*Journal of Applied Toxicology*

WILEY

Journal of AppliedToxicology

RESEARCH ARTICLE OPEN ACCESS

# A Comprehensive Study to Compare Different Compound Representations for Predicting Carcinogenicity In Vivo

Iuri Barbosa Pereira[1] | Rogerio Salvini[2] | Eloisa Dutra Caldas[1]

[1]Laboratório de Toxicologia, Faculdade de Ciências da Saúde, Universidade de Brasília, Brasília, Federal District, Brazil | [2]Instituto de Informática, Universidade Federal de Goiás, Goiânia, Goiás, Brazil

**Correspondence:** Eloisa Dutra Caldas (eloisa@unb.br)

**Received:** 10 April 2026 | **Revised:** 5 June 2026 | **Accepted:** 17 June 2026

**Keywords:** carcinogenicity | classification | computational toxicology | molecular embeddings

## ABSTRACT

Carcinogenicity evaluation is a critical component of chemical risk assessment, yet traditional in vivo testing remains time consuming, costly, and ethically challenging. Computational approaches based on machine learning offer promising alternatives, but the relative contributions of different molecular representation strategies for predicting in vivo carcinogenicity remain insufficiently explored. This study aimed to systematically evaluate the impact of molecular embeddings, classical descriptors, and toxicophore structural alerts on the performance of machine learning models for predicting in vivo carcinogenicity. A curated dataset of 2090 distinct compounds tested in vivo with rodents was assembled by integrating five major toxicological databases. Compounds were represented using classical molecular descriptors, descriptor sets enriched with structural alerts, SMILES-derived molecular embeddings, and hybrid combinations of these representations. Twenty-four machine learning classifiers were benchmarked under a 10-fold stratified cross-validation protocol. Model performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC, with statistical significance evaluated using Friedman and Nemenyi tests. Results indicated that representations combining molecular descriptors with structural alerts tend to yield the most consistent predictive performance across models. Embeddings contribute as complementary features but do not replace classical representations. These findings reinforce the central role of chemically interpretable, expert-driven descriptors, particularly those incorporating genotoxic structural alerts, in regulatory-relevant carcinogenicity modeling.

## 1 | Introduction

Carcinogenicity refers to the ability of a substance to induce cancer by promoting malignant transformation through genetic and epigenetic mechanisms. Accurate carcinogenicity assessment is essential for identifying hazardous agents, providing information to regulatory agencies during chemical registration (e.g., pesticides, food additives, and pharmaceuticals), and reducing human health risks (Benigni et al. 2020; Cordelli et al. 2021; Hartwig et al. 2020).

Computational approaches, including quantitative structure–activity relationship (QSAR) models, machine learning, deep learning architectures, read-across strategies, and systems based on structural alerts, offer faster and more cost-effective alternatives to access the carcinogenic potential of chemicals. These methods use large chemical and biological datasets to predict toxicological outcomes with increasing accuracy (Raillard et al. 2010; Yang et al. 2020).

Structural alert models, in particular, identify molecular fragments associated with genotoxic mechanisms of carcinogenicity, enabling the early screening of compounds with potential DNA-reactive properties. They support prioritizing safer candidates in safety assessments and are especially valuable given the growing number of new chemicals synthesized each year

(Blagg 2010; Hemmerich et al. 2020; Limban et al. 2018; Raies and Bajic 2016; Richard et al. 2016).

In cheminformatics, tokenization is central to converting chemical strings, such as SMILES (Simplified Molecular Input Line Entry System), into discrete, analyzable units. This process breaks complex molecular structures into smaller fragments or tokens that preserve chemical meaning (Yang et al. 2021; Zhang et al. 2020).

Natural Language Processing techniques such as Word2Vec have been widely applied to this task. Word2Vec learns continuous vector representations in which semantic similarity arises from contextual co-occurrence. Its architectures, Continuous Bag of Words (CBOW) and Skip-Gram, predict a target word from its context or predict context from a target. These embeddings capture linguistic regularities, including analogical relationships that emerge from simple vector operations (Mikolov et al. 2013).

Mol2Vec adapts this concept to the chemical domain. Molecules are represented as sentences and structural fragments, often derived from circular fingerprints such as Morgan descriptors, which act as words. Co-occurrence patterns among fragments are used to learn vector embeddings, assigning each fragment a numerical representation. Molecular embeddings are then obtained by combining the vectors of all fragments, commonly through summation or averaging. This approach places structurally or functionally similar molecules in nearby regions of the vector space, improving performance in tasks such as toxicity classification, drug discovery, and similarity analysis (Jaeger et al. 2018; Sadeghi et al. 2024).

Ensemble learning models that combine predictions from multiple classifiers further improve accuracy in cheminformatics applications (Li and Fourches 2021). When combined with Word2Vec embeddings, these models improve prediction of carcinogenic compounds by capturing complex associations between structure and biological effects.

The integration of embeddings and machine learning expands modeling capabilities by capturing nonlinear relationships between molecular structure and carcinogenic outcomes. These methods may improve predictive reliability and may decrease dependence on extensive laboratory testing, reducing both development costs and timelines (Hemmerich et al. 2020; Sharifani and Amini 2023).

This study presents a systematic evaluation of molecular embedding strategies for carcinogenicity prediction based exclusively on in vivo data, which, to the best of our knowledge, has not been previously reported. Different molecular representation strategies were investigated to assess their impact on the predictive performance of machine learning models for carcinogenicity, including classical molecular descriptors, Mol2Vec-based embeddings derived from tokenized SMILES strings, and a hybrid approach integrating both feature types. The goal was to provide a large-scale, controlled evaluation of embedding-based representations, traditional descriptors, and toxicophore alerts.

## 2 | Methods

### 2.1 | Toxicological Dataset

The dataset used in the study was constructed by integrating five publicly available toxicological databases: The Genetic Toxicology Data Bank (GENE-TOX) (US HHS 2025) and the Integrated Risk Information System (IRIS) developed by the United States Environmental Protection Agency (US EPA 2025) (, the Chemical Carcinogenesis Research Information System (CCRIS), maintained by the US National Cancer Institute, which provides curated information on the carcinogenicity, mutagenicity, tumor promotion, and genetic toxicity of chemicals (NCI 2018); the Carcinogenic Potency Database (CPDB), which aggregates results from long-term carcinogenicity bioassays, including genetic toxicity data (Gold et al. 2005); and the European Chemicals Agency (ECHA) Database, which consolidates regulatory submissions under the Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH), including experimental results for industrial chemicals (ECHA 2025). Only compounds tested in vivo were selected from each database.

The core variables shared across the databases, including chemical identity (compound name, CAS number, and SMILES) and endpoint (positive or negative for carcinogenicity), were integrated. Only in vivo data from rodents, which are the animal model for carcinogenicity, from the test system were retained for subsequent analyses. The compound identifiers were standardized by cross-checking CAS (Chemical Abstract Service) numbers and SMILES strings to ensure unique chemical mapping across the databases. This integration process produced a unified, representative dataset of 19,883 carcinogenicity evidence entries, combining experimental data from diverse sources into a single, structured resource.

A multistep curation workflow was subsequently applied to improve data consistency and chemical validity. This process involved the removal of duplicate entries, defined as structurally identical compounds appearing multiple times in the dataset (often originating from different sources or identifiers), as well as the elimination of incomplete records. In addition, entries showing conflicting results, that is, identical chemical structures associated with inconsistent or contradictory experimental outcomes across sources, were carefully analyzed and excluded. Nonorganic compounds were also removed to ensure compatibility with descriptor calculation (Ambure and Cordeiro 2020; Fourches et al. 2010). Figure 1 summarizes the process of integrating the five toxicological databases to produce the final work dataset.

The final dataset encompasses a chemically diverse collection of organic compounds, including pharmaceuticals, pesticides, industrial chemicals, environmental contaminants, natural products, and synthetic research chemicals, with distinct regulatory and scientific purposes. This variety aligns with the goal of assessing a broad range of compound representations rather than focusing on a dataset limited to a particular chemical class.

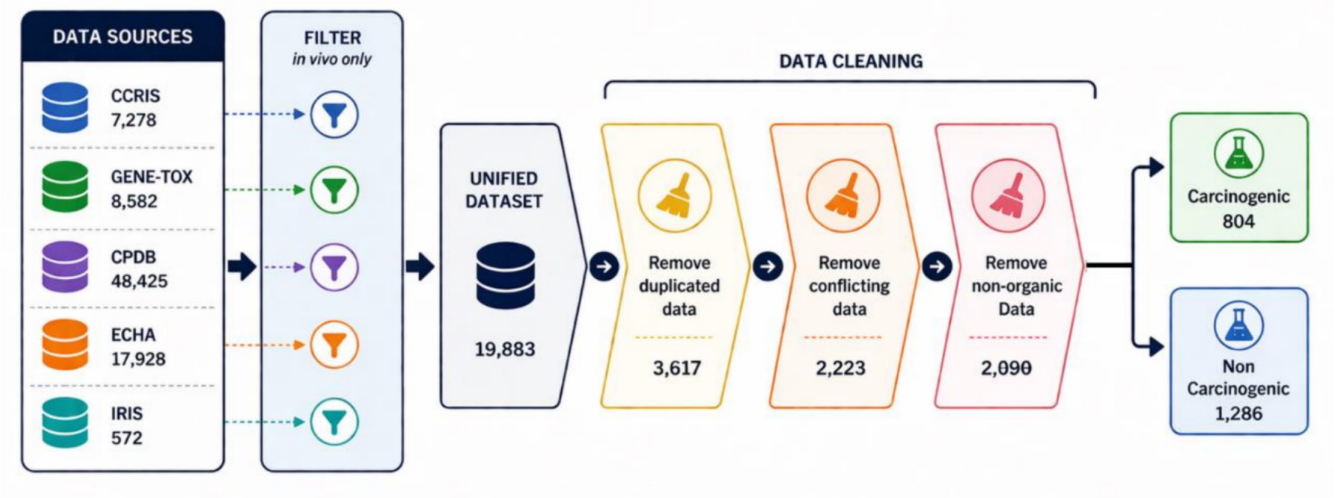


**FIGURE 1** | Integration and curation pipeline for in vivo carcinogenicity data, detailing the contribution of each source database and the progressive reduction in dataset size after removal of duplicates, conflicting records, and nonorganic compounds, resulting in the final curated dataset and class distribution.

## 2.2 | Embedding Representations

All compounds were represented using SMILES. When available, SMILES strings were obtained directly from the original database. For compounds lacking SMILES annotations, molecular structures were retrieved via the PubChem API (Power User Gateway, PUG) using their CAS identifiers. To minimize structural ambiguities related to salts, tautomers, stereochemistry, and isomeric representations, all retrieved structures were standardized using RDKit's MolStandardize (molecular structure standardization tools) functionality, including salt stripping, tautomer normalization, and charge normalization. When applicable, structures were also neutralized to their predominant neutral forms. This procedure ensured consistent and comparable molecular representation across the dataset. Compounds for which a well-defined, standardizable structure could not be obtained were excluded from subsequent analyses.

To convert the curated structure into a vector representation, each SMILES string was tokenized into chemically meaningful fragments that capture local structural environments. Using a Mol2Vec-like embedding model, each fragment was mapped to a continuous vector learned from large molecular corpora, assigning similar embeddings to fragments that occur in related chemical contexts. Then, the fragment-level embeddings were aggregated into a single vector per molecule, yielding a compact numerical representation suitable for downstream predictive modeling. This process is summarized in Figure 2.

Seven pretrained embedding models with different training corpora and objectives were tested. BERT-base (BERT) provides high-quality textual embeddings, although lacking chemical awareness, served as a baseline to highlight the limitations of purely linguistic representations for SMILES strings (Devlin et al. 2019). ChemBERTa_77_MTR (CB77-MTR), ChemBERTa_77_MLM (CB77-MLM), and ChemBERTa_100_M (CB100-M) are transformer-based architectures trained on extensive collections of SMILES strings, designed to capture both local and contextual chemical features (Chithrananda et al. 2020). Molformer (MLFM) is a transformer model pretrained on large-scale molecular graph data, focusing on structural representations of compounds (Wu et al. 2023). PubMedBERT (PMBERT) was trained on the PubMed biomedical literature corpus, capturing semantic and contextual information from biomedical texts (Gu et al. 2021). Finally, ChemBERTa_ZINC (CB-ZINC), a variant of ChemBERTa pretrained on the ZINC chemical database, emphasizes coverage of drug-like molecules (Chithrananda et al. 2020). The best-performing identified embedding model was used as the base representation to compare different experimental scenarios.

## 2.3 | Experimental Scenarios

Five different input-variable scenarios to classify the compounds were analyzed: (i) 366 classical molecular descriptors (DESC), calculated with the RDKit library, which covered physicochemical, topological, and electronic properties such as molecular weight, logP, number of rings, and connectivity indices, thereby representing well-established features traditionally used in QSAR studies; (ii) descriptors enriched with toxicological structural alerts (DEST), derived from OECD guidelines, ICH M7 documentation, IARC reports, and the seminal works of Ashby and Tennant (1988) and Benigni and Bossa (2011), which encoded known structural motifs associated with genotoxic and carcinogenic activity into binary variables (positive or negative) appended to the descriptor set, ensuring that embeddings were not compared directly on the basis of the alerts themselves; (iii) the embeddings alone (EMB), which captured structural and contextual information directly from SMILES strings; (iv) embeddings combined with descriptors (EMBD), integrating distributed representations with handcrafted features to test whether complementary information improved classification performance; and (v) embeddings combined with descriptors and toxicophore alerts (EMBDT), aiming to create the most comprehensive feature

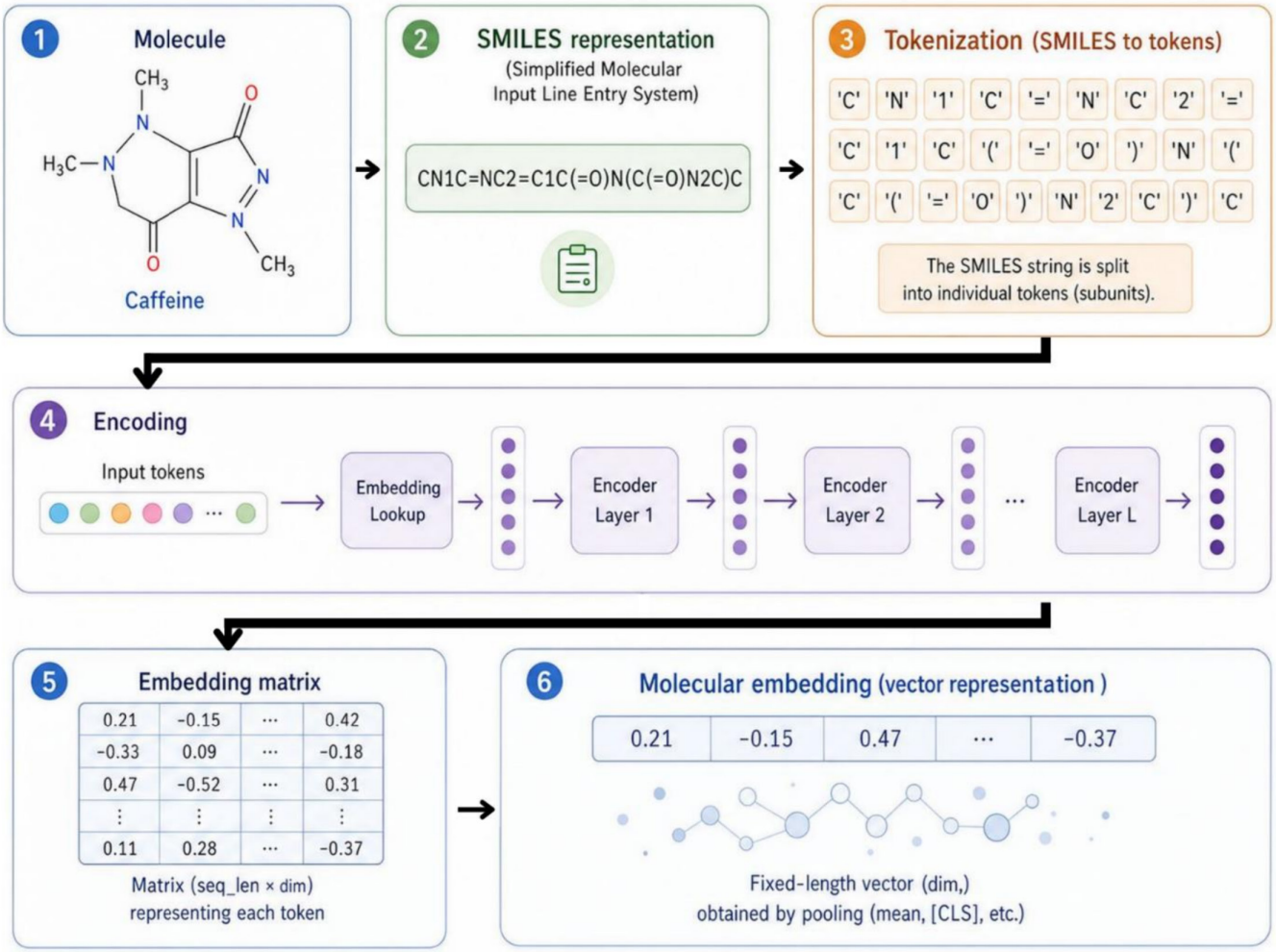


**FIGURE 2** | Representation learning pipeline from molecular structure to embeddings (using caffeine as a model), showing SMILES encoding, tokenization into chemical substructures, and transformation into numerical vectors via an embedding model.

set by uniting learned representations, classical descriptors, and expert-derived toxicological knowledge. Across all configurations, the best-performing identified embedding model was consistently adopted as the base molecular representation to enable a controlled comparison of the different experimental scenarios.

## 2.4 | Data Modeling

Benchmarked 24 machine learning classifiers were applied across all experimental scenarios to evaluate predictive performance covering a broad spectrum of classifier families. The linear models included Logistic Regression, Ridge Classifier, Stochastic Gradient Descent (SGD), PassiveAggressive, and Perceptron. For support vector machines, LinearSVC, SVC, and NuSVC were tested. The instance-based method was represented by *k*-nearest neighbors (KNN). Among tree-based algorithms, Decision Tree and ExtraTree were used, whereas the ensemble learners comprised Random Forest, ExtraTrees, Gradient Boosting, HistGradientBoosting, AdaBoost, and Bagging. Probabilistic approaches were also evaluated, namely, Gaussian Naive Bayes (GaussianNB) and Bernoulli Naive Bayes (BernoulliNB), and the Linear Discriminant Analysis (LDA). In addition, neural network classifiers (Multilayer Perceptron, MLPClassifier) and three widely used external gradient-boosting frameworks—XGBoost, LightGBM, and CatBoost—were used.

All algorithms were evaluated using the default parameters provided by their respective libraries. Preliminary tests with hyperparameter tuning did not improve performance and, in some cases, even led to inferior results (data not shown). Therefore, the default configurations were retained.

Figure 3 provides a schematic summary of the main steps of the proposed approach.

## 2.5 | Model Evaluation

The model performance was assessed using 10-fold stratified cross-validation, a resampling technique that partitions the dataset into 10 equal-sized folds while preserving the original class distribution in each fold. This method uses nine folds for training and reserves one fold for testing the classifiers. The cycle repeats until each fold has been used once as the test set. To further reduce variability associated with data partitioning,

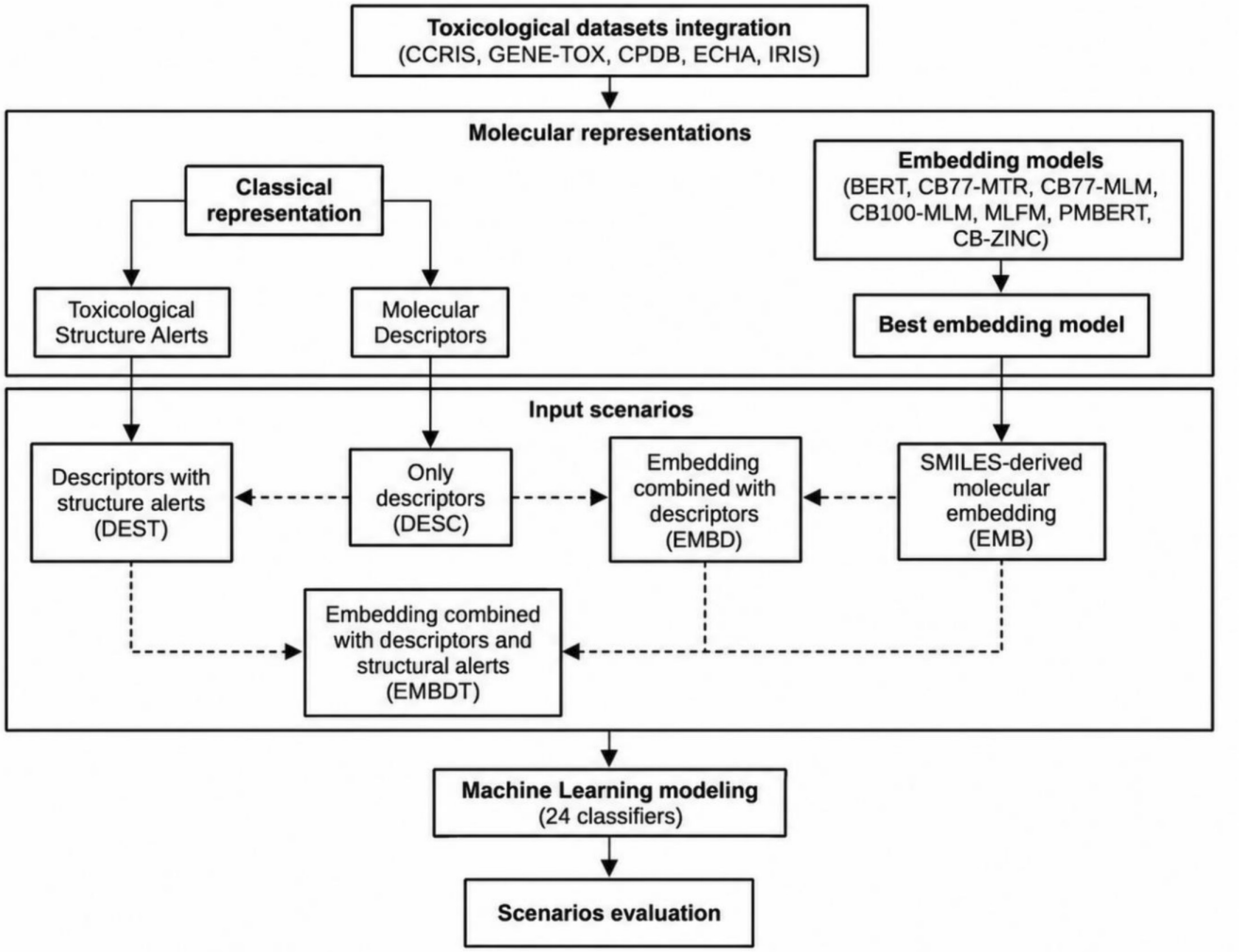


**FIGURE 3** | Schematic representation of the modeling framework, including dataset integration, generation of classical and embedding-based molecular representations, construction of input scenarios, and evaluation using multiple machine learning classifiers.

this procedure was repeated 30 times with different random seeds, yielding 30 independent 10-fold cross-validation runs. This repeated strategy reduces the variance associated with a single train-test split and ensures that both classes, *carcinogenic* (positive class) and *noncarcinogenic* (negative class), are proportionally represented in each iteration, which is particularly important for datasets with imbalanced class distributions. All models were trained and tested using the same data folds to enable performance comparison.

Computed accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) were used to evaluate the classifiers' performance. These metrics were averaged across all folds and are related to the test sets. They are summarized below, where true positive (TP) and true negative (TN) denote the numbers of positive and negative samples correctly classified as positive and negative, respectively. False positive (FP) are negative samples wrongly classified as positive, and false negative (FN) are positive samples that are wrongly classified as negative.

- **Accuracy** was defined as the proportion of instances correctly classified.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- **Precision** was calculated as the proportion of true positives among all predicted positives.

$$Precision = \frac{TP}{(TP + FN)}$$

- **Recall** (or sensitivity) was computed as the proportion of true positives among all actual positives.

$$Recall = \frac{TP}{(TP + FN)}$$

- **F1-score** was obtained as the harmonic mean of precision and recall.

$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

- **AUC-ROC** was calculated by integrating the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR) across all possible thresholds.

$$TPR = \frac{TP}{TP + FN} \qquad FPP = \frac{FP}{FP + TN}$$

$$ROC - AUC = \int_0^1 TPR(FPR)dFPR$$

The Friedman nonparametric test, as described by Demšar (2006), was used to evaluate the statistical significance of the results. The performance comparisons focused on the F1-score, which balances precision and recall. Because the harmonic mean favors smaller values, a high F1-score indicates both high precision and high recall. When the null hypothesis was rejected, the Nemenyi post hoc test was conducted to identify which values differed significantly. The significance level was set at $p = 0.05$. It allows for a more straightforward interpretation of which approaches outperform others, rather than relying solely on raw scores.

### 2.6 | Experimental Setup

The experiments were conducted in Python 3.11.9. Molecular processing was performed using RDKit 2025.03.2 (Landrum 2013). Data manipulation used NumPy 2.2.6 (Van Der Walt et al. 2011) and Pandas 2.3.3 (McKinney 2011). Deep learning components were implemented using PyTorch 2.9.1 + cu128 (Imambi et al. 2021) and the Transformers library 4.57.1 (Devlin et al. 2019). Classical machine learning models were implemented with scikit-learn 1.6.1 (Kramer 2016). Modern gradient-boosting methods relied on XGBoost 3.0.2 (Nalluri et al. 2020), LightGBM 4.6.0 (Ke et al. 2017), and CatBoost 1.2.8 (Prokhorenkova et al. 2018). Class-imbalance handling used imbalanced-learn 0.14.0 (Lemaître et al. 2017). Statistical analyses were performed with SciPy 1.15.3 (Virtanen et al. 2020) and scikit-posthocs 0.11.4 (Terpilowski 2019). Visualizations were produced with Matplotlib 3.10.3 (Bisong 2019).

## 3 | Results

### 3.1 | Embedding-Based Representations

Table 1 summarizes the F1 scores for the embedding models across 24 distinct machine learning classifiers, enabling a broad comparison of their effectiveness. The test yielded a Friedman statistic of 87.61 and a $p$-value of $9.5 \times 10^{-17}$, showing significant differences among the evaluated embedding models. The Nemenyi post hoc test identified significant differences between pairs, with a critical difference (CD) of 1.602. Differences in mean ranks between two groups exceeding the CD are considered statistically significant.

Figure 4 shows a CD diagram to visualize these differences, arranging groups from worst to best by average rank; groups connected by a horizontal line differ only slightly, whereas unconnected groups differ significantly. This ranking-based perspective provides a more reliable comparison than raw F1-scores, highlighting systematic patterns in how each representation behaves across classifiers, revealing stable advantages or weaknesses that are not evident from average metric values alone, as shown in Table 1. CB77-MTR showed a slightly higher mean F1-score than CB77-MLM and outperformed CB100-MLM and BERT, although differences were not statistically significant. CB77-MTR was then selected for subsequent analyses.

### 3.2 | Scenario Performances

Table 2 shows the results for the top 10 classification models, ranked by their average F1-score across all input-variable scenarios. In scenarios where the input data used embeddings (EMB, EMBDESC, and EMBDEST), the CB77MTR was applied.

The Friedman statistic of 50.35 and the corresponding $p$-value of $3.1 \times 10^{-5}$ indicate that the performance ranks are not equivalent across the classifiers. Figure 5 compares the input-variable scenarios. The CD for the mean ranking is 1.149.

The ranking analysis indicates that incorporating structural alerts as input enhances the performance of classification models. The DEST scenario outperforms all others, showing that alert-based information substantially enhances the discriminatory power of classical molecular descriptors. This improvement is not restricted to descriptors alone: when structural alerts are combined with embeddings (EMBDEST), performance also increases relative to the corresponding embedding-based configurations without alerts.

In contrast, embeddings used in isolation (EMB) consistently yield the worst rankings, suggesting that the latent representations captured by the embeddings are insufficient for reliable classification on their own. Hybrid scenarios without alerts (EMBDESC) show intermediate behavior, performing better than pure embeddings but still lacking stability. Overall, the results indicate that structural alerts provide robust, complementary information across representation types, whereas traditional descriptors are more reliable than embeddings when alerts are absent.

### 3.3 | Model Performance

Building on Table 2, the same analysis evaluated performance differences among classifiers across all input-variable scenarios. The Friedman test showed a significant difference ($\chi^2 = 34.60$, $p = 6.98 \times 10^{-5}$), confirming that the classifiers do not perform equivalently.

The CD analysis ($CD = 4.919$) highlights a distinct separation between the top-performing gradient-boosting models and the remaining algorithms. XGBoost achieved the best average rank, followed closely by LightGBM. These models form a stable, high-performing group, indicating that boosting-based methods are better suited to the structure and feature characteristics of this dataset. In contrast, tree ensembles such as ExtraTrees and RandomForest, along with SVC and GradientBoosting, rank among the lowest positions. The MLPClassifier remains in an intermediate position, showing moderate but less stable performance. Figure 6 presents the CD diagram, which summarizes the models' rankings.

Table 3 shows the performance metrics of the XGBoost classifier across all input-variable scenarios. Each scenario is evaluated using mean values and standard deviations, enabling

**TABLE 1** | F1-score of the embeddings across all classification models.

| Models | BERT | CB100-MLM | CB77-MLM | CB77-MTR | PMBERT | MLFM | CB-ZINC |
|---|---|---|---|---|---|---|---|
| AdaBoost | 0.551 | 0.574 | 0.570 | 0.568 | 0.412 | 0.549 | 0.541 |
| Bagging | 0.567 | 0.575 | 0.627 | 0.603 | 0.457 | 0.521 | 0.559 |
| BernoulliNB | 0.554 | 0.436 | 0.564 | 0.539 | 0.461 | 0.579 | 0.547 |
| CatBoost | 0.630 | 0.646 | 0.675 | 0.708 | 0.549 | 0.602 | 0.628 |
| DecisionTree | 0.541 | 0.538 | 0.563 | 0.589 | 0.484 | 0.527 | 0.557 |
| ExtraTree | 0.522 | 0.522 | 0.542 | 0.557 | 0.473 | 0.467 | 0.539 |
| ExtraTrees | 0.622 | 0.614 | 0.643 | 0.680 | 0.517 | 0.567 | 0.600 |
| GaussianNB | 0.588 | 0.479 | 0.585 | 0.516 | 0.404 | 0.600 | 0.570 |
| GradientBoosting | 0.613 | 0.622 | 0.635 | 0.663 | 0.501 | 0.577 | 0.603 |
| HistGradientBoosting | 0.648 | 0.652 | 0.680 | 0.704 | 0.568 | 0.625 | 0.618 |
| KNeighbors | 0.633 | 0.661 | 0.679 | 0.668 | 0.566 | 0.673 | 0.615 |
| LDA | 0.604 | 0.629 | 0.635 | 0.658 | 0.566 | 0.583 | 0.605 |
| LightGBM | 0.660 | 0.653 | 0.699 | 0.705 | 0.562 | 0.628 | 0.632 |
| LinearSVC | 0.606 | 0.614 | 0.626 | 0.649 | 0.577 | 0.567 | 0.592 |
| LogisticRegression | 0.630 | 0.656 | 0.631 | 0.642 | 0.584 | 0.590 | 0.614 |
| MLPClassifier | 0.669 | 0.674 | 0.689 | 0.702 | 0.606 | 0.648 | 0.625 |
| NuSVC | 0.653 | 0.674 | 0.703 | 0.691 | 0.607 | 0.681 | 0.647 |
| PassiveAggressive | 0.588 | 0.605 | 0.570 | 0.569 | 0.517 | 0.589 | 0.552 |
| Perceptron | 0.592 | 0.575 | 0.580 | 0.591 | 0.521 | 0.593 | 0.560 |
| RandomForest | 0.612 | 0.613 | 0.649 | 0.663 | 0.508 | 0.564 | 0.597 |
| RidgeClassifier | 0.612 | 0.640 | 0.630 | 0.646 | 0.576 | 0.584 | 0.613 |
| SGDClassifier | 0.601 | 0.599 | 0.582 | 0.577 | 0.536 | 0.579 | 0.571 |
| SVC | 0.617 | 0.613 | 0.659 | 0.633 | 0.410 | 0.635 | 0.597 |
| XGBoost | 0.651 | 0.654 | 0.692 | 0.711 | 0.573 | 0.645 | 0.635 |
| Mean | 0.607 | 0.608 | 0.627 | 0.638 | 0.594 | 0.519 | 0.593 |

Abbreviations: BERT, BERT-base (general-domain language model); CB-ZINC, ChemBERTa pretrained on the ZINC database; CB100-MLM, ChemBERTa (100k SMILES, masked language modeling); CB77-MLM, ChemBERTa (77M SMILES, masked language modeling); CB77-MTR, ChemBERTa (77M SMILES, multitask regression); MLFM, Molformer (structure-aware transformer for molecular representations); PMBERT, PubMedBERT (biomedical text corpus).

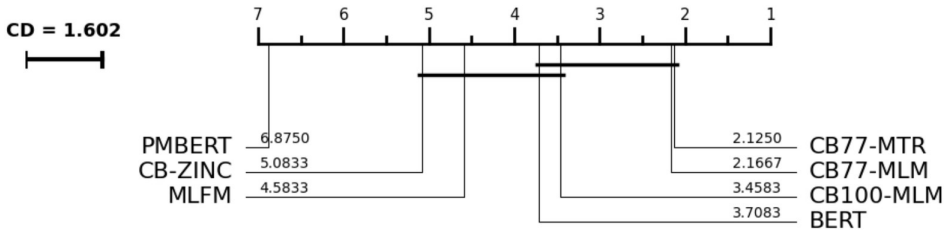


**FIGURE 4** | Critical difference (CD) diagram of embedding methods based on average ranks across datasets (Nemenyi test); methods connected by a horizontal line are not significantly different.

direct comparison of model behavior across different input representations.

XGBoost consistently performs well in all scenarios, with accuracy usually around 0.80. The DEST scenario achieves the highest F1-score (0.741) and AUC-ROC (0.872), providing the best balance between precision and recall. In embedding-based scenarios, the model increases precision but reduces recall, thereby lowering the F1-score. Despite these differences, XGBoost preserves high AUC-ROC values across all input configurations, indicating strong discriminative capability.

**TABLE 2** | F1-score for different input-variable scenarios across top 10 classifiers.

| Models | DESC | DEST | EMB | EMBDESC | EMBDEST |
|---|---|---|---|---|---|
| CatBoost | 0.729 | 0.728 | 0.708 | 0.714 | 0.720 |
| ExtraTrees | 0.715 | 0.738 | 0.680 | 0.705 | 0.718 |
| GradientBoosting | 0.685 | 0.696 | 0.663 | 0.690 | 0.692 |
| HistGradientBoosting | 0.723 | 0.730 | 0.704 | 0.715 | 0.712 |
| KNeighbors | 0.686 | 0.702 | 0.668 | 0.694 | 0.687 |
| LightGBM | 0.731 | 0.731 | 0.705 | 0.726 | 0.724 |
| MLPClassifier | 0.695 | 0.709 | 0.702 | 0.717 | 0.719 |
| NuSVC | 0.692 | 0.715 | 0.691 | 0.707 | 0.718 |
| RandomForest | 0.716 | 0.730 | 0.663 | 0.687 | 0.695 |
| XGBoost | 0.727 | 0.741 | 0.711 | 0.732 | 0.721 |
| Mean | 0.710 | 0.722 | 0.690 | 0.709 | 0.711 |

Abbreviations: DESC, molecular descriptors alone; DEST, descriptors combined with structural alerts; EMB, embedding alone; EMBDESC, embedding combined with descriptors; EMBDEST, embedding combined with descriptors and structural alerts.

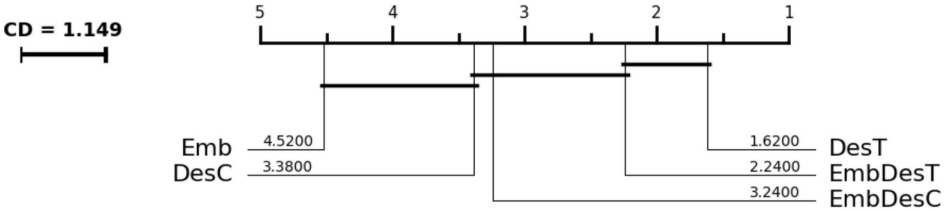


**FIGURE 5** | Critical difference (CD) diagram of input-variable scenarios based on average ranks across datasets (Nemenyi post hoc test); scenarios connected by a horizontal line do not differ statistically significantly.

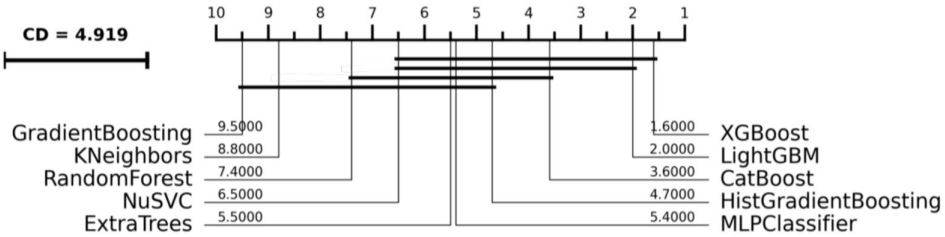


**FIGURE 6** | Critical difference (CD) diagram comparing classification models based on average ranks across datasets (Nemenyi post hoc test); models connected by a horizontal line do not differ statistically significantly.

**TABLE 3** | Performance metrics of XGBoost across all input-variable scenarios.

| Scenario | Accuracy | Precision | Recall | F1-score | AUC-ROC |
|---|---|---|---|---|---|
| DESC | 0.802 ± 0.028 | 0.773 ± 0.045 | 0.688 ± 0.044 | 0.727 ± 0.039 | 0.869 ± 0.022 |
| DEST | 0.811 ± 0.030 | 0.786 ± 0.046 | 0.703 ± 0.057 | 0.741 ± 0.044 | 0.872 ± 0.023 |
| EMB | 0.803 ± 0.019 | 0.810 ± 0.052 | 0.636 ± 0.064 | 0.711 ± 0.048 | 0.848 ± 0.029 |
| EMBDESC | 0.810 ± 0.017 | 0.799 ± 0.035 | 0.677 ± 0.042 | 0.732 ± 0.026 | 0.867 ± 0.019 |
| EMBDEST | 0.803 ± 0.019 | 0.790 ± 0.033 | 0.665 ± 0.047 | 0.721 ± 0.031 | 0.866 ± 0.018 |

Abbreviations: DESC, molecular descriptors alone; DEST, descriptors combined with structural alerts; EMB, embedding alone; EMBDESC, embedding combined with descriptors; EMBDEST, embedding combined with descriptors and structural alerts.

## 4 | Discussion

The present study evaluates 2090 distinct molecules that are exclusively associated with an in vivo carcinogenicity endpoint. Although smaller than the largest Ames-based datasets, this collection exceeds the size of most previously reported in vivo studies and matches or surpasses the effective per-endpoint sizes used in many mixed-endpoint and multitask approaches. Importantly, the dataset was not aggressively filtered to remove ambiguous or challenging cases, resulting in a heterogeneous but realistic modeling scenario. Taken together, these comparisons highlight that studies reporting large nominal datasets often restrict their scope to in vitro assays or aggregate data across multiple endpoints. When dataset size and biological relevance are considered jointly, the present work occupies a distinct position by combining a relatively large number of molecules with a single, biologically in vivo endpoint.

Given the heterogeneous origin and size of the dataset, the choice of evaluation strategy becomes particularly important. Rather than relying on a simple holdout split, the present study evaluates model performance using a robust 10-fold cross-validation framework. This approach provides more stable and reliable performance estimates, especially for imbalanced toxicological datasets, and reduces the likelihood that favorable data partitions lead to overly optimistic conclusions. The resulting performance metrics exhibit low standard deviation across folds, indicating limited sensitivity to data partitioning and reducing the likelihood that favorable splits drive the observed results.

The findings of this study differ from most previous computational studies, which primarily depend on in vitro assays such as the Ames test or on mixed in vitro/in vivo endpoints (Chen et al. 2024; Gini et al. 2019; Karamertzanis et al. 2025; Shinada et al. 2022). This distinction is central from a regulatory perspective, as in vivo outcomes are generally regarded as more directly informative for human risk assessment.

Some studies also have developed computational models for in vivo carcinogenicity. Li et al. (2015), Zhang et al. (2017), and Lagunin et al. (2018) focused on CPDB-derived rat carcinogenic data, using molecular fingerprints and descriptors. Li et al. (2021) also incorporated NCTRlcdb annotations from CPDB. Unlike previous work, this study systematically compares molecular descriptors, embeddings, hybrid features, and toxicophore-enriched data using a larger, multisource in vivo dataset and a unified validation approach.

Across the evaluated scenarios, pretrained molecular embeddings act as a complementary representation rather than a replacement for classical molecular descriptors. Models based exclusively on embeddings do not outperform descriptor-based models, but combining embeddings with descriptors consistently improves performance relative to either representation used in isolation. This result indicates that the pretrained embeddings evaluated here, which leverage molecular representations learned from large external corpora in a transfer learning setting, were informative but not sufficient as standalone features for this specific in vivo endpoint and dataset.

The important role of toxicophore alert structures deserves special emphasis. Although some carcinogenicity models used fingerprints or descriptors that implicitly encode substructural information, few studies evaluated toxicophore alerts as an explicit, separable feature block whose contribution could be compared with descriptors, embeddings, and hybrid representations. Gini et al. (2019), for example, trained deep neural networks directly on SMILES strings and two-dimensional molecular images, using structural alerts only for post hoc interpretation. Chen et al. (2024) relied exclusively on SMILES-derived embeddings, whereas Tevosyan et al. (2025) and Karamertzanis et al. (2025) employed graph neural networks that learn representations end-to-end from molecular graphs. None of these studies evaluated structural alerts as stand-alone or auxiliary features. Descriptor-based QSAR studies acknowledge structural alerts more frequently but rarely integrate them in a controlled manner. Fan et al. (2018) and Guan et al. (2018) relied on molecular descriptors and fingerprints without including explicit alert features, leaving any association between known toxicophores and model predictions implicit.

Only a small number of studies explicitly incorporate structural alerts into the modeling process. Ramesh and Veerappapillai (2021) combined molecular fingerprints with SARpy-derived genotoxic alerts and reported improved performance relative to fingerprints alone. However, this result is based on a training set of 272 compounds and does not involve systematic comparisons across multiple representation types or validation protocols. Shinada et al. (2022) included genotoxicity alerts alongside fingerprints, physicochemical descriptors, and quantum chemical properties but did not isolate the specific contribution of alerts from the broader feature set. Alert-based features encode mechanistically grounded knowledge related to electrophilic functional groups, metabolic activation pathways, and DNA-reactive moieties, which are central to carcinogenic processes. Global descriptors do not fully capture such information or what embeddings implicitly learn, especially when training data are limited. This explains why including alerts improved model performance across both descriptor-based and embedding-based representations.

Recent studies on genotoxicity have moved beyond classical molecular descriptors by adopting learned representations directly derived from molecular structure. Gini et al. (2019) demonstrated that deep learning models trained on raw representations, such as SMILES strings and graph-based images, can achieve performance comparable to or superior to state-of-the-art QSAR approaches when applied to large datasets of Ames test results. Similarly, Chen et al. (2024) proposed an SMILES-based convolutional neural network framework that achieved high predictive performance on mutagenicity data, enabling the identification of molecular-specific mutagenic fragments, thereby reinforcing the potential of sequence-based learned representations as alternatives to traditional fingerprints. Extending this line of work, Karamertzanis et al. (2025) showed that graph neural network embeddings trained on large-scale in vitro genotoxicity datasets effectively separate positive and negative compounds and cluster known genotoxicity alerts, supporting the suitability of graph-based representations for modeling genotoxicity at scale. However, in vitro genotoxicity assays capture only a subset of the mechanisms

underlying carcinogenicity and do not account for processes such as metabolic bioactivation, chronic toxicity, inflammation, or tumor promotion (EFSA Scientific Committee et al. 2017). Consequently, although comparisons with genotoxicity-based models are mechanistically and methodologically relevant, differences in absolute performance should be interpreted in light of the greater biological complexity of in vivo carcinogenicity, highlighting the need for models specifically developed for this endpoint.

Guan et al. (2018) developed a hybrid modeling framework to predict in vivo rodent carcinogenicity by combining the outputs of multiple assay-specific QSAR models, rather than by training a single model on pooled data. They independently trained QSAR models for Ames mutagenicity (6512 compounds), Syrian Hamster Embryonic cell transformation (410 compounds), GreenScreen GADD45a-GFP (1415 compounds), and an ISSCAN rodent carcinogenicity dataset (834 compounds), using conventional molecular descriptors. The authors evaluated the individual assay models using repeated 10-fold cross-validation and then integrated the best-performing models into a cascade framework to predict rodent carcinogenicity. The performance of the final cascade model was reported primarily on an external rodent carcinogenicity dataset, where it achieved 69.3% accuracy and an AUC of 0.70, without cross-validated performance estimates for the integrated model. As a result, the reported metrics reflect external prediction of the cascade rather than its internal robustness. In contrast to the present work, this approach relies exclusively on handcrafted descriptors. It provides limited validation of the final integrated model, thereby restricting conclusions about its generalization and stability across chemically diverse compounds.

Tevosyan et al. (2025) proposed a multitask deep learning framework based on graph neural networks for predicting human carcinogenicity. They incorporated auxiliary tasks, including mutagenicity, genotoxicity, animal carcinogenicity, and hormone receptor binding, to support the primary prediction task. The authors explicitly addressed class imbalance using balanced accuracy metrics. The best-performing multitask model achieved an AUC of 0.89 and a balanced accuracy of 82%, with a sensitivity of 0.75 and a specificity of 0.89. Although this approach substantially improved performance, it relies heavily on heterogeneous auxiliary endpoints, which may introduce bias and reduce interpretability when transferred to narrower problem settings, such as the one addressed in this study. Although not directly modeling carcinogenicity as a primary endpoint, Karamertzanis et al. (2025) provided an important regulatory-relevant reference in which the multitask graph neural network models predicted multiple in vitro genotoxicity assays under the REACH framework using more than 12,000 compounds. It required at least 200 positive and 200 negative samples for external validation and reported external balanced accuracies ranging from 72% to 78%. In contrast to most previous studies, which focus on specific genotoxicity endpoints, the present work evaluates carcinogenicity solely at the outcome level, without explicitly modeling individual mechanistic endpoints.

The results of this study either outperform or match those of recent modeling studies. Compared with the cascade QSAR model proposed by Guan et al. (2018), the XGBoost model improves accuracy by approximately 0.12 absolute points and increases AUC-ROC by about 0.17, indicating improved discriminative performance. Tevosyan et al. (2025) reported a slightly higher AUC using multitask graph neural networks; however, their balanced accuracy differs by less than 0.01 absolute points from the accuracy achieved in this study, despite their reliance on multiple auxiliary tasks and considerably more complex model architectures. In addition, the present work's approach exceeds the external balanced accuracies reported for graph-based multitask models by Karamertzanis et al. (2025) by approximately 0.03–0.09 absolute points, depending on the configuration. Compared to previous works, instead of a basic holdout split, this study employs a robust repeated 10-fold cross-validation to assess model performance. This method yields more stable and reliable estimates, particularly for imbalanced toxicological datasets, and minimizes the risk of overly optimistic results from favorable data partitions. The performance metrics show a low standard deviation across folds, indicating that the results are consistent regardless of how the data are split and reducing the chance that positive outcomes are due to specific data partitions.

Despite its strengths, this study has some limitations. Although the dataset is relatively large for an exclusively in vivo carcinogenicity endpoint, it remains small compared with large in vitro or multitask datasets, which constrains the expressive power of data-hungry deep learning models and partly explains the limited stand-alone performance of pretrained molecular embeddings. Carcinogenicity is modeled as a single binary endpoint, which necessarily collapses differences across species (rats or mice), target organs, exposure conditions, and underlying genotoxic and non-genotoxic mechanisms.

## 5 | Conclusion

This study presents a comprehensive evaluation of molecular representation strategies for in vivo carcinogenicity classification using an integrated dataset derived from multiple regulatory toxicology sources. Previous studies have already shown that rodent carcinogenicity can be predicted using molecular fingerprints, SAR models, and ensemble learning representations. The present work extends this literature by comparing classical molecular descriptors, molecular embeddings, hybrid representations, and toxicophore-enriched features under the same validation framework and across multiple classifiers. Under these experimental conditions, descriptor-based representations, particularly when enriched with toxicophore alerts, provided the most consistent predictive performance.

When performance is analyzed across classifiers using the F1-score, representations incorporating structural alerts consistently achieve higher average performance. The DEST scenario attains the highest mean F1-score, outperforming descriptor-only models and all embedding-based configurations. A similar trend is observed for embedding-based representations: EMBDEST improves upon embeddings used in isolation, indicating that alert-based information provides a robust and complementary signal regardless of the underlying molecular representation. In contrast, embeddings alone yield the lowest overall performance, whereas hybrid representations without

alerts exhibit intermediate behavior, reinforcing the view that embeddings contribute auxiliary information but do not replace classical descriptors.

An analysis focused on the best-performing classifier further supports these conclusions. XGBoost consistently achieves accuracy values close to 0.80 across all scenarios, with the DEST configuration providing the best balance between precision and recall and the highest AUC-ROC. In embedding-only models, XGBoost exhibits higher precision but substantially lower recall, resulting in a lower F1 Score. The inclusion of embeddings alongside descriptors partially mitigates this effect, as observed in the EMBDESC scenario, while preserving high discriminative ability, with AUC-ROC values remaining consistently high across all representations.

Taken together, these results demonstrate that molecular embeddings are most effective as complementary features for refining existing descriptor-based models, particularly when combined with toxicophore structural alerts. For in vivo carcinogenicity prediction, chemically interpretable descriptors enriched with expert knowledge remain the dominant source of predictive power, whereas embeddings provide secondary gains rather than serving as stand-alone solutions.

Future work could integrate biological and mechanistic information beyond molecular structure, such as metabolic bioactivation, toxicokinetics, inflammation, and tumor promotion, to more faithfully represent the processes underlying carcinogenic outcomes and improve predictive performance.


### Acknowledgments

We thank the Brazilian National Council for Scientific and Technological Development (CNPq) for the PhD scholarship provided to the first author. During the preparation of this work, the authors used OpenAI's ChatGPT to support text organization and language refinement. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

### Funding

This work was supported by the National Council for Scientific and Technological Development.


### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

Data will be made available on request.

### References


Ambure, P., and M. N. D. S. Cordeiro. 2020. "Importance of Data Curation in QSAR Studies Especially While Modeling Large-Size Datasets." In *Ecotoxicological QSARs*, edited by K. Roy, 97–109. Springer US. https://doi.org/10.1007/978-1-0716-0150-1_5.

Ashby, J., and R. W. Tennant. 1988. "Chemical Structure, *Salmonella* Mutagenicity and Extent of Carcinogenicity as Indicators of Genotoxic Carcinogenesis Among 222 Chemicals Tested in Rodents by the US NCI/NTP." *Mutation Research, Genetic Toxicology* 204, no. 1: 17–115.

Benigni, R., A. Bassan, and M. Pavan. 2020. "In Silico Models for Genotoxicity and Drug Regulation." *Expert Opinion on Drug Metabolism & Toxicology* 16, no. 8: 651–662.

Benigni, R., and C. Bossa. 2011. "Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review With Implications for Predictive Toxicology." *Chemical Reviews* 111, no. 4: 2507–2536.

Bisong, E. 2019. "Matplotlib and Seaborn." In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginner*. Apress.

Blagg, J. 2010. "Structural Alerts for Toxicity." In *Burger's Medicinal Chemistry and Drug Discovery*, edited by D. J. Abrahan and D. P. Rotella, 301–334. John Wiley and Sons, Inc.

Chen, C., Z. Huang, X. Zou, S. Li, D. Zhang, and S.-L. Wang. 2024. "Prediction of Molecular-Specific Mutagenic Alerts and Related Mechanisms of Chemicals by a Convolutional Neural Network (CNN) Model Based on SMILES Split." *Science of the Total Environment* 917: 170435.

Chithrananda, S., G. Grand, and B. Ramsundar. 2020. "ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction." arXiv Preprint arXiv:2010.09885.

Cordelli, E., M. Bignami, and F. Pacchierotti. 2021. "Comet Assay: A Versatile but Complex Tool in Genotoxicity Testing." *Toxicology Research* 10, no. 1: 68–78.

Demšar, J. 2006. "Statistical Comparisons of Classifiers Over Multiple Data Sets." *Journal of Machine Learning Research* 7: 1–30.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 4171–4186.

EFSA Scientific Committee; E. S., A. Hardy, D. Benford, et al. 2017. "Clarification of Some Aspects Related to Genotoxicity Assessment." *EFSA Journal* 15, no. 12: e05113.

European Chemicals Agency (ECHA). 2025. "ECHA Database on Chemicals." https://chem.echa.europa.eu/.

Fan, D., H. Yang, F. Li, et al. 2018. "In Silico Prediction of Chemical Genotoxicity Using Machine Learning Methods and Structural Alerts." *Toxicology Research* 7, no. 2: 211–220.

Fourches, D., E. Muratov, and A. Tropsha. 2010. "Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research." *Journal of Chemical Information and Modeling* 50, no. 7: 1189–1204. https://doi.org/10.1021/ci100176x.

Gini, G., F. Zanoli, A. Gamba, G. Raitano, and E. Benfenati. 2019. "Could Deep Learning in Neural Networks Improve the QSAR Models?" *SAR and QSAR in Environmental Research* 30, no. 9: 617–642.

Gold, L. S., N. B. Manley, T. H. Slone, L. Rohrbach, and G. B. Garfinkel. 2005. "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." *Toxicological Sciences* 85, no. 2: 747–808.

Gu, Y., R. Tinn, H. Cheng, et al. 2021. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing." *ACM Transactions on Computing for Healthcare (HEALTH)* 3, no. 1: 1–23.

Guan, D., K. Fan, I. Spence, and S. Matthews. 2018. "Combining Machine Learning Models of In Vitro and In Vivo Bioassays Improves Rat Carcinogenicity Prediction." *Regulatory Toxicology and Pharmacology* 94: 8–15.

Hartwig, A., M. Arand, B. Epe, et al. 2020. "Mode of Action-Based Risk Assessment of Genotoxic Carcinogens." *Archives of Toxicology* 94, no. 6: 1787–1877.

Hemmerich, J., F. Troger, B. Füzi, and G. F. Ecker. 2020. "Using Machine Learning Methods and Structural Alerts for Prediction of Mitochondrial Toxicity." *Molecular Informatics* 39, no. 5: 2000005.

Imambi, S., K. B. Prakash, and G. Kanagachidambaresan. 2021. "PyTorch." In *Programming With TensorFlow: Solution for Edge Computing Applications*, 87–104. Springer.

Jaeger, S., S. Fulle, and S. Turk. 2018. "Mol2vec: Unsupervised Machine Learning Approach With Chemical Intuition." *Journal of Chemical Information and Modeling* 58, no. 1: 27–35.

Karamertzanis, P. G., M. Rasenberg, I. Shah, and G. Patlewicz. 2025. "Modelling In Vitro Mutagenicity Using Multi-Task Deep Learning and REACH Data." *Chemical Research in Toxicology* 38, no. 8: 1382–1407.

Ke, G., Q. Meng, T. Finley, et al. 2017. "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree." In *31st Confereence on Neural Information Processing Systems*. NIPSS.

Kramer, O. 2016. "Scikit-Learn." In *Machine Learning for Evolution Strategies*, 45–53. Springer.

Lagunin, A., A. Rudik, D. Druzhilovsky, D. Filimonov, and V. Poroikov. 2018. "ROSC-Pred: Web-Service for Rodent Organ-Specific Carcinogenicity Prediction." *Bioinformatics* 34, no. 4: 710–712.

Landrum, G. 2013. "RDKit Documentation." RDKit. https://www.rdkit.org/docs/.

Lemaître, G., F. Nogueira, and C. K. Aridas. 2017. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research* 18, no. 17: 1–5.

Li, T., W. Tong, R. Roberts, Z. Liu, and S. Thakkar. 2021. "DeepCarc: Deep Learning-Powered Carcinogenicity Prediction Using Model-Level Representation." *Frontiers in Artificial Intelligence* 4: 757780.

Li, X., Z. Du, J. Wang, et al. 2015. "In Silico Estimation of Chemical Carcinogenicity With Binary and Ternary Classification Methods." *Molecular Informatics* 34, no. 4: 228–235.

Li, X., and D. Fourches. 2021. "SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning." *Journal of Chemical Information and Modeling* 61, no. 4: 1560–1569. https://doi.org/10.1021/acs.jcim.0c01127.

Limban, C., D. C. Nuţă, C. Chiriţă, et al. 2018. "The Use of Structural Alerts to Avoid the Toxicity of Pharmaceuticals." *Toxicology Reports* 5: 943–953.

McKinney, W. 2011. "pandas: A Foundational Python Library for Data Analysis and Statistics." *Python for High Performance and Scientific Computing* 14, no. 9: 1–9.

Mikolov, T., W. Yih, and G. Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 746–751.

Nalluri, M., M. Pentela, and N. R. Eluri. 2020. "A Scalable Tree Boosting System: XG Boost." *International Journal of Scientific Research in Science, Engineering and Technology* 7, no. 12: 36–51.

National Cancer Institute (NCI). 2018. "CCRIS: Chemical Carcinogenesis Research Information System." https://pubchem.ncbi.nlm.nih.gov/source/22070.

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. 2018. "CatBoost: Unbiased Boosting With Categorical Features." In *32nd Conference on Neural Information Processing Systems*. NeurIPS.

Raies, A. B., and V. B. Bajic. 2016. "In Silico Toxicology: Computational Methods for the Prediction of Chemical Toxicity." *Wiley Interdisciplinary Reviews: Computational Molecular Science* 6, no. 2: 147–172.

Raillard, S. P., J. Bercu, S. W. Baertschi, and C. M. Riley. 2010. "Prediction of Drug Degradation Pathways Leading to Structural Alerts for Potential Genotoxic Impurities." *Organic Process Research & Development* 14, no. 4: 1015–1020.

Ramesh, P., and S. Veerappapillai. 2021. "Prediction of Micronucleus Assay Outcome Using In Vivo Activity Data and Molecular Structure Features." *Applied Biochemistry and Biotechnology* 193, no. 12: 4018–4034.

Richard, A. M., R. S. Judson, K. A. Houck, et al. 2016. "ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology." *Chemical Research in Toxicology* 29, no. 8: 1225–1251.

Sadeghi, S., A. Bui, A. Forooghi, J. Lu, and A. Ngom. 2024. "Can Large Language Models Understand Molecules?" *BMC Bioinformatics* 25, no. 1: 225.

Sharifani, K., and M. Amini. 2023. "Machine Learning and Deep Learning: A Review of Methods and Applications." *World Information Technology and Engineering Journal* 10, no. 7: 3897–3904.

Shinada, N. K., N. Koyama, M. Ikemori, et al. 2022. "Optimizing Machine-Learning Models for Mutagenicity Prediction Through Better Feature Selection." *Mutagenesis* 37, no. 3–4: 191–202.

Terpilowski, M. A. 2019. "scikit-Posthocs: Pairwise Multiple Comparison Tests in Python." *Journal of Open Source Software* 4, no. 36: 1169.

Tevosyan, A., H. Yeghiazaryan, G. Tadevosyan, et al. 2025. "AI/ML Modeling to Enhance the Capability of In Vitro and In Vivo Tests in Predicting Human Carcinogenicity." *Mutation Research, Genetic Toxicology and Environmental Mutagenesis* 903: 503858.

United States Environmental Protection Agency (US EPA). 2025. "IRIS: Integrated Risk Information System." https://www.epa.gov/iris.

United States Government Department of Health and Human Services (US HHS). 2025. "GENE-TOX: Genetic Toxicology Data Bank." https://healthdata.gov/NIH/GENE-TOX-Genetic-Toxicology-Data-Bank-/mg75-uw2s/about_data.

Van Der Walt, S., S. C. Colbert, and G. Varoquaux. 2011. "The NumPy Array: A Structure for Efficient Numerical Computation." *Computing in Science & Engineering* 13, no. 2: 22–30.

Virtanen, P., R. Gommers, T. E. Oliphant, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17, no. 3: 261–272.

Wu, F., D. Radev, and S. Z. Li. 2023. "Molformer: Motif-Based Transformer on 3D Heterogeneous Molecular Graphs." *Proceedings of the AAAI Conference on Artificial Intelligence* 37, no. 4: 5312–5320. https://doi.org/10.1609/aaai.v37i4.25662.

Yang, H., C. Lou, W. Li, G. Liu, and Y. Tang. 2020. "Computational Approaches to Identify Structural Alerts and Their Applications in Environmental Toxicology and Drug Discovery." *Chemical Research in Toxicology* 33, no. 6: 1312–1322. https://doi.org/10.1021/acs.chemrestox.0c00006.

Yang, X., Z. Zhang, Q. Li, and Y. Cai. 2021. "Quantitative Structure–Activity Relationship Models for Genotoxicity Prediction Based on Combination Evaluation Strategies for Toxicological Alternative Experiments." *Scientific Reports* 11, no. 1: 8030. https://doi.org/10.1038/s41598-021-87225-8.

Zhang, L., H. Ai, W. Chen, et al. 2017. "CarcinoPred-EL: Novel Models for Predicting the Carcinogenicity of Chemicals Using Molecular Fingerprints and Ensemble Learning Methods." *Scientific Reports* 7, no. 1: 2118.

Zhang, Y.-F., X. Wang, A. C. Kaushik, et al. 2020. "SPVec: A Word2vec-Inspired Feature Representation Method for Drug-Target Interaction Prediction." *Frontiers in Chemistry* 7: 895.