKit and BiLSTM_MACCS. These results underscore the superior generalization ability of both LightGBM_RDKit and BiLSTM_MACCS on the human carcinogen test sets. The comprehensive performance of all predictive models can be found at Supporting Information.

In additional aspects, BiLSTM has demonstrated effectiveness in toxicity prediction and risk assessment, due to its ability to capture sequential dependencies and contextual patterns within molecular descriptors. Its bidirectional architecture enhances the representation of substructural interactions relevant to adverse biological effects.^{21–23} Moreover, the predictive performance observed across both the ISSCAN external validation set and the human carcinogen test set supports the reliability of the proposed predictive models. To capitalize on their complementary strengths, we integrated all three models, including LightGBM_RDKit, BiLSTM_MACCS, and RF_E-state, into a consensus framework to enhance the robustness and credibility of carcinogenicity prediction, particularly for compounds classified as IARC class 2B.

3.3. Performance Assessment with Train/Test and 5-Fold CV. Table 2 provides a detailed comparison of the predictive performance across the top three models—LightGBM_RDKit, RF_E-state, and BiLSTM_MACCS under different validation strategies, while confusion matrices for all models and validation strategies are presented in Figure 5. Each model demonstrates unique strengths that together offer complementary perspectives for the task of carcinogenicity prediction.

The LightGBM_RDKit model achieves the highest training accuracy (0.985 ± 0.000) and an almost perfect AUROC (1.000 ± 0.000), indicating its strong ability to learn complex patterns from the training data. Notably, while its performance decreases on the test set, the model still maintains a competitive AUROC (0.810 ± 0.001) along with well-balanced sensitivity and specificity. These characteristics make it especially suitable for scenarios where reducing false negatives and ensuring reliable overall classification are essential.

The RF_E-state model also shows strong performance, with excellent consistency on the training set (accuracy: 0.977 ± 0.000 , AUROC: 0.999 ± 0.000), and slightly better general-

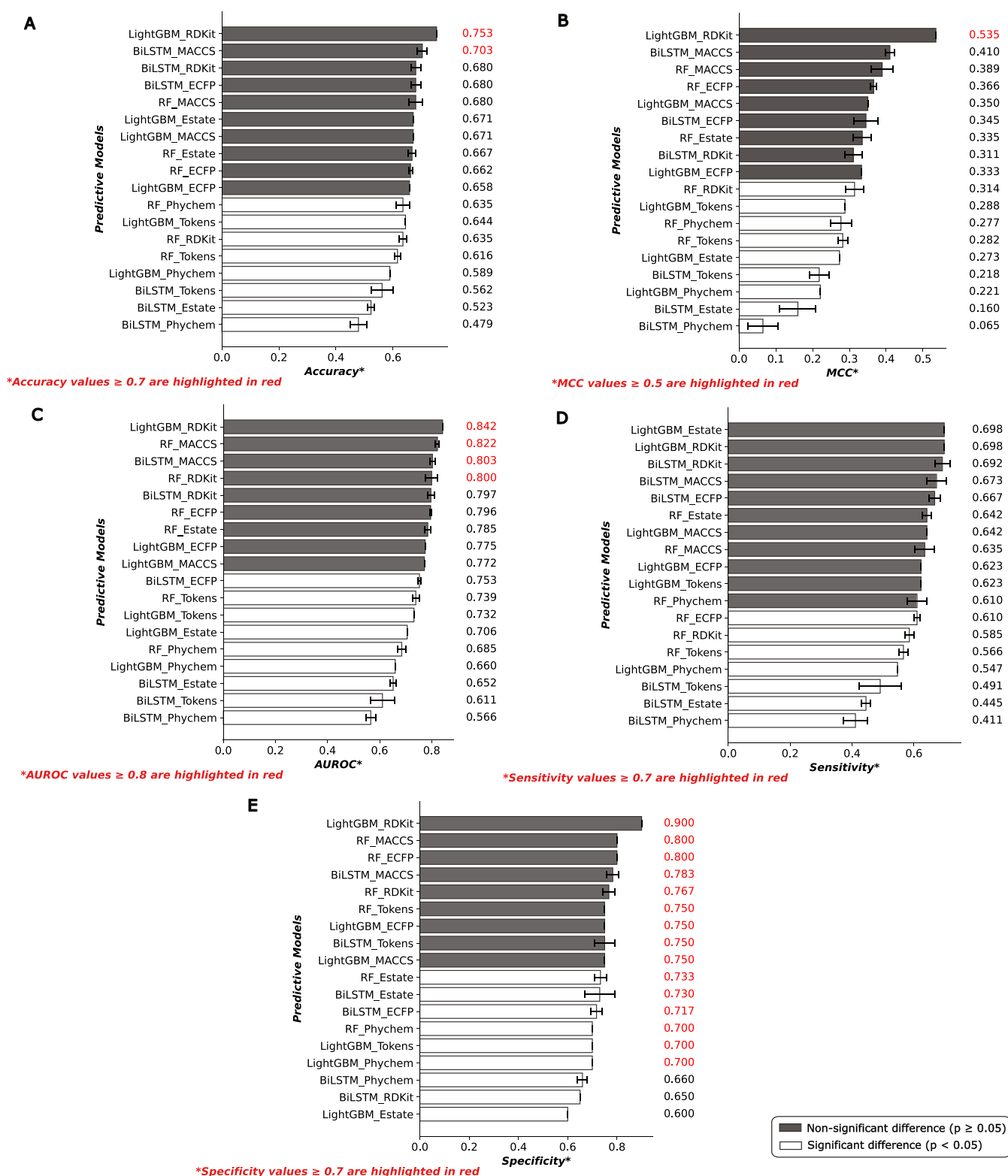


Figure 4. Evaluation metrics of predictive models on human carcinogen set ($n = 73$). (A) Accuracy; (B) MCC; (C) AUROC; (D) Sensitivity; (E) Specificity. All values were obtained from three independent runs.

ization to the test set compared to LightGBM_RDKit. Notably, it achieves the highest test accuracy (0.748 ± 0.005) and specificity (0.775 ± 0.003) among all traditional models, highlighting its effectiveness in correctly identifying inactive compounds—an essential feature for reducing false positives in high-throughput screening workflows.

Meanwhile, although the BiLSTM_MACCS model shows slightly lower performance on the training set, it offers several practical advantages. It performs competitively on the test set (accuracy: 0.732 ± 0.018 ; AUROC: 0.786 ± 0.008) and achieves the highest specificity (0.802 ± 0.056), which highlights its strength in reducing false positives. In addition,

Table 2. Performance Metrics for both Validation Strategies from the Three Top-Performing Models

models	accuracy	MCC	AUROC	sensitivity	specificity
Training Subset					
LightGBM_RDKit	0.985 ± 0.000	0.970 ± 0.000	1.000 ± 0.000	0.987 ± 0.002	0.984 ± 0.002
RF_E-state	0.977 ± 0.000	0.954 ± 0.000	0.999 ± 0.000	0.979 ± 0.002	0.976 ± 0.002
BiLSTM_MACCS	0.873 ± 0.021	0.750 ± 0.036	0.945 ± 0.003	0.813 ± 0.083	0.923 ± 0.035
Test Subset					
LightGBM_RDKit	0.729 ± 0.001	0.455 ± 0.003	0.810 ± 0.001	0.713 ± 0.009	0.743 ± 0.008
RF_E-state	0.748 ± 0.005	0.491 ± 0.011	0.809 ± 0.004	0.715 ± 0.009	0.775 ± 0.003
BiLSTM_MACCS	0.732 ± 0.018	0.462 ± 0.035	0.786 ± 0.008	0.648 ± 0.107	0.802 ± 0.056
5-Fold CV					
LightGBM_RDKit	0.732 ± 0.017	0.460 ± 0.035	0.793 ± 0.022	0.682 ± 0.055	0.774 ± 0.033
RF_E-state	0.717 ± 0.020	0.429 ± 0.041	0.783 ± 0.025	0.664 ± 0.029	0.761 ± 0.035
BiLSTM_MACCS	0.684 ± 0.015	0.368 ± 0.032	0.745 ± 0.008	0.691 ± 0.051	0.678 ± 0.035

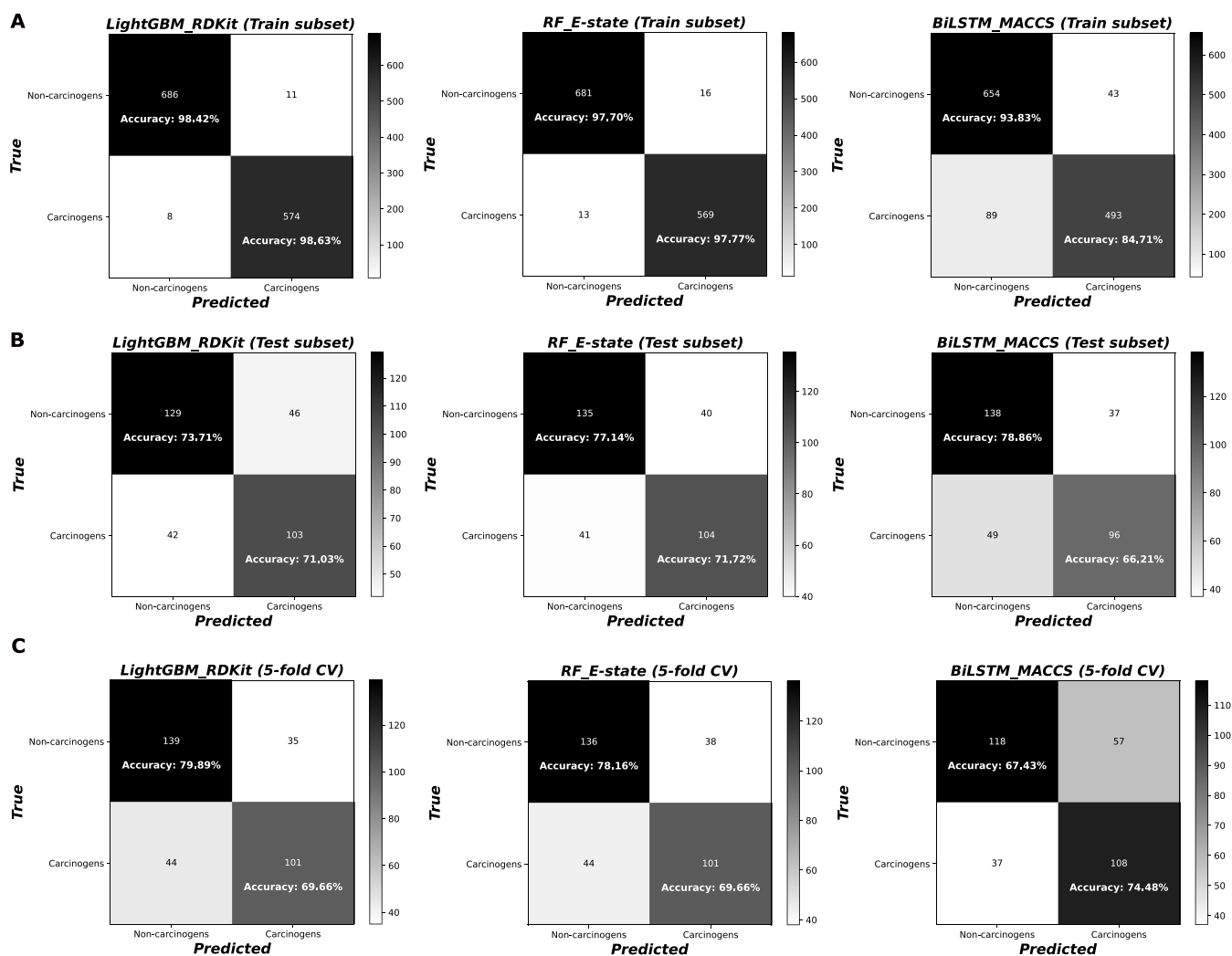


Figure 5. Confusion matrices for the three best-performing models and validation strategies. (A) Training subsets ($n = 1,279$); (B) Test subsets ($n = 320$); (C) 5-fold CV. All experiments were conducted using the full original training data set.

the model demonstrates consistent results across 5-fold CV, as reflected in its low standard deviations. This indicates good stability across different data partitions, making it a reliable and adaptable choice for real-world use.

In summary, the three models exhibit complementary strengths: LightGBM_RDKit excels in learning capacity, RF_E-state ensures strong generalization, and BiLSTM_MACCS delivers stable, fold-consistent performance. Their

lower test performance, despite strong training results, reflects a typical generalization gap and suggests room for further optimization. These findings support the potential of a consensus-based model selection or integration strategy to enhance predictive reliability in practical deployment scenarios.

3.4. AD for Three Best-Performing Models. To balance the trade-off between performance improvement and the retention of test samples within the AD, we determined the

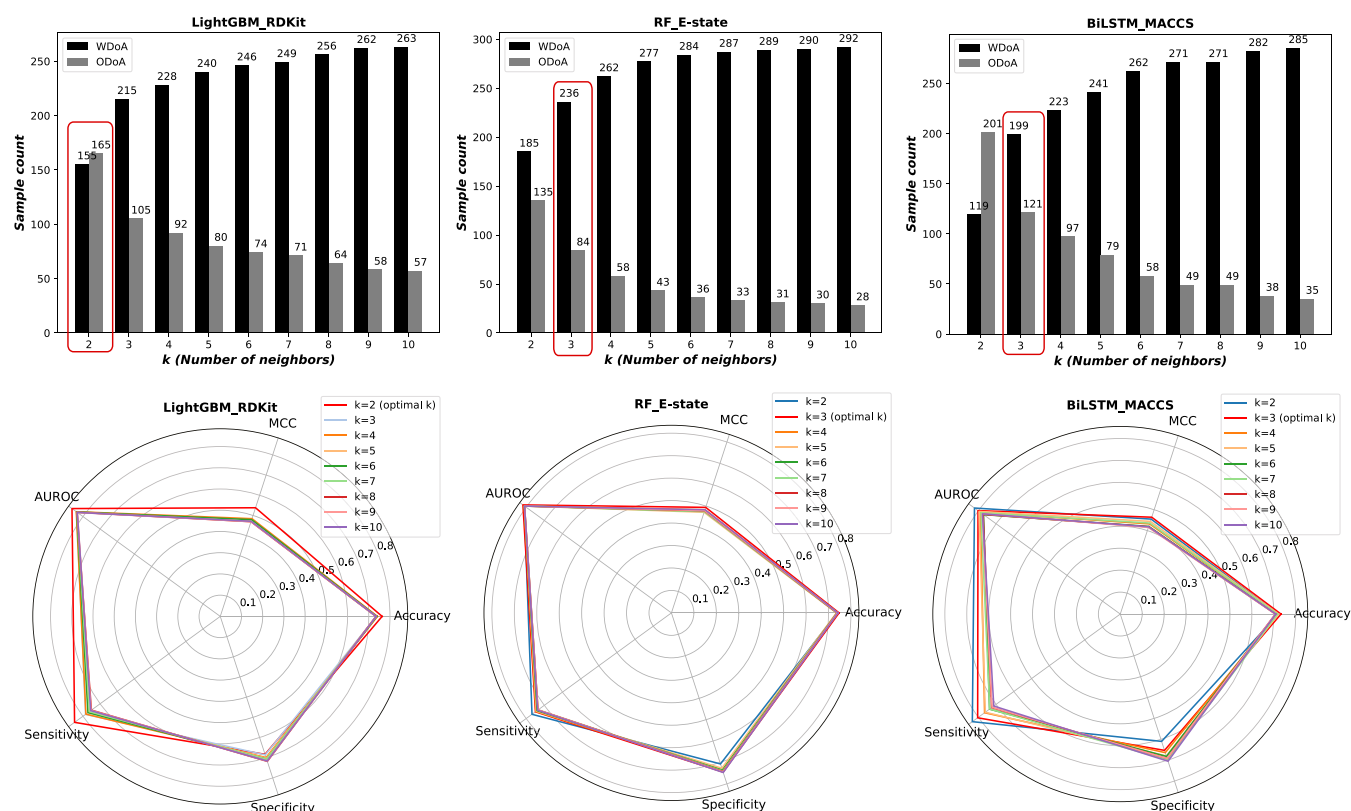


Figure 6. AD of LightGBM_RDKit, RF_E-state, and BiLSTM_MACCS. (A) Number of within-domain and out-of-domain samples. (B) Performance metrics on refined test sets (k ranging from 2 to 10), with the optimal k highlighted.

optimal value of k for each model–descriptor pair based on the minimal exclusion of out-of-domain instances while maximizing performance metrics. Specifically, the optimal k was found to be 2 for LightGBM_RDKit, and 3 for both RF_E-state and BiLSTM_MACCS. At the optimal k value, the LightGBM_RDKit model exhibited a substantial improvement, with the MCC increasing by 18.24%, the AUROC by 6.67%, and the overall accuracy by 4.66%. In contrast, the RF_E-state model showed only marginal gains: MCC improved by 0.61%, AUROC by 1.11%, while the overall accuracy remained nearly unchanged (0.748 for the original data vs 0.747 for the refined data). Meanwhile, the BiLSTM_MACCS model achieved a moderate performance boost, with MCC rising by 8.67%, AUROC by 1.91%, and overall accuracy by 2.66%. The number of within-domain and out-of-domain samples across the full k range (2–10), along with corresponding performance metrics on the refined test sets, are comprehensively summarized in Figure 6 with the selected optimal k values clearly annotated.

The selected k values were subsequently employed to delineate the within-domain compounds from the IARC class 2B data set. This filtering step was performed prior to the consensus-based carcinogenicity assessment, thereby ensuring that only structurally relevant and reliable predictions were included in the final evaluation.

3.5. Activity Cliff Network Analysis across Fingerprint Types. We constructed activity cliff networks for the ISSCAN and human carcinogen data sets using RDKit, E-state, and MACCS fingerprints at a Tanimoto similarity threshold of $\tau = 0.85$. In the ISSCAN data set, none of the three fingerprints, RDKit, E-state, or MACCS, identified any activity cliffs. This absence suggests a well-behaved and consistent activity landscape, in which structurally similar compounds tend to

exhibit similar biological responses. Such a profile is highly favorable for machine learning, as it supports the core assumption underlying most predictive models in cheminformatics.

In contrast, the human carcinogen data set revealed a more complex activity landscape. Among seven highly similar molecular pairs, three were identified as activity cliffs—all exclusively detected by the E-state fingerprint. Neither RDKit nor MACCS detected any cliffs under the same threshold, suggesting a more stable structure–activity relationship. While E-state may capture subtle structural variations that lead to divergent activities, this sensitivity also exposes regions of the chemical space where predictive modeling is likely to struggle. In comparison, the consistency observed with RDKit and MACCS is likely to support more reliable model performance.

The corresponding network visualizations (Figure 7) provide deeper insights into the interplay between structural features and biological activity. Carcinogenic and noncarcinogenic compounds segregate into distinct yet interconnected clusters (green and pink nodes, respectively). Red edges represent activity cliffs, while blue edges denote high structural similarity coupled with consistent activity annotations. Furthermore, representative compound pairs forming activity cliffs within the E-state space of the human carcinogen data set are extracted and illustrated in Figure 8, where similar structural motifs are highlighted and the true activity labels of each compound are explicitly annotated.

We further evaluated the predicted outcomes of six structurally related compounds (Figure 8) derived from the identified active cliff pair using a consensus framework. This framework integrates prediction probabilities derived from the RF_E-state model alongside BiLSTM_MACCS and Light-

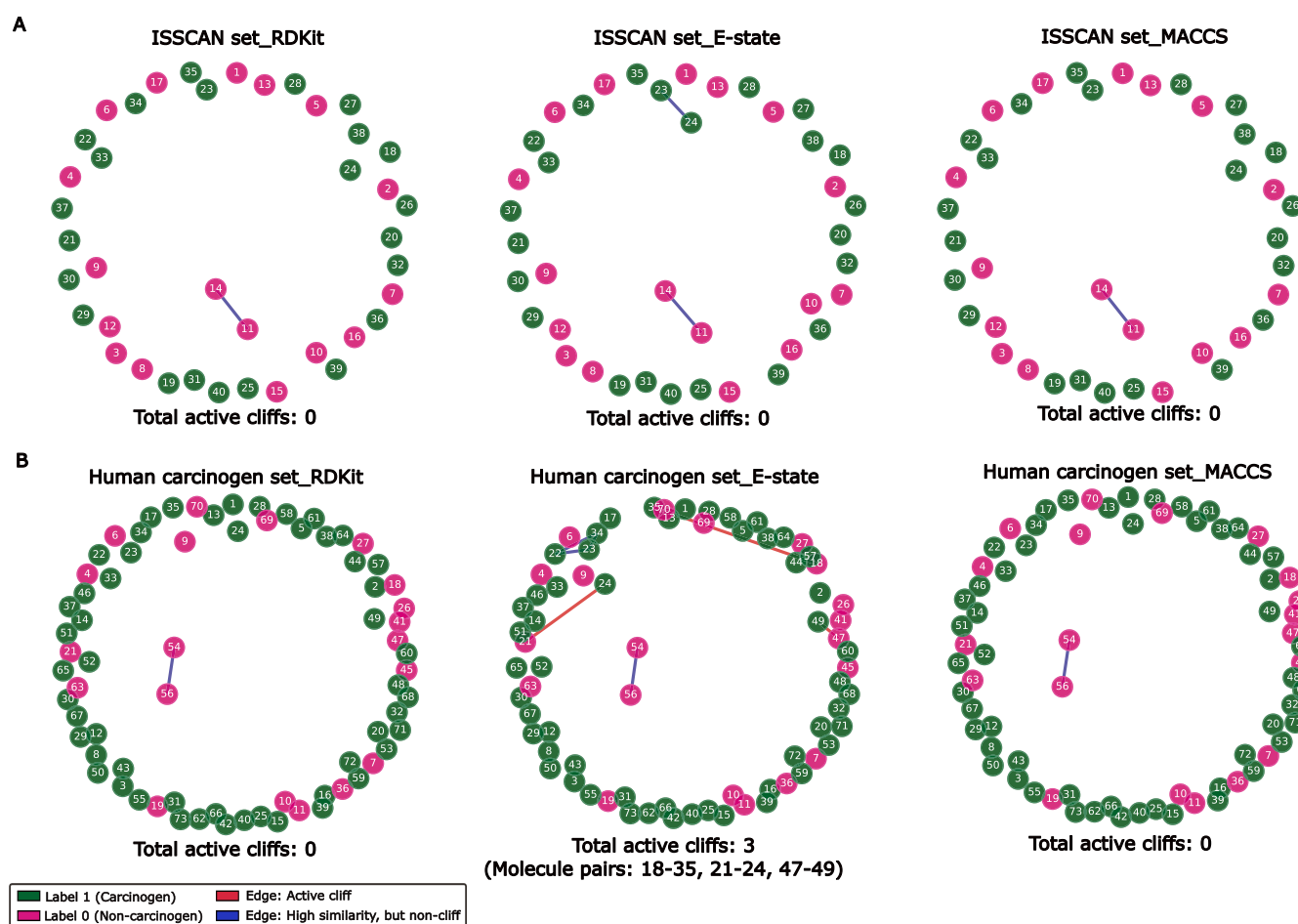


Figure 7. Activity cliff networks for ISSCAN data set and human carcinogen data set using RDKit, E-state, and MACCS fingerprints (similarity threshold $\tau = 0.85$). Nodes are colored by carcinogenicity label (green = label 1 and pink = label 0). Red edges indicate activity cliffs; blue edges indicate high similarity, but noncliffs. (A) ISSCAN data set ($n = 40$) and (B) Human carcinogen data set ($n = 73$).

GBM_RDKit. Owing to inconsistencies among the individual model outputs, we adopted a consensus rule: if all models yielded prediction probabilities ≥ 0.5 , the compound was classified as “Carcinogen”, whereas probabilities < 0.5 across all models led to the assignment of “Noncarcinogen”. In cases where model predictions were discordant, the compound was designated as “Inconclusive”. Using this consensus strategy, three compounds were correctly classified in agreement with their experimental labels, while three were deemed inconclusive. The results are summarized in Table 3. Notably, only one compound within the active cliff pair was correctly predicted, underscoring the intrinsic challenge posed by active cliffs. This observation highlights the necessity of carefully scrutinizing predicted labels in such cases and demonstrates the value of consensus modeling in reconciling conflicting outputs from individual algorithms. Accordingly, the principles of this consensus approach were extended to subsequent analyses in this study.

3.6. Assessment of Potential Carcinogenicity in IARC Class 2B Compounds Using a Consensus Framework. We first identified the within-domain samples from the IARC Group 2B data set for each selected individual model, utilizing their respective optimal k values determined during AD analysis. Notably, both the LightGBM_RDKit and BiLSTM_MACCS models yielded an identical number of within-domain compounds ($n = 66$); however, the specific agents comprising

each set were not completely overlapping. To further investigate consensus predictions, we focused on the intersection of test compounds simultaneously falling within the AD of all three models, resulting in a subset of 47 within-domain samples. A comprehensive summary of these findings is presented in Table 4.

In this analysis, a consensus modeling approach was employed by integrating predictions from three distinct models: LightGBM_RDKit, BiLSTM_MACCS, and RF_E-state, to assess the potential carcinogenicity of 47 within-domain compounds from the IARC class 2B data set. The SMILES representations of the compounds were first standardized and then transformed into corresponding chemical descriptors according to each model's feature requirements. For each compound, predicted probabilities were generated independently by all three models.

The consensus framework was defined as follows:

1. A compound was classified as a **Potential carcinogen** if all three models predicted a probability ≥ 0.5 .
2. A compound was classified as a **Presumed non-carcinogen** if all three models predicted a probability < 0.5 .
3. If there was any disagreement among the models (i.e., at least one probability differed from the others in classification), the compound was labeled as **Inconclusive**.

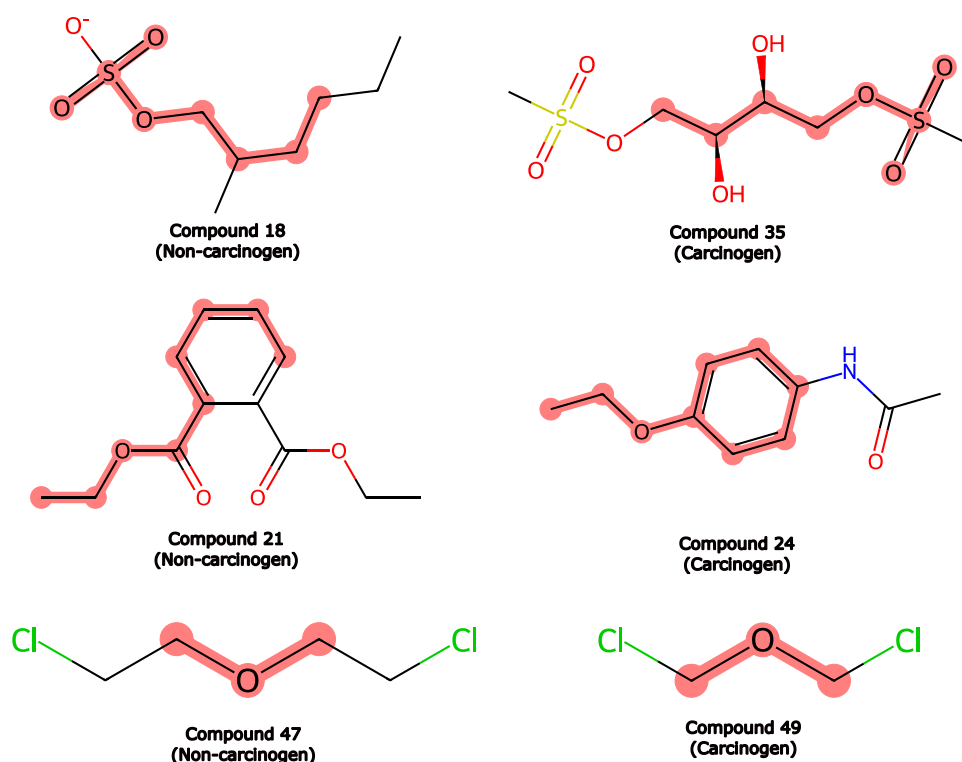


Figure 8. Representative activity-cliff compound pairs (18–35, 21–24, and 47–49) from the E-state space of the human carcinogen data set, with highlighted structural similarities and their corresponding true activity labels.

Table 3. Predicted Outcomes for Six Compounds Structurally Related to the Active Cliff Molecule Pair, as Obtained from the Consensus Framework

compound	true label	RF (E-state)	BiLSTM (MACCS)	LGBM (RDKit)	predicted label	assessment
18	noncarcinogen	0.263	0.148	0.321	noncarcinogen	true
21	noncarcinogen	0.145	0.328	0.294	noncarcinogen	true
24	carcinogen	0.377	0.172	0.575	inconclusive	inconclusive
35	carcinogen	0.261	0.369	0.506	inconclusive	inconclusive
47	noncarcinogen	0.918	0.971	0.36	inconclusive	inconclusive
49	carcinogen	0.904	0.979	0.646	carcinogen	true

Table 4. Number of Within-Domain and Out-of-Domain Samples among IARC 2B Compounds, Identified by Three Models Using Domain Applicability Filtering

IARC 2B samples	within-domain samples	out-of-domain samples	model
104	66	39	LightGBM_RDKit
	73	32	RF_E-state
	66	39	BiLSTM_MACCS
	47	57	common to all three models

In parallel, the average and standard deviation of the predicted probabilities were calculated for each compound across the three models. This value reflects the extent to which individual model predictions deviate from their mean. A higher standard deviation indicates greater variation and less agreement among the models, while a lower value suggests consistent predictions. In this context, the standard deviation serves as a measure of uncertainty in the classification outcome for each compound.

Based on this strategy, out of 47 evaluated compounds, 16 were identified as **Potential carcinogens**, 8 as **Presumed noncarcinogens**, and 23 were categorized as **Inconclusive**.

Comprehensive results of these three subsets are provided in the [Supporting Information](#).

Among the compounds predicted as **Potential carcinogens**, the 16 candidates from the IARC class 2B data set are summarized in [Table 5](#). These compounds exhibited the highest average predicted probabilities and the greatest level of consensus across the LightGBM_RDKit, BiLSTM_MACCS, and RF_E-state models, underscoring strong alignment in their carcinogenic risk assessment. Ranked in descending order by mean probability scores, these compounds represent the most confidently and consistently predicted carcinogenic candidates. Similarly, [Table 6](#) highlights all 8 compounds classified as **Presumed noncarcinogens**, all of which showed consistently low individual and mean probabilities (<0.5), reflecting agreement on their likely noncarcinogenicity. These prioritized lists serve as valuable references for subsequent toxicological validation and experimental follow-up.

Nonetheless, from the IARC class 2B data set, a total of 23 compounds were identified as **Inconclusive** due to inconsistent predictions across the three developed models. These instances exhibited conflicting classification outputs, indicating a lack of consensus regarding their carcinogenic potential. Compounds with high predictive uncertainty, ranked by their standard

Table 5. 16 Potential Carcinogens from IARC Class 2B Data Sets, Identified by Predictive Models, Ranked by Mean Probabilities in Descending Order

agent	LGBM	RF	BiLSTM	mean	Std	predictive label
lasiocarpine	0.9995	0.98	0.986	0.9885	0.01	potential carcinogen
phenazopyridine hydrochloride	0.9995	0.9517	0.9796	0.9769	0.024	potential carcinogen
2-amino-3-methyl-9H-pyrido[2,3-b]indole	0.9983	0.9342	0.9744	0.9689	0.0324	potential carcinogen
benz[j]aceanthrylene	0.9987	0.8543	0.9399	0.931	0.0726	potential carcinogen
phenoxybenzamine hydrochloride	0.9981	0.8967	0.8303	0.9083	0.0845	potential carcinogen
1,2-dichloroethane	0.822	0.8339	0.9371	0.8643	0.0633	potential carcinogen
sterigmatocystin	0.9998	0.8983	0.6854	0.8612	0.1605	potential carcinogen
1,3-dichloropropene	0.8281	0.8273	0.8635	0.8396	0.0207	potential carcinogen
4-methylimidazole	0.9853	0.7647	0.7434	0.8311	0.134	potential carcinogen
chloroprene	0.9696	0.7433	0.706	0.8063	0.1427	potential carcinogen
bromodichloromethane	0.7875	0.7168	0.9132	0.8059	0.0995	potential carcinogen
benzo[c]phenanthrene	0.65	0.8883	0.8508	0.7963	0.1281	potential carcinogen
vinylidene chloride	0.7618	0.7558	0.8458	0.7878	0.0503	potential carcinogen
1,1,1,2-tetrachloroethane	0.709	0.7268	0.8522	0.7627	0.0781	potential carcinogen
chloroform	0.7299	0.743	0.7718	0.7482	0.0214	potential carcinogen
1-bromopropane	0.7264	0.5533	0.9122	0.7306	0.1794	potential carcinogen

Table 6. All 8 Presumed Non-Carcinogens from IARC Class 2B Data Sets, Identified by Predictive Models, Ranked by Mean Probabilities in Descending Order

agent	LGBM	RF	BiLSTM	mean	std	predictive label
butyl methacrylate	0.2719	0.4622	0.2774	0.3372	0.1083	presumed noncarcinogen
dimethyl hydrogen phosphite	0.1195	0.4153	0.4709	0.3352	0.1889	presumed noncarcinogen
dichloroacetic acid	0.3218	0.35	0.2829	0.3182	0.0337	presumed noncarcinogen
1,3-dichloro-2-propanol	0.0867	0.4936	0.1731	0.2511	0.2144	presumed noncarcinogen
quinoline	0.0076	0.435	0.1829	0.2085	0.2149	presumed noncarcinogen
isoeugenol	0.0293	0.1864	0.3773	0.1977	0.1743	presumed noncarcinogen
1,4-dichloro-2-nitrobenzene	0.001	0.2308	0.2013	0.1444	0.125	presumed noncarcinogen
2,4-dichloro-1-nitrobenzene	0.0002	0.2308	0.2013	0.1441	0.1255	presumed noncarcinogen

Table 7. 23 Agents with Inconsistent Model Predictions from IARC Class 2B Data Sets, Identified by Predictive Models, Ranked by Standard Deviation in Descending Order

agent	LGBM	RF	BiLSTM	mean	std	predictive label
<i>para</i> -nitroanisole	0.0236	0.9301	0.606	0.5199	0.4594	inconclusive
anthracene	0.8725	0.1133	0.8508	0.6122	0.4322	inconclusive
leucomalachite green (see also malachite green)	0.5056	0.1872	0.945	0.5459	0.3805	inconclusive
<i>ortho</i> -phenylenediamine dihydrochloride	0.9868	0.2612	0.4978	0.582	0.3701	inconclusive
benzyl violet 4B	0.9976	0.5667	0.3525	0.6389	0.3286	inconclusive
sodium <i>ortho</i> -phenylphenate	0.9869	0.7858	0.3546	0.7091	0.3231	inconclusive
diphenylamine	0.0234	0.6102	0.5413	0.3916	0.3208	inconclusive
zalcitabine	0.9971	0.7527	0.3626	0.7041	0.32	inconclusive
1-bromo-3-chloropropane	0.8868	0.435	0.9839	0.7686	0.2929	inconclusive
daunomycin (daunorubicin)	0.9982	0.84	0.4417	0.76	0.2868	inconclusive
CI basic red 9	0.9959	0.9167	0.4741	0.7956	0.2812	inconclusive
dibromoacetic acid	0.9817	0.4492	0.5683	0.6664	0.2795	inconclusive
magenta	0.0058	0.2992	0.5423	0.2824	0.2686	inconclusive
<i>Fusarium moniliforme</i> , toxins derived from (fumonisin B1, fumonisin B2, and fusarin C)	0.9957	0.7067	0.4709	0.7244	0.2628	inconclusive
<i>N,N</i> -dimethyl- <i>p</i> -toluidine	0.8887	0.4412	0.6351	0.655	0.2244	inconclusive
carbazole	0.2877	0.2633	0.6013	0.3841	0.1885	inconclusive
nitromethane	0.2542	0.3322	0.6105	0.399	0.1873	inconclusive
acetamide	0.6179	0.4723	0.7946	0.6283	0.1614	inconclusive
β -myrcene	0.252	0.5534	0.3339	0.3798	0.1558	inconclusive
cumene	0.6319	0.4467	0.3622	0.4803	0.1379	inconclusive
<i>para</i> -chloroaniline	0.5847	0.3597	0.4208	0.455	0.1163	inconclusive
acetaldehyde	0.5846	0.3623	0.4829	0.4766	0.1113	inconclusive
chlordane	0.4753	0.3983	0.5886	0.4874	0.0957	inconclusive

deviation rather than mean values in predicted probabilities, are presented in Table 7. This ranking highlights compounds with

the highest variability among model predictions, serving as potential candidates for further examination.

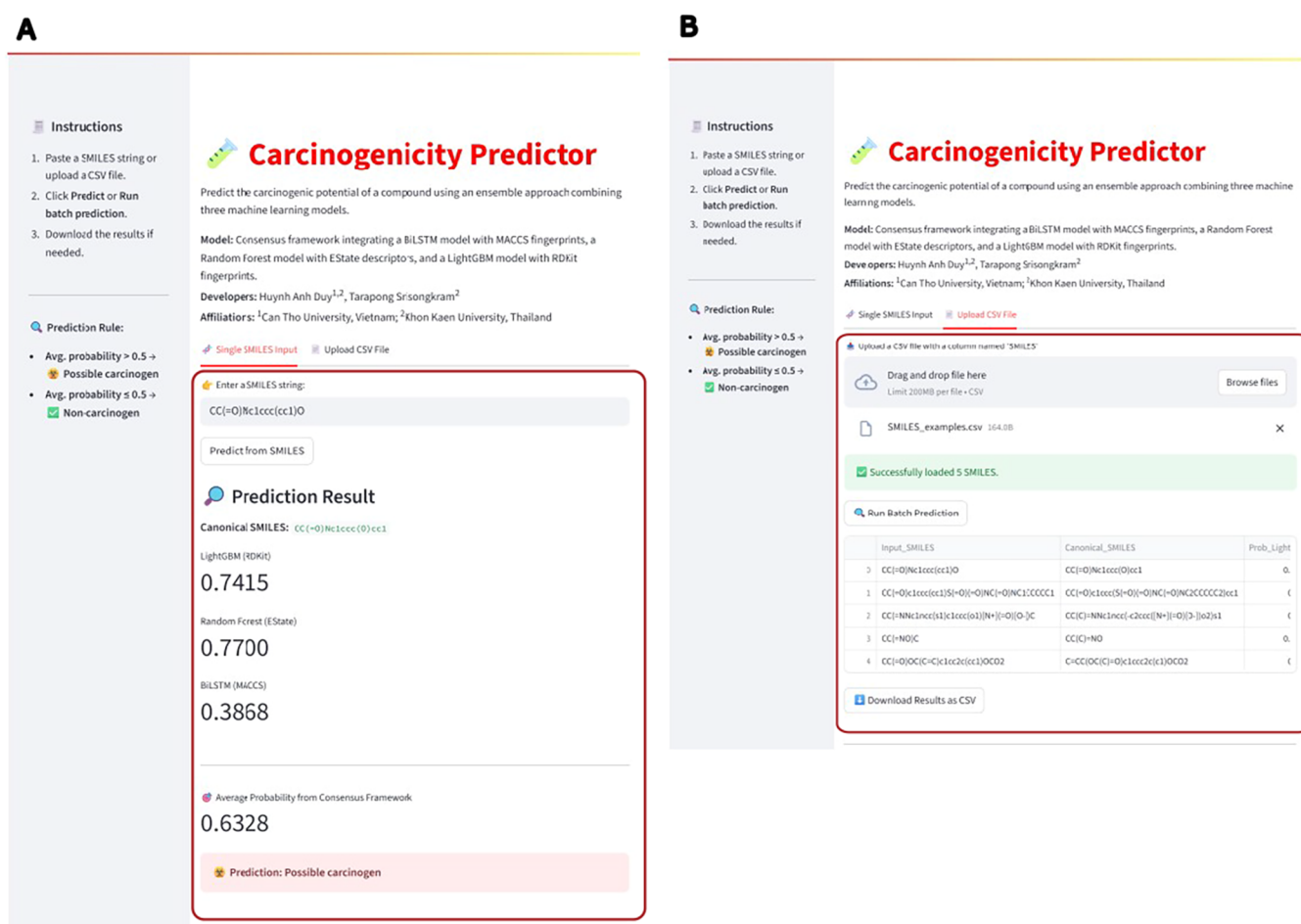


Figure 9. Example of prediction results for SMILES input via (A) single-input SMILES entry and (B) batch-input CSV file.

3.7. Web-Server Tool for Carcinogen Predictor Based on a Consensus Framework. Subsequently, we employed this consensus framework, combining the three best-performing models (i.e., LightGBM_RDKit, BiLSTM_MACCS, and RF_Estate), to deploy a web-based tool for carcinogenicity prediction. A publicly accessible web interface was developed to allow real-time prediction of carcinogenic potential directly from SMILES input. The platform is available at <https://carcinogenicity-predictor.streamlit.app/> and supports both single-compound and batch-mode analysis with an emphasis on usability and data confidentiality. Each SMILES string, after preprocessing, was transformed into three distinct molecular fingerprint representations: MACCS keys, RDKit fingerprints, and E-state descriptors. These representations served as input features for three independently trained models: a BiLSTM neural network (MACCS), a LightGBM model (RDKit), and a RF classifier (E-state). For each compound, the final carcinogenicity prediction was derived by averaging the output probabilities of the three models. A compound was classified as a possible carcinogen if the mean probability exceeded 0.5; otherwise, it was categorized as noncarcinogenic.

In single-input mode, users can submit an individual SMILES string to obtain a binary classification alongside a confidence probability score. In batch-processing mode, a comma separate value (CSV) file containing multiple SMILES entries can be uploaded to perform high-throughput screening. The prediction results are rendered dynamically within the interface and are also available for download in a structured tabular format for further

analysis. Figure 9A,B showcase examples illustrating the web server's functionality and usability in both prediction modes: single-input SMILES and batch-input CSV file formats.

3.8. Substructural Profiling of Carcinogens Using SHAP Values Derived from MACCS Keys Fingerprints. Substructural profiling of carcinogenic compounds was performed using SHAP values derived from MACCS keys fingerprints. This approach allowed us to identify specific molecular substructures that had the greatest influence on the predictions made by the BiLSTM_MACCS model.²¹ The SHAP-based interpretation revealed consistent patterns of substructures and functional groups commonly linked to carcinogenicity. These informative features may serve as interpretable indicators to support chemical risk assessment and prioritization. Figure 10A illustrates the top 20 most influential features contributing to the prediction of chemical carcinogenicity, ranked in descending order based on the absolute values of their mean SHAP scores. In contrast, the SHAP summary plot in Figure 10B provides a more detailed view of how each feature impacts the model's output, highlighting whether its contribution is predominantly positive or negative toward carcinogenic classification.

Based on these observations, we subsequently identified the MACCS features that positively contributed to the prediction of carcinogenicity. Specifically, nine key MACCS fingerprint features were filtered as the most influential substructures. Detailed information on these selected features, including their corresponding SMARTS patterns and substructure interpreta-

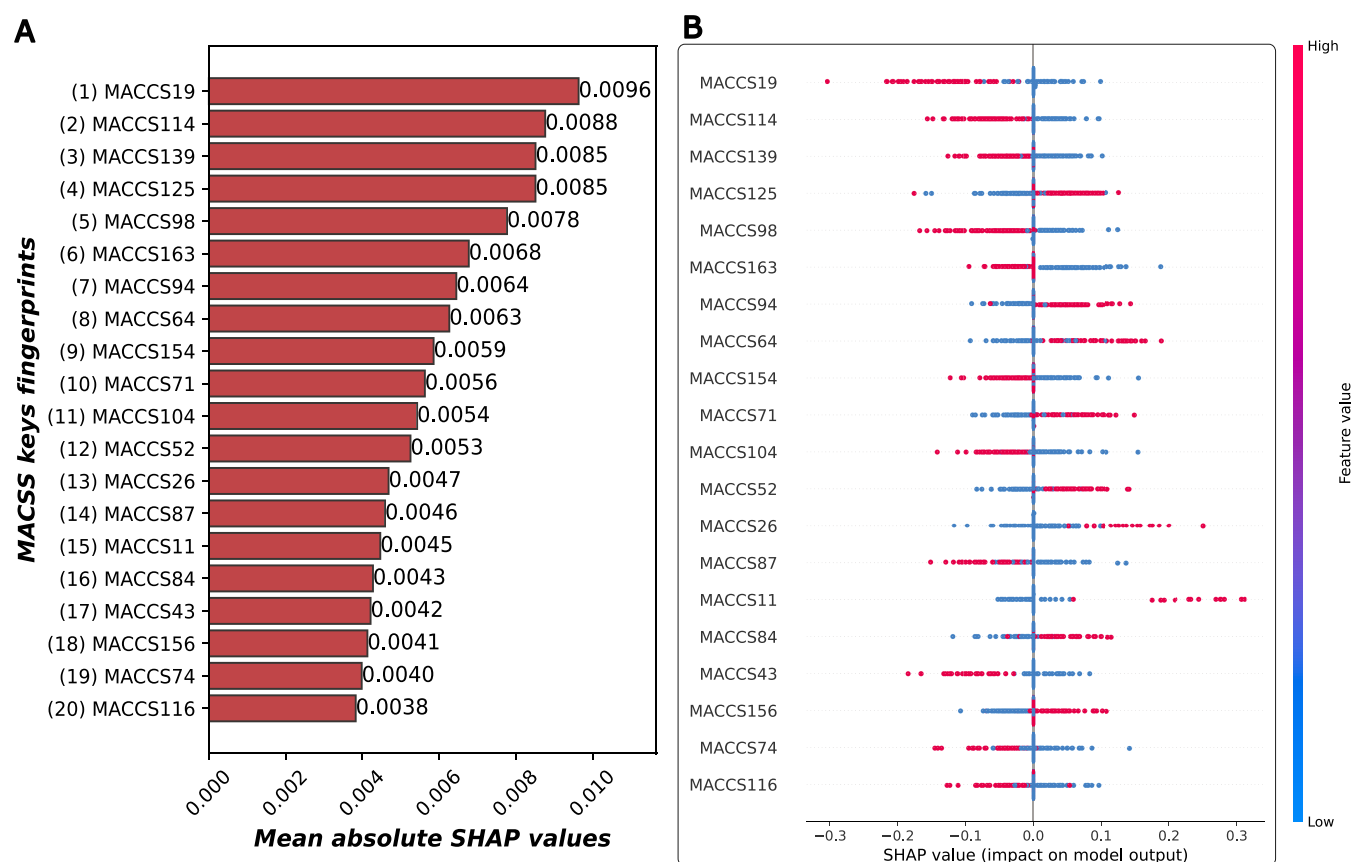


Figure 10. Crucial MACCS key fingerprints influencing carcinogenicity predictions. (A) Mean absolute SHAP values of the top 20 features in descending order. (B) SHAP summary plot showing features with positive and negative impacts on carcinogenicity predictions.

Table 8. MACCS Keys Fingerprints Positively Influencing Carcinogenicity Predictions

MACCS keys	SMARTS patterns	interpretations	possible chemical class
MACCS125	?	aromatic ring >1	polycyclic aromatic system
MACCS94	[!#6;!#1]~[#7]	N–X bond (X ≠ C, H)	nitrogen–heteroatom bond
MACCS64	*@*!@[#16]	A\$A!S	sulfur-substituted aromatic ring
MACCS71	[#7]~[#8]	N–O bond	nitroso or hydroxylamine group
MACCS52	[#7]~[#7]	N–N bond	azo or hydrazine group
MACCS26	[#6]=;@[#6](@*)@*	C=C(A)A	substituted alkene
MACCS11	*1~*~*~*~*1	4-membered ring	cyclobutane or azetidine ring
MACCS84	[NH2]	NH2	primary amine
MACCS156	[#7]~*(~*)~*	NA(A)A	tertiary-substituted nitrogen

tions, is summarized in Table 8. Notably, MACCS125, which indicates the presence of multiple aromatic rings, was emerged as the most important feature. This finding aligns with established toxicological evidence, as polycyclic aromatic compounds are strongly associated with carcinogenic potential.^{40,41} In addition, several nitrogen-containing substructures (MACCS94, MACCS71, MACCS52, MACCS84, and MACCS156), as well as sulfur-containing groups (i.e., MACCS64), were also identified as significant contributors, further supporting their relevance in predicting carcinogenicity.

We propose the assignment of chemical class labels based on the definitions of MACCS structural keys, corresponding precisely to each SMARTS pattern. These assignments are intended to provide a consistent, interpretable mapping between substructural fingerprints and recognizable chemical functionalities. The resulting chemical class annotations have been systematically integrated into Table 8, thereby enhancing the

semantic clarity of the molecular features extracted and facilitating downstream interpretability in cheminformatics modeling.

To gain deeper insights into the model's learned representations, we curated a subset of carcinogenic compounds from the training data that prominently exhibited the most impactful MACCS features identified during interpretation. Consequently, 11 representative carcinogens were selected, showcasing distinct structural patterns, as visualized with their corresponding MACCS features in Figure 11.

The relative contribution of each feature was assessed by computing mean absolute SHAP values across all samples, and then dividing each by the sum of all features to obtain percentage contributions. Table 9 summarizes the relative impact of MACCS structural keys on carcinogenicity prediction, with larger percentages reflecting stronger effects on model performance. Notably, MACCS key 19 emerged as the most influential

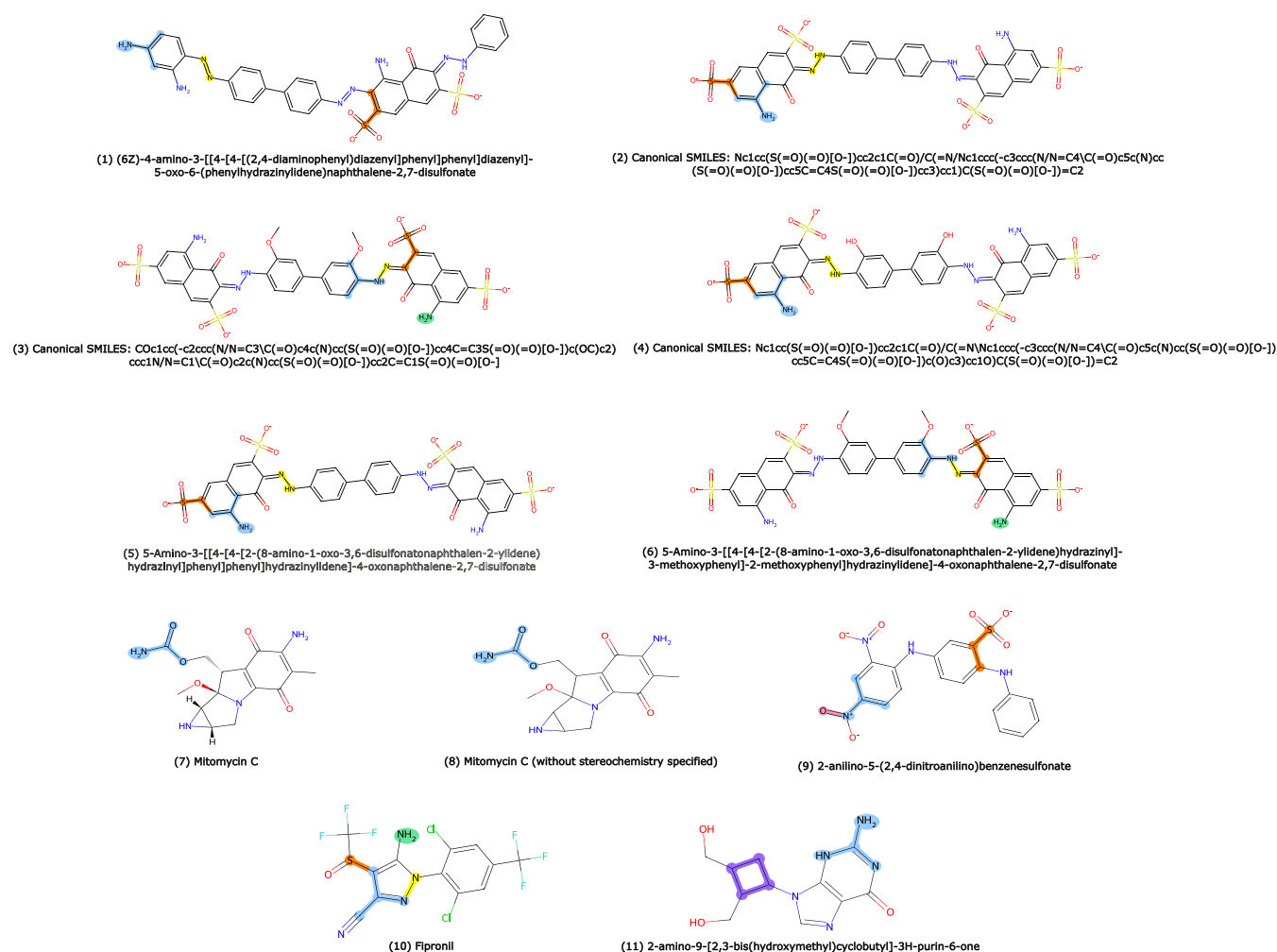


Figure 11. MACCS visualizations of selected carcinogens from the training data set ($n = 11$). Highlighted substructures include: MACCS64 (A\$A!S, orange), MACCS71 (NO, peach-pink), MACCS52 (NN, yellow), MACCS11 (4-membered ring, lavender), MACCS84 (NH₂, mint green), and MACCS156 (NA(A)A, light blue). Based on the canonical SMILES, compound names were annotated according to the PubChem database. Samples without an assigned compound name were annotated using their corresponding canonical SMILES.

Table 9. Percentage Contribution of Top MACCS Features Based on SHAP Values

no.	feature	contribution (%)	no.	feature	contribution (%)
1	MACCS19	2.65	11	MACCS104	1.49
2	MACCS114	2.4	12	MACCS52	1.45
3	MACCS139	2.31	13	MACCS26	1.29
4	MACCS125	2.26	14	MACCS87	1.24
5	MACCS98	2.14	15	MACCS11	1.23
6	MACCS163	1.86	16	MACCS43	1.16
7	MACCS94	1.73	17	MACCS84	1.15
8	MACCS64	1.68	18	MACCS156	1.12
9	MACCS154	1.6	19	MACCS74	1.06
10	MACCS71	1.55	20	MACCS116	1.05

feature (mean SHAP value = 0.0096; 2.65% contribution), whereas among the positively oriented features, MACCS125 showed the highest relevance (mean SHAP value = 0.0085; 2.26% contribution).

Additionally, based on the predictive outcomes for the IARC class 2B data set shown in Table 5, we identified four compounds, specially lasiocarpine, phenazopyridine hydrochloride, 2-amino-3-methyl-9H-pyrido[2,3-*b*]indole, and benz-

[*j*]aceanthrylene, as the most probable carcinogens, each displaying a mean predictive probability greater than 0.9 and demonstrating strong agreement across the consensus framework. Impressively, these compounds contain multiple aromatic rings or bear critical carcinogenic substructures emphasized by SHAP analysis, such as MACCS125 (aromatic ring >1), MACCS52 (N–N bond), MACCS84 (NH₂), and MACCS156 (N-A(A)A). Their molecular structures, along with the highlighted MACCS fingerprint substructures, are illustrated in Figure 12.

3.9. Distribution of Bemis–Murcko Scaffolds across Carcinogenic and Noncarcinogenic Groups. To elucidate structural patterns underlying carcinogenicity, we performed a scaffold enrichment analysis on the training data set using Bemis–Murcko framework extraction. A total of 601 unique scaffolds were identified, among which nine exhibited statistically significant differences in their distribution between carcinogenic and noncarcinogenic compounds (Fisher's exact test, $p < 0.05$). Notably, seven of these scaffolds were predominantly associated with carcinogens, whereas the remaining two were significantly enriched in noncarcinogens. The comprehensive results are summarized in Table 10.

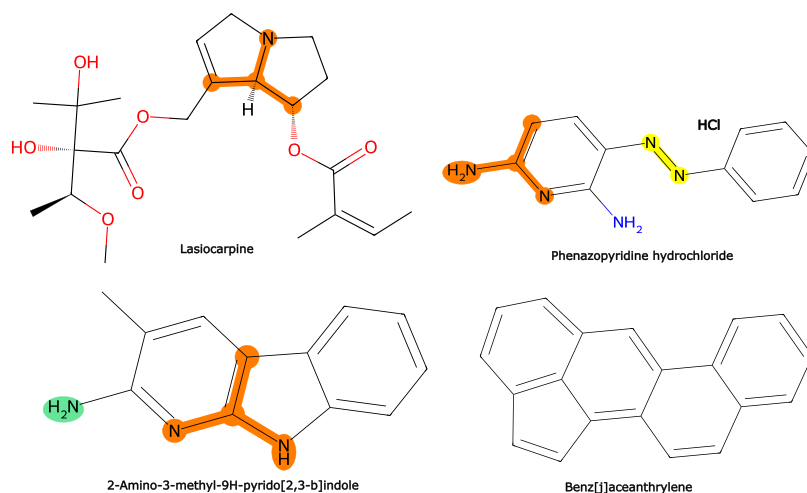


Figure 12. MACCS visualizations of the top potential carcinogens from the IARC class 2B data set ($n = 4$). Highlighted substructures include: MACCS52 (NN, yellow), MACCS84 (NH_2 , mint green), and MACCS156 (NA(A)A, amber orange).

Table 10. Representative Scaffolds Most Enriched in Carcinogen and Non-Carcinogen Groups, Identified Based on Odds Ratios (ORs) and Statistically Significant Differences between the Two Groups^{a,b}

Bemis–Murcko scaffold	name of scaffold	CGs	NCGs	ORs	<i>p</i> -value ^c
c1ccccc1	benzene	91	168	0.6	0.00029
c1ccc(N=Nc2ccccc2)cc1	azobenzene	6	0	ND	0.00873
c1ccc(-c2ccccc2)cc1	1,1'-biphenyl	25	13	2.35	0.01291
c1coc(-c2cscn2)c1	4-(furan-2-yl)-1,3-thiazole	8	1	9.69	0.01381
O=C1c2ccccc2C(=O)c2ccccc21	anthracene-9,10-dione	8	1	9.69	0.01381
O=C1OC2(c3ccccc3Oc3ccccc32)c2ccccc21	fluoran	0	6	0	0.03483
c1ccc2cc3c(ccc4ccccc43)cc2c1	benzo[<i>a</i>]anthracene	4	0	ND	0.04254
C1=Cc2c(ccc3c2ccc2ccccc23)CC1	1,2-dihydrochrysene	4	0	ND	0.04254
c1ncncn1	1,2,4-triazine	4	0	ND	0.04254

^aCGs = carcinogens; NCGs = noncarcinogens. ^bND: Not detectable. ^cFisher's exact test, *p*-values <0.05.

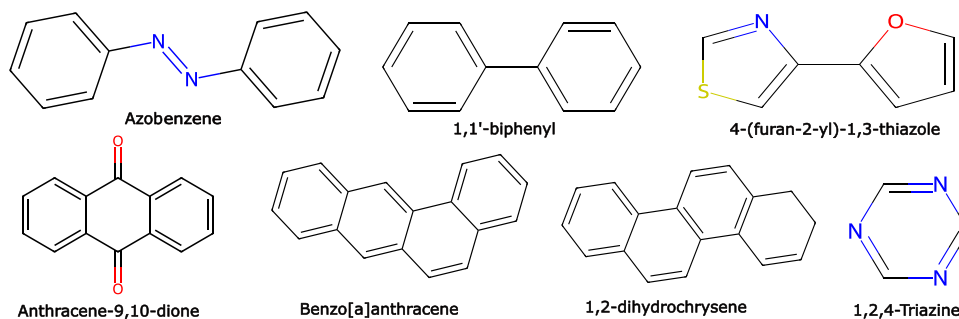


Figure 13. Key scaffolds of carcinogens identified from the training data set.

The representative chemical structures of carcinogen-associated scaffolds are depicted in Figure 13, while those linked to noncarcinogens are shown in Figure 14. Notably, the seven carcinogen-enriched scaffolds largely consist of fused polycyclic aromatic hydrocarbons (PAHs) (that is, benzo[*a*]anthracene, 1,2-dihydrochrysene, and anthracene-9,10-dione),⁴² heterocyclic aromatic systems (that is, 4-(furan-2-yl)-1,3-thiazole and 1,2,4-triazine), or simple aromatic structures (that is, azobenzene and 1,1'-biphenyl). These findings reinforce the association between specific structural motifs and carcinogenic potential, corroborating earlier insights from SHAP-based interpretation with MACCS fingerprints, which identified multiple aromatic rings, nitrogen- or sulfur-containing moieties as pivotal contributors directly associated with carcinogenic compounds.

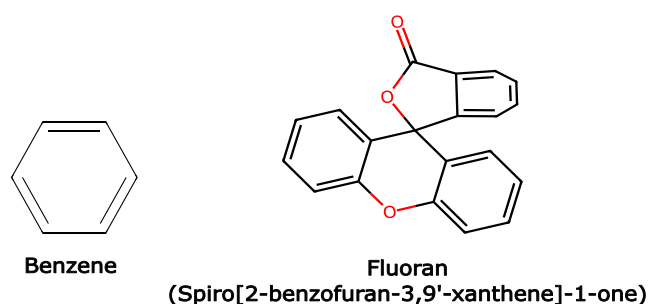


Figure 14. Key scaffolds of noncarcinogens identified from the training data set.

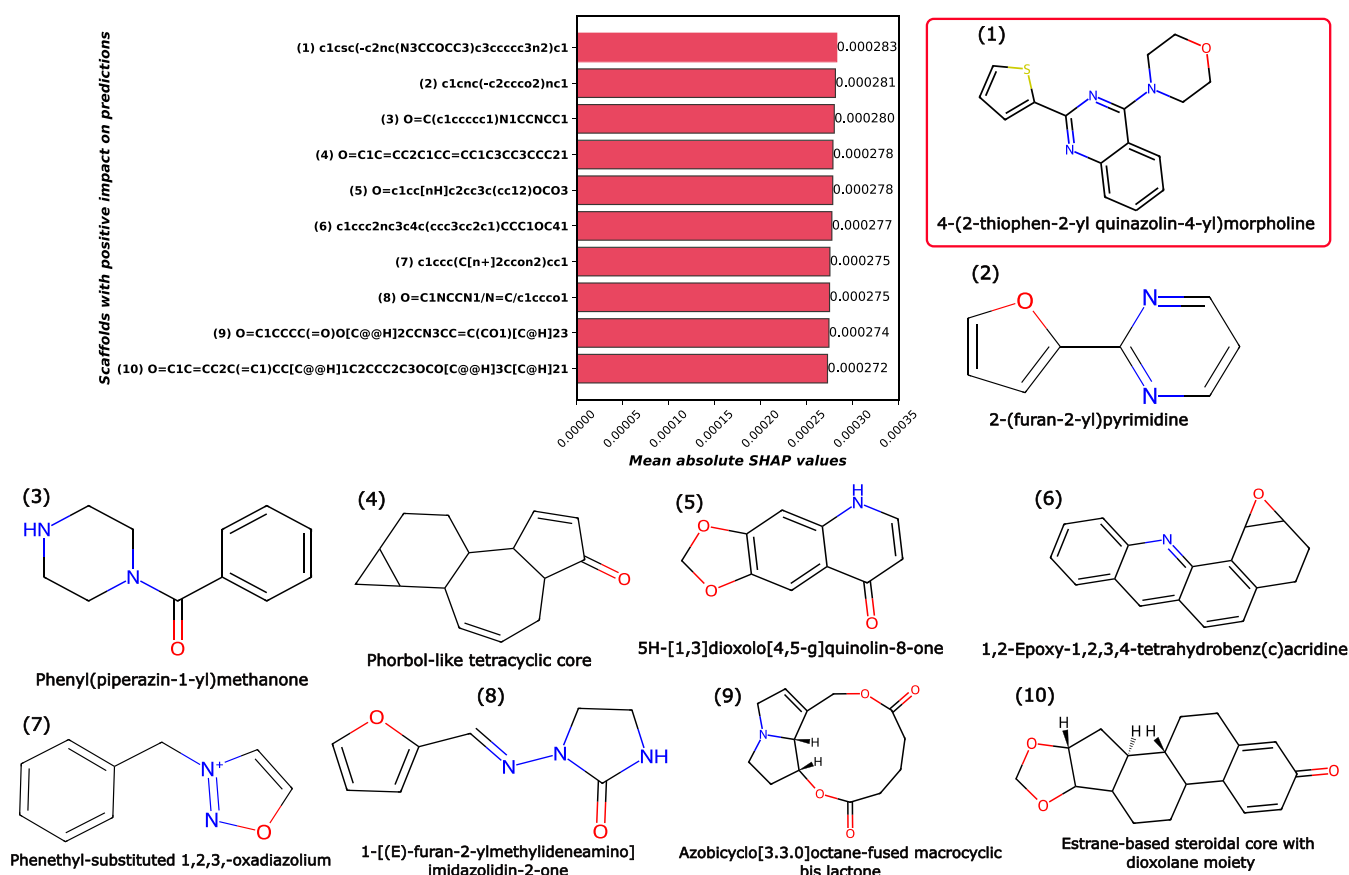


Figure 15. Ten key scaffolds with a positive impact on the prediction as revealed by SHAP analysis. (1) 4-(2-thiophen-2-yl quinazolin-4-yl)morpholine. (2) 2-(furan-2-yl)pyrimidine. (3) Phenyl(piperazin-1-yl)methanone. (4) Phorbol-like tetracyclic core. (5) 5H-[1,3]dioxolo[4,5-g]quinolin-8-one. (6) 1,2-Epoxy-1,2,3,4-tetrahydrobenz(c)acridine. (7) Phenethyl-substituted 1,2,3-oxadiazolium. (8) 1-[(E)-furan-2-ylmethylideneamino]imidazolidin-2-one. (9) Azobicyclo[3.3.0]octane-fused macrocyclic bis lactone. (10) Estrane-based steroidal core with dioxolane moiety.

Unlike the complex and multiring architectures prevalent in carcinogenic compounds, the predominant scaffolds among noncarcinogens tend to exhibit reduced structural complexity, often consisting of a single aromatic unit (e.g., benzene) or a spirocyclic framework such as fluoran (spiro[2-benzofuran-3,9'-xanthene]-1-one).

3.10. Scaffold-Level Profiling of Model Predictions Via BiLSTM-Driven SHAP Analysis. Enhancing the interpretability of model predictions, a scaffold-level analysis based on BiLSTM-derived SHAP values was performed. This approach revealed core structural scaffolds that play a pivotal role in compound classification, offering mechanistic insights into the model's decision-making process. Among the 601 Bemis–Murcko scaffolds identified from the training set, SHAP analysis revealed that only a subset exhibited meaningful contributions to model predictions. Specifically, 113 scaffolds exhibited positive SHAP values, suggesting a strong association with carcinogenic predictions, whereas 137 scaffolds showed negative contributions, indicating relevance to noncarcinogenic classifications. The remaining 351 scaffolds exerted minimal influence on the model's outputs and were therefore excluded from further interpretation.

Figure 15 illustrates the top 10 scaffolds with the highest positive SHAP values alongside their corresponding chemical structures. Notably, 4-(2-thiophen-2-yl quinazolin-4-yl)-morpholine emerged as the most influential scaffold driving carcinogenic predictions. Conversely, (1Z)-1-benzylidenein-

dene exerted the strongest negative SHAP effect, aligning with noncarcinogenic profiles; the top 10 negatively contributing scaffolds are depicted in Figure 16.

Subsequently, we selected representative compounds associated with two scaffolds exhibiting the highest absolute SHAP values from both the positive and negative impact groups. From the positively contributing scaffold, 4-(2-thiophen-2-yl quinazolin-4-yl)-morpholine, we identified the compound 4-[2-(5-nitrothiophen-2-yl)quinazolin-4-yl]morpholine, which is experimentally labeled as a carcinogen. Conversely, the negatively contributing scaffold (1Z)-1-benzylideneindene led to the identification of Sulindac sulfone, a compound known to be noncarcinogenic. These findings underscore the model's ability to highlight crucial scaffolds that are strongly linked to ground-truth labels. Notably, the scaffold 4-(2-thiophen-2-yl quinazolin-4-yl)morpholine also contains the substructure MACCS156 (NA(A)A), previously identified as a key carcinogenic motif in SHAP-based substructure analysis. Figure 17 presents the representative compounds harboring scaffolds with the most prominent SHAP contributions in each category. In aggregate, these findings support the robustness and interpretability of the proposed framework.

3.11. Bemis–Murcko Scaffolds of the Top Four Potential Carcinogens Identified from IARC Class 2B Data Set. In this analysis, we extracted the Bemis–Murcko scaffolds corresponding to the four top-ranked compounds from the IARC class 2B data set, which are classified as potential

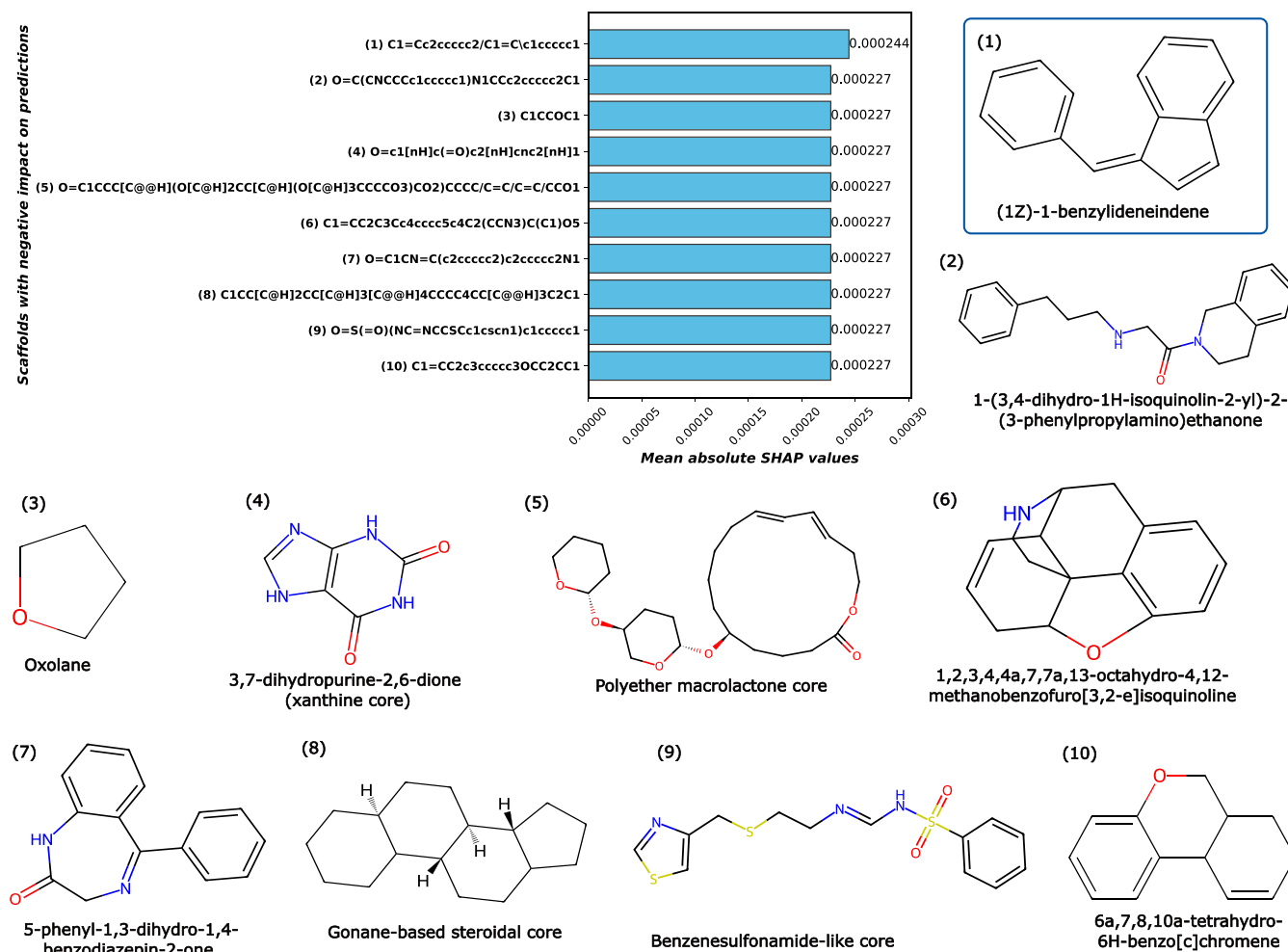


Figure 16. Ten key scaffolds with a negative impact on the prediction as revealed by SHAP analysis. (1) (1Z)-1-benzylideneindene. (2) 1-(3,4-dihydro-1H-isoquinolin-2-yl)-2-(3-phenylpropylamino)ethanone. (3) Oxolane. (4) 3,7-dihydropurine-2,6-dione. (5) Polyether macrolactone core. (6) 1,2,3,4,4a,7,7a,13-octahydro-4,12-methanobenzofuro[3,2-e]isoquinoline. (7) 5-phenyl-1,3-dihydro-1,4-benzodiazepin-2-one. (8) Gonane-based steroidal core. (9) Benzenesulfonamide-like core. (10) 6a,7,8,10a-tetrahydro-6H-benzo[c]chromene.

carcinogens. As previously noted, these compounds were consistently predicted as carcinogenic across all individual models, each exhibiting mean predicted probabilities above 0.9, highlighting the strong agreement within the consensus framework. The four resulting scaffolds, representing the core structural frameworks underlying these high-confidence predictions, are depicted in Figure 18.

Importantly, the identified scaffolds contain critical substructural motifs revealed by SHAP interpretation, such as MACCS125 (aromatic ring count >1), MACCS52 (N–N bond), and MACCS156 (NA(A)A), which have been recurrently linked to carcinogenic activity. Notably, (8S)-2,3,5,8-tetrahydro-1H-pyrrolizine and phenyl(pyridin-3-yl)-diazene correspond to positively contributing scaffolds observed in the training data, reinforcing the consistency of our structural interpretation. In contrast, 9H-pyrido[2,3-b]indole and benzo-[j]aceanthrylene were not captured among the positively contributing scaffolds, suggesting potential scaffold-level novelty or underrepresentation in the training distribution.

4. DISCUSSION

Based on the predictive outcomes summarized in Table 5, we further performed a chemical classification of 16 potential

carcinogens from the IARC class 2B, as identified by our consensus framework. The resulting distribution of chemical classes and representative compounds is presented in Table 11. The compounds are ordered from top to bottom according to decreasing carcinogenic risk. Accordingly, we provided critical insights and warnings regarding their potential carcinogenic properties in real-world scenarios.

Lasiocarpine (No. 1 in Table 11) emerged as the top-ranked compound in our consensus prediction, exhibiting both the highest mean predicted probability and unanimous agreement across all individual models in its classification as a carcinogen. This naturally occurring pyrrolizidine alkaloid (PA) is well-documented for its potent hepatotoxicity.⁴³ In previous research, lasiocarpine has been reported to induce liver and skin tumors in rat models.⁴⁴ Recent toxicological data indicate that, among structurally related PAs, lasiocarpine exhibits the highest level of toxicity, as evidenced by its lowest benchmark dose lower confidence limit for a 10% response (BMDL10) value of 0.07 mg/kg bw/day.⁴⁵ Noted that the BMDL10 is a toxicology and risk assessment statistical value representing the lower 95% confidence limit of a benchmark dose (BMD) associated with a 10% increase in a specific adverse effect.⁴⁶ This finding highlights lasiocarpine as the most hazardous PA among

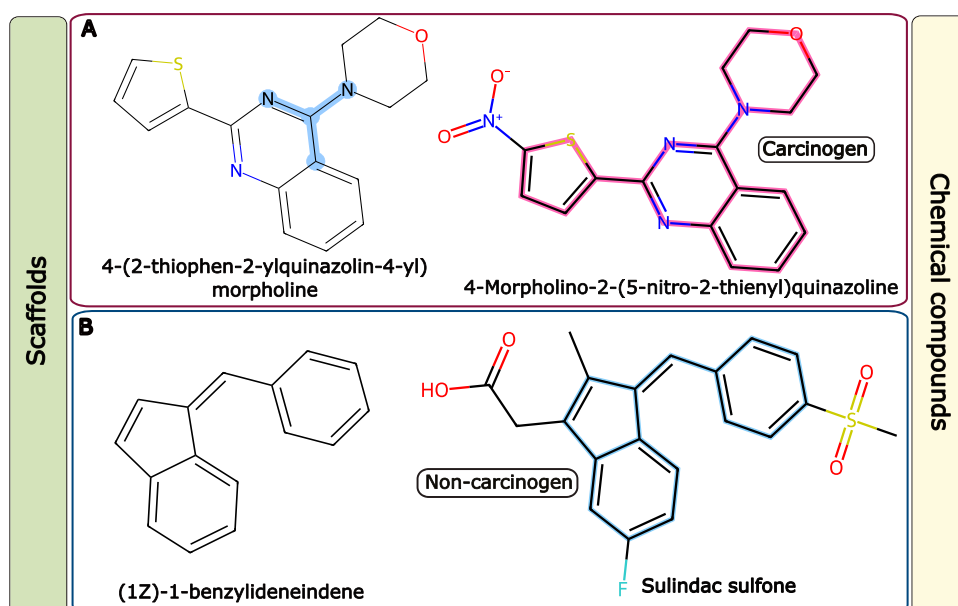


Figure 17. Representative compounds containing scaffolds with the highest SHAP values in each group. (A) Carcinogenic compound featuring a scaffold containing the MACCS156 substructure (NA(A)A, highlighted in light blue) and (B) Noncarcinogenic compound with a negatively contributing scaffold.

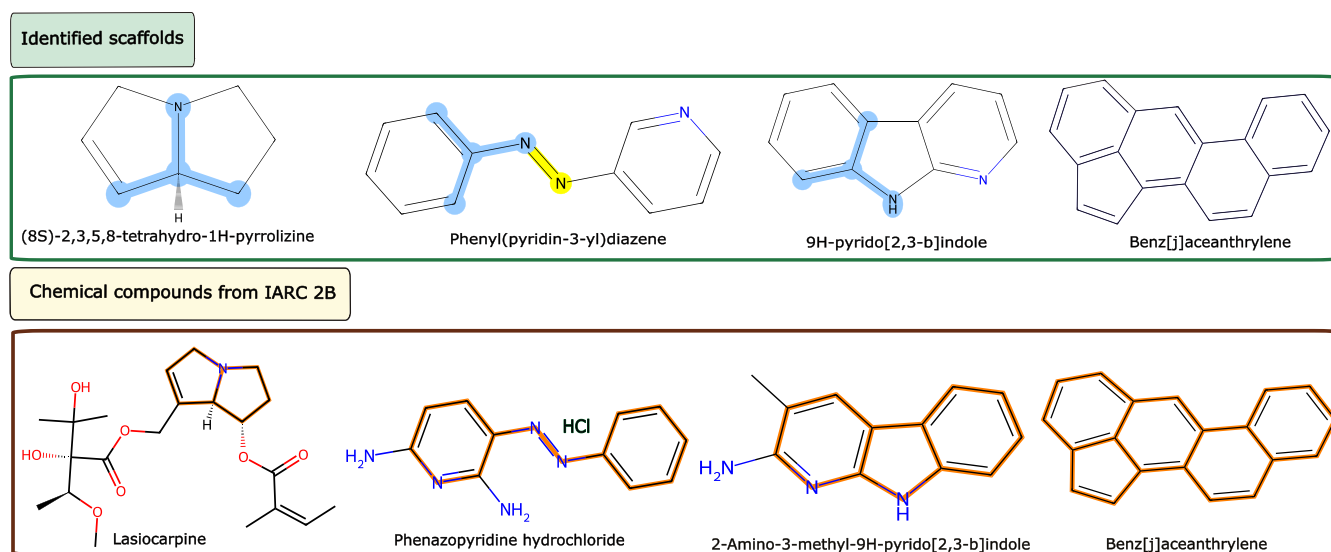


Figure 18. Bemis–Murcko scaffolds of the four most probable carcinogens predicted from the IARC class 2B data set, with MACCS52 (NN, yellow) and MACCS156 (NA(A)A, light blue) substructures highlighted.

those with available tumorigenicity data. In general, pyrrolizidine alkaloids are well recognized for their toxic potential and their capacity to pose significant risks to both human and animal health, particularly when consumed chronically at high levels.⁴⁷ Notably, human health concerns associated with dietary exposure to PAs, especially through contaminated honey, tea, herbal infusions, and food supplements, have been increasingly documented and emphasized in recent risk assessments.⁴⁸

Subsequently, 2-Amino-3-methyl-9H-pyrido[2,3-b]indole (No. 3 in Table 11), a representative heterocyclic aromatic amine (HAA), was identified as one of the carcinogenic compounds predicted by our model. HAAs are a class of chemical compounds formed during the cooking of meat at high temperatures, such as grilling, broiling, pan-frying, or barbecuing. Several epidemiological studies have reported a positive

association between the consumption of well-done or high-temperature cooked meat and the risk of cancers, particularly colorectal cancer.^{49–51} However, conflicting findings have also been observed, with some studies reporting no significant association.⁵² A major limitation contributing to this inconsistency is the difficulty in accurately assessing dietary HAA exposure. Due to the role of metabolic activation in HAA induced carcinogenesis, recent studies have examined genetically susceptible individuals, particularly those with high activity CYP1A2 and/or N-acetyltransferase 2 (NAT2) phenotypes. While no interaction between red meat intake and NAT2 genotype was found in Whites, a significant gene diet interaction was observed in Japanese and African Americans, populations with higher colorectal cancer rates and a greater prevalence of the rapid NAT2 phenotype. In these groups, red meat

Table 11. Chemical Classification of the 16 Predicted Carcinogens of IARC Class 2B from the Consensus Framework

no.	agent	chemical classification
1	lasiocarpine	pyrrolizidine alkaloid
2	phenazopyridine hydrochloride	azo-containing aromatic pyridine derivative
3	2-amino-3-methyl-9H-pyrido[2,3-b]indole	heterocyclic aromatic amine (pyridoindole-based derivative)
4	benz[<i>j</i>]aceanthrylene	polycyclic aromatic hydrocarbon
5	phenoxybenzamine hydrochloride	haloalkylamine derivative
6	1,2-dichloroethane	halogenated hydrocarbon
7	sterigmatocystin	xanthone derivative
8	1,3-dichloropropene	halogenated hydrocarbon
9	4-methylimidazole	imidazole derivative
10	chloroprene	halogenated hydrocarbon
11	bromodichloromethane	halogenated hydrocarbon
12	benzo[<i>c</i>]phenanthrene	polycyclic aromatic hydrocarbon
13	vinylidene chloride	halogenated hydrocarbon
14	1,1,1,2-tetrachloroethane	halogenated hydrocarbon
15	chloroform	halogenated hydrocarbon
16	1-bromopropane	halogenated hydrocarbon

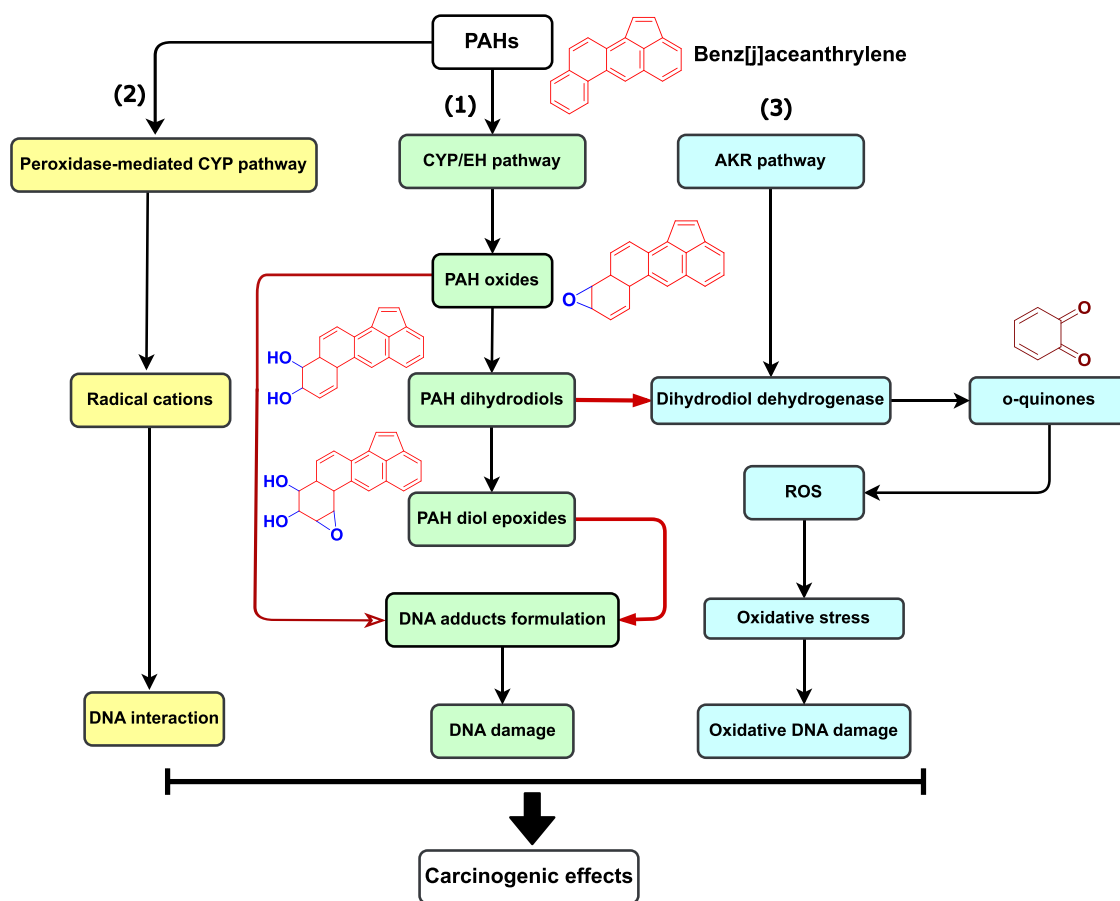


Figure 19. Diverse biological pathways responsible for the carcinogenic potential of PAHs. (1) the cytochrome P450 1A1/1B1 with epoxide hydrolase pathway (CYP/EH), (2) the peroxidase-mediated CYP pathway, and (3) the aldo-keto reductase (AKR) pathway.

consumption was most strongly linked to cancer risk among rapid acetylators, less so in intermediates, and not significant in slow acetylators.⁵³ In light of emerging evidence on the carcinogenic potential of HAAs in certain population groups combined with the warnings of HAA-related compounds by our model, these findings may help raise awareness of HAAs as possible dietary risk factors. These findings may serve as a helpful reference for public health efforts, supporting early recognition of potential dietary risks associated with HAAs.

Polycyclic aromatic hydrocarbons (PAHs) represent a class of organic compounds characterized by the presence of two or more fused aromatic rings, typically originating from benzene structures.⁵⁴ Recognized among the earliest identified environmental carcinogens,⁵⁵ many PAHs exhibit strong carcinogenic properties, either directly or following metabolic activation. Due to their widespread occurrence and persistence, PAHs pose significant public health concerns, having been associated with a range of adverse effects, including carcinogenesis as well as

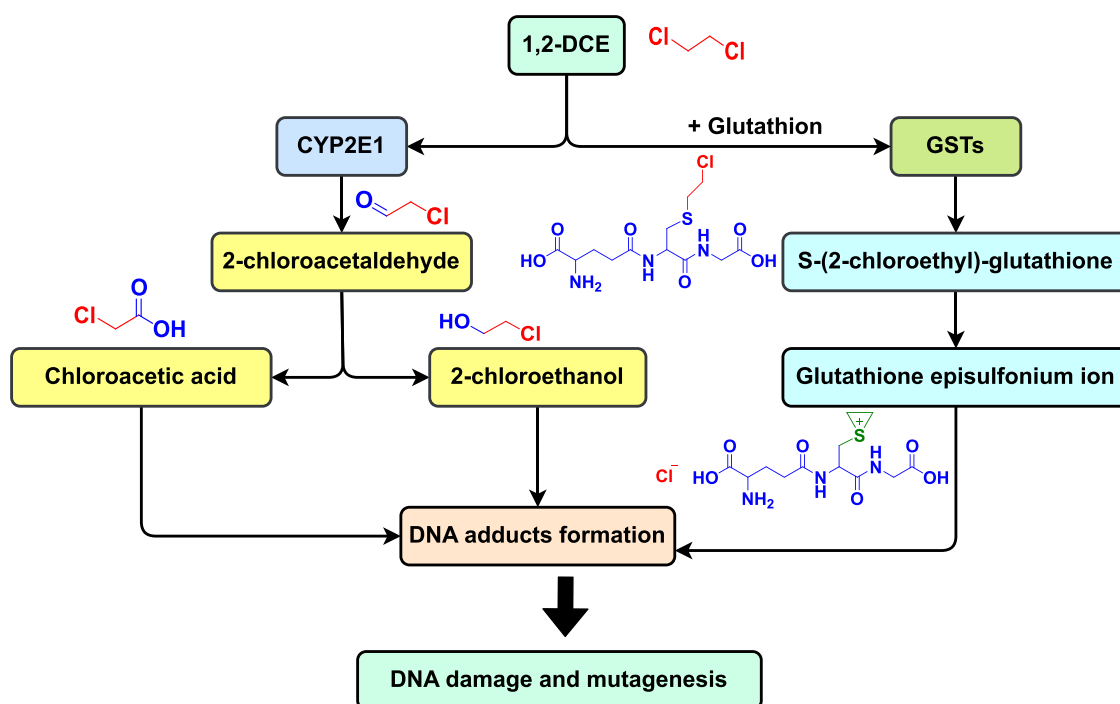


Figure 20. Carcinogenic mechanism of 1,2-DCE. CYP2E1-mediated metabolism produces reactive metabolites (chloroacetic acid and 2-chloroethanol), while GST conjugation generates episulfonium ions. Both pathways induce DNA adducts and damage, driving mutagenesis and thereby promoting carcinogenesis.

cardiovascular, pulmonary, and reproductive toxicity.^{55,56} Notably, several PAHs function as procarcinogens, requiring enzymatic biotransformation within the body to yield active carcinogenic metabolites.⁵⁷ A recent meta-analysis evaluating the carcinogenic effects of PAHs reported a 10.75% increased overall cancer risk and an 8.40% higher incidence of breast cancer in individuals exposed to PAHs compared to nonexposed groups.⁵⁸ These findings further emphasize the importance of raising awareness about the potential health risks associated with PAH exposure. Specifically, benz[*j*]aceanthrylene (No. 4 in Table 11) and benzo[*c*]phenanthrene (No. 12 in Table 11), two representative PAHs, were identified by our framework as carcinogenic compounds.

PAHs impact various biological systems, from the respiratory to the endocrine, via multiple pathways. These include binding to intracellular targets, altering gene transcription, and interfering with hormonal signaling cascades.⁵⁹ Due to their lipophilic nature, PAHs readily permeate cell membranes through passive diffusion, particularly following inhalation. While the parent PAH molecules are not directly genotoxic, they act as procarcinogens; their metabolic derivatives are responsible for initiating carcinogenesis.⁴⁰ In hepatic metabolism, PAHs are primarily processed through three enzymatic routes: (1) the cytochrome P450 1A1/1B1 with epoxide hydrolase pathway (CYP/EH), (2) the peroxidase-mediated CYP pathway, and (3) the aldo-keto reductase (AKR) pathway.^{40,59} These pathways successively convert PAHs into reactive intermediates. Among these, the CYP/EH pathway is the most prominent, in which PAHs undergo oxidation to form PAH oxides that are subsequently converted into dihydrodiols. These intermediates are further enzymatically activated to generate diol epoxides. Both oxides and diol epoxides are capable of forming covalent adducts with DNA, thereby exerting carcinogenic effects. If this DNA damage is not efficiently

repaired, it may initiate mutagenic processes that contribute to tumorigenesis.⁶⁰ The second metabolic route involves cytochrome P450 enzymes with peroxidase-like activity, which oxidize PAHs to form highly reactive, short-lived radical cations. These electrophilic intermediates readily interact with DNA, thereby inducing mutagenic lesions. In the third pathway, dihydrodiol dehydrogenase (a member of the aldo-keto reductase superfamily) catalyzes the conversion of PAH dihydrodiols to o-quinones. Subsequent redox cycling of these quinones generates reactive oxygen species (ROS), which exacerbate oxidative stress and promote carcinogenesis through oxidative DNA damage.⁴⁰ The multifaceted mechanisms through which PAHs exert their carcinogenic effects are summarized in Figure 19, using benz[*j*]aceanthrylene as an example.⁶¹

Additionally, our predictive framework identified eight of the 16 compounds with carcinogenic potential as halogenated hydrocarbons, a chemical class historically linked to elevated cancer risks and adverse health effects upon chronic exposure.⁶² The aliphatic or olefinic halogenated hydrocarbons identified in this study are characterized by the presence of chlorine and bromine substituents. Within this group, 1,2-dichloroethane (1,2-DCE) (No. 6 in Table 11) emerged as the compound with the highest predicted carcinogenic risk. 1,2-DCE is a high-volume industrial solvent primarily used in vinyl chloride production and is frequently detected in both air and water environments. Extensive research has demonstrated its genotoxicity across diverse test systems and exposure routes, with effects mediated through metabolic activation by key enzymes, particularly P450 2E1 (CYP2E1) and glutathione S-transferases (GSTs), leading to DNA adduct formation, gene mutations, and chromosomal aberrations. Collectively, these findings indicate that 1,2-DCE induces genetic damage through enzyme-dependent bioactivation pathways, despite negative

outcomes in certain clastogenicity assays.^{63,64} Proposed carcinogenic mechanism of 1,2-DCE is illustrated in Figure 20. Furthermore, another industrial solvent, chloroform (No. 15 in Table 11), was also identified within the predicted subset of carcinogens. Chloroform is widely used in organic extraction and purification processes and has long been under regulatory scrutiny due to its adverse health effects.⁶⁵ The accurate identification of such well-established hazardous compounds underscores both the robustness and the practical relevance of our model, highlighting the continued need for vigilance regarding carcinogenic risks posed by commonly used industrial solvents.

Mounting evidence highlights the profound and long-lasting health consequences of early life exposure to carcinogenic substances, which substantially elevates cancer risk in later stages of life. During childhood and adolescence, the heightened cellular proliferation and ongoing organogenesis render individuals particularly susceptible to the genotoxic effects of carcinogens. Among these, PAHs are a ubiquitous class of well-characterized carcinogens to which young populations are routinely exposed through contaminated food, drinking water, ambient air, and various environmental media.⁶⁶ In our model, benz[*j*]aceanthrylene and benzo[*c*]phenanthrene, both members of the PAH family, were confidently identified as carcinogenic compounds with high predictive scores. Furthermore, cumulative risk assessment conducted in 2018 for the occurrence of contaminants in United States drinking water (2010–2017) estimated that over 100,000 lifetime cancer cases could be attributable to chronic exposure to carcinogenic chemicals in tap water.⁶⁷ Notably, chloroform, a disinfection byproduct, was identified as one of the primary contributors. Based on risk thresholds defined by the California Office of Environmental Health Hazard Assessment, a chloroform concentration of 0.4 $\mu\text{g/L}$ in drinking water corresponds to a 10^{-6} lifetime cancer risk. With an estimated exposed population of 203 million individuals, this translates to approximately 8,687 lifetime cancer cases. As corroborated in our model, chloroform was classified as a high-confidence carcinogenic compound, reinforcing its public health relevance. These findings underscore the urgent need to identify and characterize potential carcinogens with high correctness. Scientifically robust and practically applicable predictive tools are essential to support regulatory decision-making aimed at alerting vulnerable populations to carcinogens.

The present study systematically addresses the OECD principles in the context of carcinogenicity prediction. In line with OECD Principle 1, carcinogenic potential is defined as the end point of interest, reflecting the urgent need to address ongoing cancer risks and to provide complementary evidence for compounds that remain controversial within the IARC Group 2B classification. Consistent with OECD Principle 2, the framework leverages ML and DL approaches, which have demonstrated strong capacity in toxicity modeling and in capturing QSAR relationships,⁶⁸ thereby providing a scientifically sound rationale for linking chemical structure to predicted toxicity. To satisfy OECD Principle 3, the AD is explicitly delineated using Euclidean distance-based *k*-nearest neighbor (*k*-NN) methods, enabling robust confidence assessment of predictions. In accordance with OECD Principle 4, model performance is rigorously evaluated through a comprehensive suite of statistical metrics, including accuracy, MCC, AUROC, sensitivity, and specificity. Finally, addressing OECD Principle 5, cheminformatics profiling is incorporated to probe structural

associations between carcinogenicity and chemical features, with particular emphasis on substructures and scaffolds, thereby enhancing transparency and interpretability. As a whole, this alignment with OECD principles underscores the scientific robustness of our predictive framework and highlights its potential regulatory relevance.

The models developed in this study were primarily designed to predict and classify carcinogenicity using molecular descriptors and structural features, showing strong promise as practical tools for compound screening and prioritization. As such, they can help identify and compare chemical compounds, enabling the early elimination of candidates with high carcinogenic potential and indirectly guiding the design of safer analogs. In addition, the cheminformatics analyses presented here provide detailed insights into substructures and scaffolds that distinguish carcinogens from noncarcinogens, offering valuable direction for optimizing molecular design toward safer compounds. Nonetheless, the *de novo* design of low-toxicity compounds remains a complex challenge that requires multiobjective optimization and experimental validation. As such, our model can be served as a supportive tool within the broader context of drug discovery and safety evaluation. Overall, these findings underscore the potential of integrating predictive modeling with cheminformatics to accelerate the development of safer compounds and improve early stage safety assessment in drug discovery.

4.1. Advantages and Challenges. In this study, a consensus-based framework was constructed by integrating three independently trained models: LightGBM utilizing RDKit molecular fingerprints, RF employing E-state descriptors, and a BiLSTM network operating on MACCS keys. To improve the reliability of carcinogenicity predictions, a probabilistic voting strategy was employed, in which a compound was classified as carcinogenic only when all three models yielded predicted probabilities greater than 0.5. This stringent decision criterion was designed to enhance confidence in the outcomes by requiring agreement across models that capture complementary aspects of molecular structure, including topological connectivity, physicochemical properties, and substructural sequence patterns. By enforcing agreement among all predictive models, the framework effectively reduces the risk of false-positive classifications that may arise from biases or overfitting within any single model. This conservative strategy enhances the overall robustness and interpretability of the predictions, particularly when applied to structurally heterogeneous compounds such as those in the IARC class 2B data set. As a result, the proposed framework serves as a reliable decision-support tool for identifying and prioritizing potential carcinogens with increased predictive certainty.

A limitation of this study arises from the overlap between certain chemicals in the IARC class 2B and 3 classifications and those included in the CPDB training data set. To ensure the integrity of model evaluation and prevent bias due to prior exposure during training, these overlapping compounds were excluded from the IARC-based test set during data preprocessing. As a result, although not present in the final evaluation phase, these chemicals did contribute to model learning through their inclusion in the training set based on animal carcinogenicity data. For example, aspartame was used as a training compound labeled as noncarcinogenic in animals, despite ongoing scientific debate and insufficient evidence regarding its carcinogenicity in humans. The exclusion of such

compounds, while necessary for methodological rigor, reduced the number of chemicals available for IARC-based assessment.

4.2. Future Perspectives. Although the proposed consensus framework demonstrated promising performance in carcinogenicity prediction, future research could further enhance its predictive capability and generalizability by incorporating additional molecular representations, such as graph-based neural networks or quantum chemical descriptors, given their growing potential to capture intricate structural and electronic features relevant to chemical toxicity.^{69–71} Furthermore, expanding the training set with well-curated human carcinogenicity data, when available, would also improve relevance to real-world risk assessment. Finally, the application of this approach to broader chemical spaces, such as food additives, environmental contaminants, and pharmaceuticals, could facilitate high-throughput screening and prioritization in regulatory toxicology.

5. CONCLUSIONS

In this study, we developed a robust consensus framework for the prediction of chemical carcinogenicity, with a particular emphasis on compounds classified as possibly carcinogenic to humans (IARC class 2B). Three complementary ML and DL models, including BiLSTM, LightGBM, and RF, were integrated following their training on benchmark data sets. The application of this framework to the IARC 2B compounds enabled meaningful stratification into potential carcinogens, presumed noncarcinogens, and inconclusive cases, offering a data-driven basis for prioritizing further toxicological investigation. Furthermore, SHAP-based interpretability analysis revealed chemically intuitive and mechanistically relevant features, including MACCS substructures and Bemis–Murcko scaffolds, that substantially contributed to model predictions. To facilitate practical adoption and future research, a publicly accessible web server was deployed, providing an interactive platform for carcinogenicity prediction. In summary, these findings underscore the potential of interpretable ML to support early hazard identification and inform regulatory decision-making in chemical safety assessment.

■ ASSOCIATED CONTENT

Data Availability Statement

All data used in this study are provided in the [Supporting Information](#). Source code utilized in this study are openly accessible at <https://github.com/taraponglab/carcinogen-consensus-model>. A freely accessible web server for carcinogen prediction is available at <https://carcinogenicity-predictor.streamlit.app>

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.5c01873>.

The training data set, external test data sets, complete performance metrics of all predictive models, three subsets of outcomes identified from the IARC 2B data set (47 within-domain compounds), and full SHAP values of all MACCS keys fingerprints (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Tarapong Srisongkram — Division of Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Khon Kaen

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.5c01873>

Author Contributions

Conceptualization, T.S.; methodology, H.A.D. and T.S.; data curation, T.S.; 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

H.A.D. acknowledges the financial support received from the KKU Scholarship for ASEAN and GMS Countries Personnel, provided by the Faculty of Pharmaceutical Sciences, Khon Kaen University, Thailand.

■ REFERENCES

- (1) Siegel, R. L.; Giaquinto, A. N.; Jemal, A. Cancer statistics. *Ca-Cancer J. Clin.* **2024**, *74*, 12–49.
- (2) Madia, F.; Worth, A.; Whelan, M.; Corvi, R. Carcinogenicity assessment: Addressing the challenges of cancer and chemicals in the environment. *Environ. Int.* **2019**, *128*, 417–429.
- (3) ECHA. *Guidance on Information Requirements and Chemical Safety Assessment: Chapter R. 6: QSARs and Grouping of Chemicals*, Helsinki, Finland; 2008.
- (4) OECD. *Guidance Document on the Validation of (Quantitative) Structure-activity Relationship [(Q) SAR] Models*, Organisation for Economic Co-operation and Development 2014.
- (5) Stepišnik, T.; Škrlj, B.; Wicker, J.; Kocev, D. A comprehensive comparison of molecular feature representations for use in predictive modeling. *Comput. Biol. Med.* **2021**, *130*, No. 104197.
- (6) Srisongkram, T. Ensemble quantitative read-across structure-activity relationship algorithm for predicting skin cytotoxicity. *Chem. Res. Toxicol.* **2023**, *36*, 1961–1972.
- (7) Srisongkram, T.; Syahid, N. F.; Tookkane, D.; Weerapreeyakul, N.; Puthongking, P. Stacked ensemble learning on HaCaT cytotoxicity for skin irritation prediction: A case study on dipterocarpol. *Food Chem. Toxicol.* **2023**, *181*, No. 114115.
- (8) 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*, No. 109480.
- (9) Zetta, D. N.; Srisongkram, T. Stacking Ensemble Neural Network for Chemical Safety Assessment: A Case Study of Thyroid Peroxidase and Natural Product Screening. *ACS Omega* **2025**, *10*, 30450–30466, DOI: [10.1021/acsomega.5c02188](https://doi.org/10.1021/acsomega.5c02188).
- (10) Srisongkram, T. DeepRA: A novel deep learning-read-across framework and its application in non-sugar sweeteners mutagenicity prediction. *Comput. Biol. Med.* **2024**, *178*, No. 108731.
- (11) Zhang, L.; Ai, H.; Chen, W.; Yin, Z.; Hu, H.; Zhu, J.; Zhao, J.; Zhao, Q.; Liu, H. CarcinoPred-EL: Novel models for predicting the

carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. *Sci. Rep.* **2017**, *7*, No. 2118.

(12) Chen, Z.; Zhang, L.; Sun, J.; Meng, R.; Yin, S.; Zhao, Q. DCAMCP: A deep learning model based on capsule network and attention mechanism for molecular carcinogenicity prediction. *J. Cell. Mol. Med.* **2023**, *27*, 3117–3126.

(13) Wang, Y.-W.; Huang, L.; Jiang, S.-W.; Li, K.; Zou, J.; Yang, S.-Y. CapsCarcino: A novel sparse data deep learning tool for predicting carcinogens. *Food Chem. Toxicol.* **2020**, *135*, No. 110921.

(14) Rusyn, I.; Wright, F. A. Ten years of using key characteristics of human carcinogens to organize and evaluate mechanistic evidence in IARC Monographs on the identification of carcinogenic hazards to humans: Patterns and associations. *Toxicol. Sci.* **2024**, *198*, 141–154.

(15) Hayes, A. W.; Kobets, T. *Hayes' Principles and Methods of Toxicology*; CRC Press, 2023.

(16) Goodman, J. E.; Boon, D. N.; Jack, M. M. Perspectives on recent reviews of aspartame cancer epidemiology. *Global Epidemiol.* **2023**, *6*, No. 100117.

(17) National Center for Biotechnology Information. Compound Summary for CID 441140, Griseofulvin 2025 <https://pubchem.ncbi.nlm.nih.gov/compound/Griseofulvin>. (Accessed July 25, 2025).

(18) IARC Working Group on the Evaluation of Carcinogenic Risks to Humans Some Drugs and Herbal Products. 2016.

(19) Barber, C.; Heghes, C.; Johnston, L. A framework to support the application of the OECD guidance documents on (Q) SAR model validation and prediction assessment for regulatory decisions. *Comput. Toxicol.* **2024**, *30*, No. 100305.

(20) OECD. *QSAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions* OECD Publishing: Paris; 2023.

(21) Duy, H. A.; Srisongkram, T. Comparative analysis of recurrent neural networks with conjoint fingerprints for skin corrosion prediction. *J. Chem. Inf. Model.* **2025**, *65*, 1305–1317.

(22) Duy, H. A.; Srisongkram, T. Bidirectional Long Short-Term Memory (BiLSTM) Neural Networks with Conjoint Fingerprints: Application in Predicting Skin-Sensitizing Agents in Natural Compounds. *J. Chem. Inf. Model.* **2025**, *65*, 3035–3047.

(23) Duy, H. A.; Srisongkram, T. Protecting your skin: a highly accurate LSTM network integrating conjoint features for predicting chemical-induced skin irritation. *J. Cheminf.* **2025**, *17*, No. 39.

(24) Gold, L. S.; Manley, N. B.; Slone, T. H.; Rohrbach, L.; Garfinkel, G. B. Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature through 1997 and by the National Toxicology Program in 1997–1998. *Toxicol. Sci.* **2005**, *85*, 747–808.

(25) Benigni, R.; Bossa, C.; Richard, A. M.; Yang, C. A novel approach: chemical relational databases, and the role of the ISSCAN database on assessing chemical carcinogenicity. *Ann. Ist. Super. Sanita* **2008**, *44*, 48–56.

(26) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; et al. PubChem substance and compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213.

(27) Dharmaratne, H. R. W.; Nanayakkara, N. D.; Khan, I. A. Kavalactones from Piper methysticum, and their ¹³C NMR spectroscopic analyses. *Phytochemistry* **2002**, *59*, 429–433.

(28) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. 2014, arXiv:1412.6980v9. arXiv.org e-Print archive. <https://doi.org/10.48550/arXiv.1412.6980>.

(29) Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, *21*, No. 6.

(30) Richardson, E.; Trevizani, R.; Greenbaum, J. A.; Carter, H.; Nielsen, M.; Peters, B. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns* **2024**, *5*, No. 100994, DOI: 10.1016/j.patter.2024.100994.

(31) Srisongkram, T.; Tookkane, D. Insights into the structure-activity relationship of pyrimidine-sulfonamide analogues for targeting BRAF V600E protein. *Biophys. Chem.* **2024**, *307*, No. 107179.

(32) Duy, H. A.; Srisongkram, T. A hybrid framework of generative deep learning for antiviral peptide discovery. *Sci. Rep.* **2025**, *15*, No. 25554.

(33) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(34) Danishuddin; Haque, M. A.; Madhukar, G.; Jamal, Q. M. S.; Kim, J.-J.; Ahmad, K. Machine Learning-Driven Consensus Modeling for Activity Ranking and Chemical Landscape Analysis of HIV-1 Inhibitors. *Pharmaceuticals* **2025**, *18*, No. 714.

(35) Salih, A. M.; Raisi-Estabragh, Z.; Galazzo, I. B.; Radeva, P.; Petersen, S. E.; Lekadir, K.; Menegaz, G. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv. Intell. Syst.* **2025**, *7*, No. 2400304.

(36) Li, J. Area under the ROC Curve has the most consistent evaluation for binary classification. *PLoS One* **2024**, *19*, No. e0316019.

(37) Muschelli, J., III ROC and AUC with a binary predictor: a potentially misleading metric. *J. Classif.* **2020**, *37*, 696–708.

(38) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.

(39) White, N.; Parsons, R.; Collins, G.; Barnett, A. Evidence of questionable research practices in clinical prediction models. *BMC Med.* **2023**, *21*, No. 339.

(40) Moorthy, B.; Chu, C.; Carlin, D. J. Polycyclic aromatic hydrocarbons: from metabolism to lung cancer. *Toxicol. Sci.* **2015**, *145*, 5–15.

(41) Grimmer, G. Environmental Carcinogens. In *Polycyclic Aromatic Hydrocarbons*; CRC Press, 2018.

(42) Lawal, A. T. Polycyclic aromatic hydrocarbons. A review. *Cogent Environ. Sci.* **2017**, *3*, No. 1339841.

(43) Fashe, M. M.; Juvonen, R. O.; Petsalo, A.; Rasanen, J.; Pasanen, M. Species-specific differences in the in vitro metabolism of lasiocarpine. *Chem. Res. Toxicol.* **2015**, *28*, 2034–2044.

(44) Svoboda, D. J.; Reddy, J. K. Malignant tumors in rats given lasiocarpine. *Cancer Res.* **1972**, *32*, 908–913.

(45) Chen, L.; Mulder, P. P.; Louise, J.; Peijnenburg, A.; Wesseling, S.; Rietjens, I. M. C. M. Risk assessment for pyrrolizidine alkaloids detected in (herbal) teas and plant food supplements. *Regul. Toxicol. Pharmacol.* **2017**, *86*, 292–302.

(46) More, S. J.; Bampidis, V.; Benford, D.; Bragard, C.; Halldorsson, T. I.; Hernández-Jerez, A. F.; Bennekou, S. H.; Koutsoumanis, K.; Lambré, C.; EFSA Scientific Committee; et al. Guidance on the use of the benchmark dose approach in risk assessment. *EFSA J.* **2022**, *20*, No. e07584.

(47) Tábuas, B.; Barros, S. C.; Diogo, C.; Cavaleiro, C.; Silva, A. S. Pyrrolizidine alkaloids in foods, herbal drugs, and food supplements: chemistry, metabolism, toxicological significance, analytical methods, occurrence, and challenges for future. *Toxins* **2024**, *16*, No. 79.

(48) Knutsen, H. K.; Alexander, J.; Barregård, L.; Bignami, M.; Brüschweiler, B.; Ceccatelli, S.; Cottrill, B.; Dinovi, M.; Edler, L.; EFSA Panel on Contaminants in the Food Chain (CONTAM); et al. Risks for human health related to the presence of pyrrolizidine alkaloids in honey, tea, herbal infusions and food supplements. *EFSA J.* **2017**, *15*, No. e04908.

(49) Rohrmann, S.; Hermann, S.; Linseisen, J. Heterocyclic aromatic amine intake increases colorectal adenoma risk: findings from a prospective European cohort study. *Am. J. Clin. Nutr.* **2009**, *89*, 1418–1424.

(50) Turesky, R. J.; Le Marchand, L. Metabolism and biomarkers of heterocyclic aromatic amines in molecular epidemiology studies: lessons learned from aromatic amines. *Chem. Res. Toxicol.* **2011**, *24*, 1169–1214.

(51) Zheng, W.; Lee, S.-A. Well-done meat intake, heterocyclic amine exposure, and cancer risk. *Nutr. Cancer* **2009**, *61*, 437–446.

(52) Iwasaki, M.; Tsugane, S. Dietary heterocyclic aromatic amine intake and cancer risk: epidemiological evidence from Japanese studies. *Genes and Environment* **2021**, *43*, No. 33.

(53) Le Marchand, L. The role of heterocyclic aromatic amines in colorectal cancer: the evidence from epidemiologic studies. *Genes Environ.* **2021**, *43*, 20.

- (54) Kim, K.-H.; Jahan, S. A.; Kabir, E.; Brown, R. J. A review of airborne polycyclic aromatic hydrocarbons (PAHs) and their human health effects. *Environ. Int.* **2013**, *60*, 71–80.
- (55) Bai, L.; Geng, X.; Liu, X. Review of polycyclic aromatic hydrocarbons pollution characteristics and carcinogenic risk assessment in global cooking environments. *Environ. Pollut.* **2024**, *361*, No. 124816.
- (56) Baird, W. M.; Hooven, L. A.; Mahadevan, B. Carcinogenic polycyclic aromatic hydrocarbon-DNA adducts and mechanism of action. *Environ. Mol. Mutagen.* **2005**, *45*, 106–114.
- (57) Cheng, T.; Gamage, S. M. K.; Lu, C.-T.; Aktar, S.; Gopalan, V.; Lam, A. K.-y. Polymorphisms in PAH metabolising enzyme CYP1A1 in colorectal cancer and their clinicopathological correlations. *Pathol., Res. Pract.* **2022**, *231*, No. 153801.
- (58) Arif, I.; Adams, M. D.; Johnson, M. T. A meta-analysis of the carcinogenic effects of particulate matter and polycyclic aromatic hydrocarbons. *Environ. Pollut.* **2024**, *351*, No. 123941.
- (59) Sombiri, S.; Balhara, N.; Attri, D.; Kharb, I.; Giri, A. An overview on occurrence of polycyclic aromatic hydrocarbons in food chain with special emphasis on human health ailments. *Discover Environ.* **2024**, *2*, No. 87.
- (60) Jameson, C. W. *Tumour Site Concordance and Mechanisms of Carcinogenesis*; Baan, R. A.; Stewart, B. W.; Straif, K., Eds.; IARC Scientific Publications, 2019.
- (61) Newcomb, K.; Sangaiah, R.; Gold, A.; Ball, L. Activation and metabolism of benz [j] aceanthrylene-9, 10-dihydrodiol, the precursor to bay-region metabolism of the genotoxic cyclopenta-PAH benz [j] aceanthrylene. *Mutat. Res., Fundam. Mol. Mech. Mutagen.* **1993**, *287*, 181–190.
- (62) Weisburger, E. K. Carcinogenicity studies on halogenated hydrocarbons. *Environ. Health Perspect.* **1977**, *21*, 7–16.
- (63) Tang, H.; Sun, Q.; Wang, T.; Liao, Y.; Wang, G.; Zhao, F.; Jin, Y. Upregulation of CYP2E1 expression causes oxidative damage induced by 2-chloroethanol in primary cultured rat astrocytes. *Neurotoxicology* **2019**, *75*, 233–244.
- (64) Gwinn, M. R.; Johns, D. O.; Bateson, T. F.; Guyton, K. Z. A review of the genotoxicity of 1, 2-dichloroethane (EDC). *Mutat. Res., Rev. Mutat. Res.* **2011**, *727*, 42–53.
- (65) Guillen, V. M.; Irizarry, L.; Connolly, M. K. *StatPearls [Internet]*; StatPearls Publishing, 2024.
- (66) Carpenter, D. O.; Bushkin-Bedient, S. Exposure to chemicals and radiation during childhood and risk for cancer later in life. *J. Adolesc. Health* **2013**, *52*, S21–S29.
- (67) Evans, S.; Campbell, C.; Naidenko, O. V. Cumulative risk analysis of carcinogenic contaminants in United States drinking water. *Heliyon* **2019**, *5*, No. e02314, DOI: [10.1016/j.heliyon.2019.e02314](https://doi.org/10.1016/j.heliyon.2019.e02314).
- (68) Wang, J.; Li, J. The digital evolution in toxicology: pioneering computational education for emerging challenges. *BMC Med. Educ.* **2024**, *24*, No. 1204.
- (69) Fradkin, P.; Young, A.; Atanackovic, L.; Frey, B.; Lee, L. J.; Wang, B. A graph neural network approach for molecule carcinogenicity prediction. *Bioinformatics* **2022**, *38*, i84–i91.
- (70) Reenu; Vikas. Exploring the role of quantum chemical descriptors in modeling acute toxicity of diverse chemicals to *Daphnia magna*. *J. Mol. Graphics Modell.* **2015**, *61*, 89–101.
- (71) Schieferdecker, S.; Vock, E. Quantum chemical evaluation and QSAR modeling of N-nitrosamine carcinogenicity. *Chem. Res. Toxicol.* **2025**, *38*, 325–339.