Hybrid Graph Attention Network and XGBoost Framework for Interpretable Molecular Toxicity Prediction

Stow, May* and Asotekari Angel Jombo**

Department of Computer Science and Informatics, Federal University Otuoke, Nigeria*
Department of Computer Science and Informatics, Federal University Otuoke, Nigeria **

maystow@gmail.com*, jomboaa@fuotuoke.edu.ng**
Orcid ID: https://orcid.org/0009-0006-8653-8363*

DOI: 10.29322/IJSRP.16.01.2026.p16927
https://dx.doi.org/10.29322/IJSRP.16.01.2026.p16927

Paper Received Date:25th December 2025
Paper Acceptance Date: 25th January 2026
Paper Publication Date: 29th January 2026

ABSTRACT

Traditional experimental toxicity testing is costly, time-consuming, and raises ethical concerns, creating an urgent need for computational alternatives that can rapidly screen chemical safety while maintaining interpretability for regulatory acceptance. This study develops a hybrid framework combining Graph Attention Networks (GAT) with XGBoost for molecular toxicity prediction on the Tox21 benchmark dataset. The method leverages GAT's attention mechanism to learn toxicity-relevant substructures from molecular graphs encoded with 11-dimensional atom-level features (atomic number, degree, formal charge, aromaticity, hydrogen count, implicit valence, and hybridization states), subsequently combining these learned embeddings with six physicochemical descriptors (molecular weight, lipophilicity, TPSA, hydrogen bond donors, hydrogen bond acceptors, rotatable bonds) for classification. SMOTE oversampling applied exclusively to the training set addresses severe class imbalance while preventing data leakage. The framework achieved 0.892 ROC-AUC on the SR-MMP endpoint (mitochondrial membrane potential disruption), outperforming Random Forest with ECFP fingerprints (0.810) and GAT encoder alone (0.723). Comparison with a descriptor-only XGBoost baseline (0.887) reveals a modest improvement, though SHAP analysis demonstrates that GAT-learned features rank among the top 15 predictors, validating their complementary value. The model identified 121 of 155 toxic compounds (78% recall) at 46% precision, prioritizing sensitivity for screening applications. Cross-endpoint validation demonstrated ROC-AUC values above 0.74 for 11 of 12 Tox21 assays, establishing generalization across diverse biological mechanisms. SHAP-based explainability analysis revealed that lipophilicity and molecular weight dominate toxicity predictions, with multiple GAT-learned features capturing structural patterns complementary to traditional descriptors. This interpretable hybrid approach contributes to green chemistry by enabling data-driven toxicological assessment that balances predictive performance with mechanistic transparency.

Keywords: *Graph Attention Networks, Molecular Toxicity Prediction, Explainable Artificial Intelligence, SHAP Analysis, Gradient Boosting, Tox21 Dataset*

I. INTRODUCTION

Traditional approaches to chemical safety assessment rely heavily on experimental animal testing, a methodology fraught with ethical concerns, substantial costs, and time-intensive protocols that can delay product development and regulatory approval [1]. The European Union's Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) regulation and similar global frameworks demand comprehensive toxicity data for thousands of chemical substances, creating an urgent need for reliable computational alternatives [2]. Predictive toxicology has emerged as a critical field at the intersection of computational chemistry, machine learning, and regulatory science, offering the potential to accelerate chemical screening while reducing dependence on animal testing [3].

Quantitative structure-activity relationship (QSAR) models have been the traditional computational approach for toxicity prediction, utilizing molecular descriptors and fingerprints to correlate chemical structure with biological activity [4]. While QSAR methods have demonstrated utility in pharmaceutical drug discovery and environmental risk assessment, they are fundamentally limited by their reliance on hand-crafted features that may not capture the complex topological and electronic properties governing molecular toxicity [5]. Molecular fingerprints, such as Extended Connectivity Fingerprints (ECFP), encode structural information as fixed-length bit vectors but inherently lose spatial and stereochemical information during the conversion process [6]. This information loss becomes particularly problematic when predicting endpoints that depend on three-dimensional molecular interactions, such as receptor binding or membrane disruption mechanisms.

This publication is licensed under Creative Commons Attribution CC BY.

Recent advances in deep learning have catalyzed a paradigm shift in molecular property prediction, with graph neural networks (GNNs) emerging as a natural framework for representing chemical structures [7]. Unlike traditional fingerprints, GNNs operate directly on molecular graphs where atoms are represented as nodes and chemical bonds as edges, preserving the inherent graph structure of molecules [8]. Message-passing neural networks, which iteratively aggregate information from neighboring atoms, have shown promising results in predicting molecular properties across diverse chemical spaces [9]. Among GNN architectures, Graph Attention Networks (GATs) introduce an attention mechanism that learns to weigh the importance of neighboring atoms, enabling the model to identify toxicity-relevant substructures without explicit feature engineering [10].

The Tox21 Data Challenge, launched by the National Institutes of Health in collaboration with the Environmental Protection Agency and the Food and Drug Administration, established a benchmark dataset for computational toxicology comprising over 10,000 compounds tested across 12 toxicity assays [11]. This dataset has become a standard benchmark for evaluating machine learning approaches to toxicity prediction, with reported performance varying significantly across endpoints due to differences in biological mechanisms, data quality, and class imbalance [12]. Conventional machine learning methods applied to Tox21, including random forests and support vector machines with molecular fingerprints, typically achieve area under the receiver operating characteristic curve (ROC-AUC) values ranging from 0.70 to 0.75 for most endpoints [13]. More recent graph-based deep learning approaches have pushed performance to 0.75-0.82 ROC-AUC, demonstrating the value of learning representations directly from molecular graphs [14].

Despite these advances, several critical limitations persist in current approaches to toxicity prediction. First, purely graph-based methods, while effective at capturing local structural patterns, may fail to encode global physicochemical properties that influence absorption, distribution, metabolism, and excretion (ADME) characteristics relevant to toxicity [15]. Second, severe class imbalance in toxicity datasets, where toxic compounds often represent less than 15% of samples, poses significant challenges for classifier training and can lead to models biased toward predicting the majority class [16]. Third, the black-box nature of deep neural networks limits their regulatory acceptance, as interpretability and mechanistic understanding remain essential requirements for toxicological risk assessment [17]. Finally, most published studies focus on single-model architectures without exploring hybrid approaches that combine the complementary strengths of graph-based and descriptor-based representations.

Green chemistry principles emphasize the design of chemical products and processes that minimize hazardous substance generation and environmental impact [18]. Computational toxicity prediction aligns directly with these principles by enabling early-stage identification of potentially hazardous compounds, thereby guiding synthetic chemistry toward safer molecular designs [19]. However, the integration of machine learning models into green chemistry workflows requires not only high predictive accuracy but also interpretability that can inform rational molecular modifications to reduce toxicity while maintaining desired functionality [20].

This work addresses the identified limitations through the development and validation of a hybrid machine learning framework that integrates Graph Attention Networks with gradient boosting for interpretable molecular toxicity prediction. The proposed approach leverages GAT's ability to learn attention-weighted molecular representations that identify toxicity-relevant substructures, while incorporating physicochemical descriptors to capture global molecular properties influencing bioavailability and cellular uptake. To address class imbalance, the framework employs Synthetic Minority Oversampling Technique (SMOTE) in conjunction with class weighting during classifier training. Model interpretability is achieved through SHapley Additive exPlanations (SHAP) analysis, enabling identification of molecular features driving toxicity predictions and providing actionable insights for molecular design.

The primary contributions of this research are fourfold. First, a hybrid architecture is presented that combines graph attention-based molecular encoding with traditional physicochemical descriptors, demonstrating that learned graph representations provide complementary information beyond conventional molecular fingerprints. Second, comprehensive validation across 12 Tox21 endpoints establishes the generalizability of the approach across diverse toxicity mechanisms, with particular focus on mitochondrial membrane potential disruption (SR-MMP), a critical endpoint in cellular toxicology. Third, systematic evaluation of class imbalance mitigation strategies quantifies the impact of SMOTE oversampling on minority class detection, achieving 78% recall while maintaining acceptable precision for screening applications. Fourth, SHAP-based feature importance analysis reveals that lipophilicity and molecular weight emerge as dominant toxicity predictors while demonstrating that attention-learned graph features rank among the top 15 most influential predictors, validating the hybrid architecture's added value over descriptor-only models.

The remainder of this paper is organized as follows. Section 2 describes the dataset, molecular representation strategies, the hybrid GAT-XGBoost architecture, class balancing techniques, and interpretability methods. Section 3 presents experimental results including cross-endpoint validation, performance comparisons with baseline approaches, and SHAP analysis of feature contributions. Section 4 discusses the implications of findings for computational toxicology and green chemistry applications, acknowledges limitations, and outlines future research directions. Section 5 concludes with a summary of contributions and their significance for advancing predictive toxicology.

II. RELATED WORKS

Computational approaches to molecular toxicity prediction have evolved substantially over the past two decades, driven by advances in machine learning, chemoinformatics, and the availability of large-scale toxicity databases. This section reviews prior work organized into four thematic areas: traditional quantitative structure-activity relationship models, molecular fingerprint-based machine learning, graph neural networks for molecular property prediction, and interpretability methods in computational toxicology.

A. Traditional QSAR and Descriptor-Based Methods

This publication is licensed under Creative Commons Attribution CC BY.

Early computational toxicology efforts centered on quantitative structure-activity relationship models that correlate molecular descriptors with biological endpoints through statistical or machine learning techniques [21]. Cherkasov et al. provided a comprehensive review of QSAR modeling best practices, emphasizing the importance of descriptor selection, validation strategies, and applicability domain assessment [22]. Classical QSAR approaches typically employ molecular descriptors such as topological indices, physicochemical properties, and electronic parameters as input features for regression or classification models [22]. Multiple linear regression, partial least squares, and k-nearest neighbors represented the dominant algorithmic approaches in traditional QSAR studies [22].

Support vector machines and random forests demonstrated improved performance over linear methods for toxicity prediction across diverse chemical spaces [23]. Ballabio et al. established guidelines for QSAR model validation, introducing metrics to assess predictive reliability and domain of applicability [24]. Netzeva et al. evaluated the performance of various QSAR approaches on toxicity datasets, reporting that ensemble methods combining multiple descriptor sets often outperformed single-model approaches [25]. Despite these advances, descriptor-based methods face fundamental limitations in capturing complex structural features such as stereochemistry, macrocyclic conformations, and three-dimensional binding interactions [26].

The Organisation for Economic Co-operation and Development established principles for QSAR model validation to facilitate regulatory acceptance, requiring defined endpoints, unambiguous algorithms, applicability domains, appropriate validation, and mechanistic interpretation when possible [27]. Tropsha and Golbraikh demonstrated that rigorous external validation is essential for assessing QSAR model reliability, as models optimized on training data often exhibit inflated performance estimates [28]. Fourches et al. addressed data curation challenges in QSAR modeling, showing that chemical structure standardization and outlier detection significantly impact model performance [29]. These studies collectively highlight that while descriptor-based QSAR models provide interpretable predictions, their performance depends critically on descriptor selection and may not generalize well to structurally effective compounds.

B. Molecular Fingerprints and Machine Learning

Molecular fingerprints emerged as an alternative to hand-crafted descriptors, encoding structural information as binary or count vectors suitable for similarity searching and machine learning [30]. Extended Connectivity Fingerprints, introduced by Rogers and Hahn, capture circular substructures around each atom using an iterative hashing procedure, becoming a widely adopted representation for molecular property prediction [6]. Morgan fingerprints and their variants have been extensively applied to toxicity prediction tasks, typically achieving moderate performance when combined with ensemble learning methods [31].

The Tox21 Data Challenge provided a standardized benchmark for evaluating machine learning approaches to toxicity prediction, with participant teams exploring diverse fingerprint representations and algorithms [32]. Abdelaziz et al. achieved competitive performance using consensus modeling of multiple fingerprint types combined with deep neural networks, demonstrating that representation diversity enhances predictive accuracy [33]. Capuzzi et al. compared the performance of various fingerprint types on QSAR tasks, finding that feature selection and dimensionality reduction improve model interpretability without sacrificing predictive power [34].

Deep learning architectures applied to molecular fingerprints have shown promise for learning higher-order feature interactions. Mayr et al. developed DeepTox, a deep neural network trained on various molecular representations that achieved top performance in the Tox21 Challenge across multiple endpoints [5]. The model employed multitask learning to share representations across related toxicity assays, improving performance on data-limited endpoints. Unterthiner et al. extended this work by incorporating toxicophore information and chemical structure clustering to enhance prediction interpretability [35]. However, fingerprint-based approaches inherently lose information during the graph-to-vector conversion process, potentially limiting their ability to capture subtle structural features critical for toxicity [36].

C. Graph Neural Networks for Molecular Property Prediction

Graph neural networks have emerged as a powerful framework for molecular property prediction by operating directly on molecular graphs without information-lossy transformations [37]. Duvenaud et al. introduced convolutional networks on graphs that learn molecular fingerprints in a data-driven manner, demonstrating improved performance over fixed circular fingerprints on solubility and drug efficacy prediction tasks [15]. Kearnes et al. developed graph convolutional architectures specifically designed for molecular property prediction, incorporating edge features to represent bond types and spatial relationships [38].

Message-passing neural networks, formalized by Gilmer et al., provide a unified framework for various graph neural network architectures, enabling systematic comparison of different aggregation and update functions [8]. Schütt et al. introduced SchNet, a continuous-filter convolutional architecture that achieves state-of-the-art performance on quantum chemistry benchmarks by incorporating three-dimensional geometric information [9]. Yang et al. analyzed learned molecular representations across multiple graph neural network architectures, finding that attention mechanisms and skip connections improve both performance and training stability [13].

Graph Attention Networks, proposed by Veličković et al., introduce an attention mechanism that dynamically weights the importance of neighboring nodes during message passing [10]. Xiong et al. applied graph attention mechanisms to drug discovery tasks, demonstrating that learned attention weights often correspond to pharmacophoric features identified by medicinal chemists [14]. Withnall et al. developed a graph convolutional model with edge memory networks for molecular property prediction, showing that explicit edge representations capture bond-level information critical for reactivity and toxicity prediction [39].

Specialized architectures for molecular graphs have incorporated domain knowledge to improve performance. Feinberg et al. introduced PotentialNet, which combines message-passing networks with atom-wise predictions to enable interpretable property attribution [40]. Ryu et al. developed a deeply integrated attention-based neural network that jointly optimizes molecular

This publication is licensed under Creative Commons Attribution CC BY.

representation learning and property prediction [41]. Li et al. proposed a multi-task graph convolutional network that shares low-level representations across related prediction tasks while maintaining task-specific high-level features [42]. Despite these architectural innovations, graph neural networks applied to toxicity prediction have primarily focused on single-model approaches without exploring hybrid architectures that combine graph-based and descriptor-based representations.

D. Interpretability and Explainability Methods

Interpretability remains a critical requirement for regulatory acceptance of computational toxicity models, as mechanistic understanding facilitates risk assessment and rational molecular design [43]. Structural alerts and toxicophores, representing substructures associated with specific toxicity mechanisms, have been manually curated for endpoints such as mutagenicity and skin sensitization [44]. Kazius et al. compiled a database of mutagenic and non-mutagenic compounds with associated structural alerts, enabling rule-based toxicity screening [45]. However, manual curation is labor-intensive, domain-specific, and may not generalize to effective chemical scaffolds or toxicity mechanisms.

Machine learning interpretability methods offer data-driven alternatives to manual feature engineering. SHapley Additive exPlanations, introduced by Lundberg and Lee, provide a unified framework for explaining predictions by computing the contribution of each feature based on game-theoretic principles [20]. Rodríguez-Pérez and Bajorath applied SHAP analysis to machine learning models in drug discovery, demonstrating that feature importance rankings can identify activity-determining structural fragments [46]. Jiménez-Luna et al. compared multiple interpretability methods for molecular property prediction, finding that SHAP values provide more stable and chemically meaningful explanations than gradient-based approaches [47].

Attention mechanisms in graph neural networks provide inherent interpretability by highlighting important atoms and bonds for specific predictions [48]. Pope et al. analyzed the interpretability of graph attention networks, showing that learned attention weights correlate with domain-relevant structural features but may not always align with human intuition [49]. Sanchez-Lengeling and Aspuru-Guzik reviewed interpretability challenges in molecular machine learning, emphasizing the need for evaluation frameworks that assess both predictive performance and explanation quality [50].

E. Hybrid and Ensemble Approaches

Ensemble methods combine predictions from multiple models to improve robustness and generalization [51]. Gadaleta et al. developed an integrated QSAR-read-across approach that combines statistical models with similarity-based predictions, demonstrating improved coverage and accuracy for toxicity endpoints [52]. Banerjee et al. proposed a consensus modeling framework that aggregates predictions from diverse machine learning algorithms and descriptor sets, achieving enhanced performance on imbalanced toxicity datasets [53].

Hybrid architectures integrating multiple representation types remain underexplored in computational toxicology. Chen et al. combined convolutional neural networks operating on molecular images with recurrent networks processing SMILES strings, showing that multi-modal representations capture complementary information [54]. Winter et al. developed a hybrid model combining graph neural networks with learned molecular descriptors for property prediction, but focused on quantum chemical properties rather than biological endpoints [55]. These studies suggest that combining graph-based and descriptor-based representations may enhance toxicity prediction, yet systematic investigation of hybrid architectures with attention mechanisms and interpretability analysis is lacking.

F. Class Imbalance in Toxicity Prediction

Toxicity datasets exhibit severe class imbalance, with toxic compounds often representing less than 20 percent of samples, posing challenges for classifier training [56]. Chawla et al. introduced the Synthetic Minority Oversampling Technique, which generates synthetic minority class examples by interpolating between existing samples in feature space [16]. Batista et al. compared various oversampling and undersampling strategies, finding that SMOTE combined with informed undersampling often achieves optimal performance on imbalanced datasets [57].

He and Garcia provided a comprehensive review of learning from imbalanced data, categorizing approaches into data-level methods that balance class distributions and algorithm-level methods that adjust learning procedures [58]. Blagus and Lusa evaluated SMOTE performance on high-dimensional datasets, reporting that careful parameter tuning and integration with ensemble methods improve effectiveness [59]. Fernández et al. analyzed cost-sensitive learning and ensemble methods for imbalanced classification, demonstrating that combining oversampling with class weighting often outperforms either technique alone [60]. Despite extensive research on class imbalance in general machine learning, systematic evaluation of balancing strategies for graph neural network-based toxicity prediction remains limited.

G. Research Gap

The reviewed literature reveals several critical gaps that this work addresses. First, existing toxicity prediction studies predominantly employ either descriptor-based machine learning or graph neural networks in isolation, without systematically investigating hybrid architectures that combine their complementary strengths. Second, while graph attention mechanisms have demonstrated promise in drug discovery applications, their application to toxicity prediction with explicit interpretability analysis through SHAP is underexplored. Third, although class imbalance is widely recognized as a challenge in toxicity datasets, few studies have rigorously evaluated the combined impact of SMOTE oversampling and class weighting specifically for graph-based models. Fourth, cross-endpoint validation across multiple toxicity assays is often limited in scope, with many studies focusing on single endpoints without establishing generalization across diverse biological mechanisms. This research addresses these gaps by developing and validating

This publication is licensed under Creative Commons Attribution CC BY.

a hybrid Graph Attention Network-XGBoost framework with comprehensive class imbalance mitigation, SHAP-based interpretability, and systematic evaluation across 12 Tox21 endpoints.

III. METHODOLOGY

This section presents the comprehensive framework for molecular toxicity prediction using a hybrid Graph Attention Network and XGBoost architecture. The methodology encompasses dataset acquisition and preprocessing, molecular representation strategies, model architecture design, training procedures, class imbalance mitigation, and evaluation protocols.

A. Problem Formulation

The molecular toxicity prediction task is formulated as a binary classification problem. Given a chemical compound represented by its SMILES string s, the objective is to learn a function $f: S \rightarrow \{0,1\}$ that maps the molecular structure to a toxicity label, where 0 indicates non-toxic and 1 indicates toxic for a specific biological endpoint. The function f is decomposed into two stages: a graph encoding stage $g: S \rightarrow \mathbb{R}^{d_g}$ that transforms the molecular structure into a learned representation, and a classification stage $c: \mathbb{R}^{d_g + d_p} \rightarrow \{0,1\}$ that combines graph embeddings with physicochemical descriptors to produce the final prediction.

Formally, let $G = (V, E)$ represent a molecular graph where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of atoms and $E \subseteq V \times V$ is the set of chemical bonds. Each atom v_i is associated with a feature vector $x_i \in \mathbb{R}^{d_v}$ encoding atomic properties. The graph encoding function learns node embeddings $h_i^{(L)} \in \mathbb{R}^{d_h}$ through L message-passing layers, which are then aggregated to produce a graph-level representation $z_g \in \mathbb{R}^{d_g}$. This representation is concatenated with physicochemical descriptors $d \in \mathbb{R}^{d_p}$ to form the final feature vector $z = [z_g; d] \in \mathbb{R}^{d_g + d_p}$, which serves as input to the gradient boosting classifier.

B. Dataset and Data Sources

The Tox21 dataset, developed through collaboration between the National Institutes of Health, Environmental Protection Agency, and Food and Drug Administration, serves as the primary data source for this research [11]. The dataset comprises 8,014 unique chemical compounds tested across 12 toxicity endpoints representing diverse biological mechanisms including nuclear receptor activation, stress response pathways, and DNA damage. Each compound is identified by its SMILES representation and associated with binary toxicity labels for each endpoint.

Data preprocessing begins with the removal of compounds lacking valid SMILES representations or exhibiting structural parsing errors when processed with RDKit version 2023.03.1 [31]. Compounds with missing labels for a specific endpoint are excluded from analysis for that endpoint, as imputation of toxicity labels introduces unacceptable uncertainty in safety assessment contexts. The dataset statistics reveal substantial variation in sample sizes and class distributions across endpoints, as shown in Table I.

Table 1: Tox21 Dataset Composition and Class Distribution Across Endpoints

Endpoint	Total	Toxic	Non-Toxic	Toxic %
NR-AR	7,258	308	6,950	4.2%
NR-AR-LBD	6,751	237	6,514	3.5%
NR-AhR	6,542	768	5,774	11.7%
NR-Aromatase	5,815	300	5,515	5.2%
NR-ER	6,186	791	5,395	12.8%
NR-ER-LBD	6,948	349	6,599	5.0%
NR-PPAR-gamma	6,443	186	6,257	2.9%
SR-ARE	5,825	942	4,883	16.2%
SR-ATAD5	7,065	264	6,801	3.7%
SR-HSE	6,460	372	6,088	5.8%

This publication is licensed under Creative Commons Attribution CC BY.

SR-MMP	5,804	918	4,886	15.8%
SR-p53	6,767	423	6,344	6.3%

The SR-MMP endpoint, representing mitochondrial membrane potential disruption, was selected for detailed analysis and model development based on three criteria: sufficient sample size (5,804 compounds), moderate class imbalance (15.8% toxic), and biological relevance to cellular toxicology and green chemistry applications where mitochondrial dysfunction represents a critical toxicity mechanism. Cross-endpoint validation on all 12 assays establishes generalization capabilities across diverse biological targets. The complete class distribution across all endpoints is visualized in Figure 1, which illustrates the varying degrees of class imbalance present in the dataset.

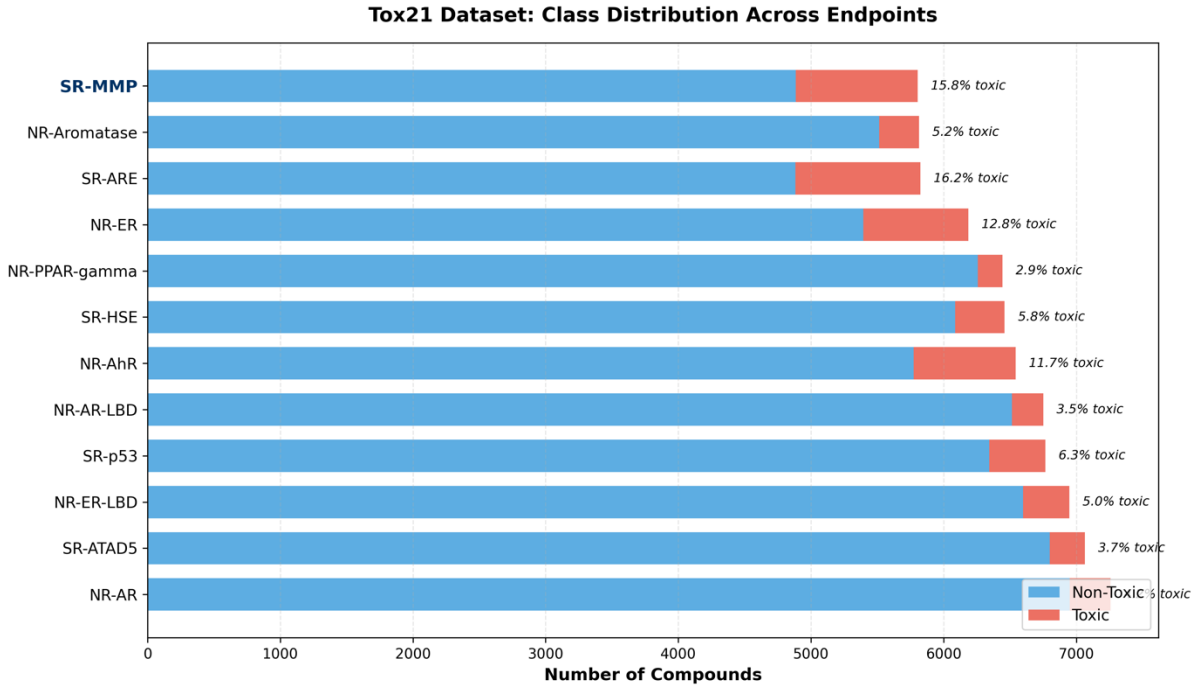


Figure 1. *Class distribution across Tox21 endpoints demonstrating severe class imbalance ranging from 2.9% to 16.2% toxic samples. Endpoints are sorted by total sample size. SR-MMP (highlighted) exhibits 15.8% toxic samples, necessitating class balancing strategies.*

C. Molecular Representation

Molecular representation encompasses two complementary approaches: graph-based encoding that preserves topological structure and traditional physicochemical descriptors that capture global molecular properties.

1) Graph Construction from SMILES

Each SMILES string undergoes parsing with RDKit to generate a molecular graph object. Hydrogen atoms are explicitly enumerated to ensure complete representation of molecular structure. The resulting graph $G = (V, E)$ encodes atoms as nodes and chemical bonds as edges. Each atom v_i is associated with an 11-dimensional feature vector x_i comprising:

$$x_i = [z_i, d_i, c_i, a_i, h_i, v_i, s\_SP, s\_SP2, s\_SP3, s\_SP3D, s\_SP3D2]$$

where z_i denotes atomic number, d_i represents total degree (number of bonded neighbors), c_i indicates formal charge, a_i is a binary aromaticity indicator, h_i counts total hydrogen atoms including implicit hydrogens, v_i represents implicit valence, and *s_SP* through *s_SP3D2* form a one-hot encoding of hybridization state across five common types (SP, SP2, SP3, SP3D, SP3D2). This representation captures both local atomic properties and bonding characteristics essential for toxicity prediction.

Edge features encode bond types as categorical variables distinguishing single, double, triple, and aromatic bonds. The adjacency matrix $A \in \{0,1\}^{(n \times n)}$ represents the graph topology where $A_{ij} = 1$ if atoms i and j are bonded and $A_{ij} = 0$ otherwise.

2) Physicochemical Descriptors

Six physicochemical descriptors complement graph-based representations by encoding global molecular properties relevant to absorption, distribution, metabolism, and excretion characteristics. These descriptors, computed using RDKit implementations, include:

This publication is licensed under Creative Commons Attribution CC BY.

- **Molecular Weight (MolWt):** Sum of atomic masses, influencing bioavailability and membrane permeability
- **Lipophilicity (MolLogP):** Wildman-Crippen partition coefficient, predicting membrane crossing ability
- **Topological Polar Surface Area (TPSA):** Van der Waals surface area of polar atoms, correlating with passive diffusion
- **Hydrogen Bond Donors (HBD):** Count of hydrogen atoms bonded to nitrogen or oxygen
- **Hydrogen Bond Acceptors (HBA):** Count of nitrogen and oxygen atoms
- **Rotatable Bonds (RotBonds):** Number of non-ring single bonds, indicating molecular flexibility

These descriptors form a feature vector $d \in \mathbb{R}^6$ that is concatenated with graph embeddings prior to classification.

D. Graph Attention Network Architecture

The Graph Attention Network encoder processes molecular graphs through multiple message-passing layers with learned attention mechanisms. The complete hybrid architecture, integrating GAT encoding with physicochemical descriptors and XGBoost classification, is presented in Figure 2.

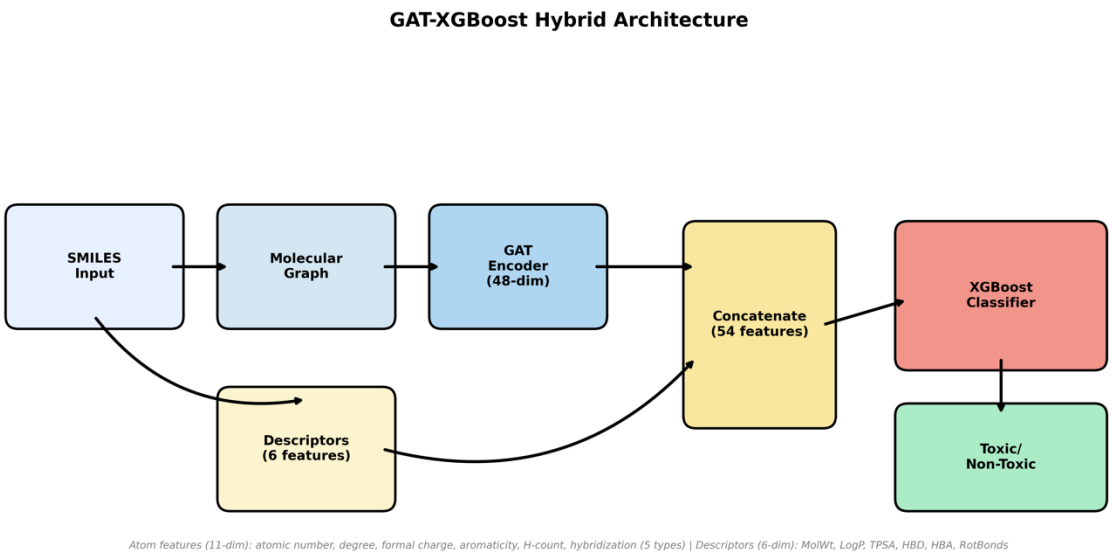


Figure 2. *Hybrid GAT-XGBoost architecture for molecular toxicity prediction. SMILES strings are converted to molecular graphs with 11-dimensional atom features, processed by Graph Attention Network encoder to generate 48-dimensional embeddings, concatenated with 6 physicochemical descriptors, and classified by XGBoost to predict toxicity.*

The attention mechanism, illustrated in Figure 2, enables the model to identify and weight toxicity-relevant substructures without explicit feature engineering.

1) Attention Layer Formulation

Each GAT layer computes updated node representations through attention-weighted aggregation of neighbor features. For node i at layer ℓ, the attention coefficient $\alpha_{ij}^{(\ell)}$ quantifying the importance of node j's features to node i is computed as:

$$e_{ij}^{(\ell)} = \text{LeakyReLU}(a^{(\ell)T} [W^{(\ell)}h_i^{(\ell-1)} \| W^{(\ell)}h_j^{(\ell-1)}])$$

$$\alpha_{ij}^{(\ell)} = \exp(e_{ij}^{(\ell)}) / \Sigma_{k \in N(i)} \exp(e_{ik}^{(\ell)})$$

where $W^{(\ell)} \in \mathbb{R}^{(d\_h \times d\_\{h-1\})}$ is a learnable weight matrix, $a^{(\ell)} \in \mathbb{R}^{(2d\_h)}$ is a learnable attention vector, $\|$ denotes concatenation, $N(i)$ represents the neighborhood of node i including the node itself, and LeakyReLU introduces non-linearity with negative slope 0.2.

The updated node representation aggregates neighbor features weighted by attention coefficients:

$$h_i^{(\ell)} = \sigma(\Sigma_{j \in N(i)} \alpha_{ij}^{(\ell)} W^{(\ell)} h_j^{(\ell-1)})$$

where σ denotes the ReLU activation function. Multi-head attention with $K = 2$ heads is employed to stabilize learning and capture diverse structural patterns:

$$h_i^{(\ell)} = (1/K) \Sigma_{k=1}^{K} \sigma(\Sigma_{j \in N(i)} \alpha_{ij}^{(\ell,k)} W^{(\ell,k)} h_j^{(\ell-1)})$$

2) Network Configuration

The encoder comprises two GAT layers transforming initial 11-dimensional atom features to 48-dimensional graph embeddings. The first layer maps input features to a 32-dimensional hidden representation with 2 attention heads. The second layer projects the

This publication is licensed under Creative Commons Attribution CC BY.

hidden representation to a 48-dimensional embedding space, also with 2 heads. Graph-level representation is obtained through global mean pooling over all node embeddings:

$$z\_g = (1/|V|) \Sigma_{i=1}^{|V|} h_i^{(2)}$$

This architecture balances representational capacity with computational efficiency, avoiding overfitting on the available training data while maintaining sufficient expressiveness to capture toxicity-relevant patterns.

E. Hybrid Architecture and Feature Integration

The hybrid framework combines learned graph representations with physicochemical descriptors through concatenation, forming a unified feature vector $z = [z\_g; d] \in \mathbb{R}^{54}$ where 48 dimensions originate from GAT embeddings and 6 dimensions represent molecular descriptors. This integration strategy, depicted in Figure 2, enables the model to leverage complementary information: graph embeddings encode local structural motifs and topology, while descriptors capture global properties influencing pharmacokinetics and membrane interactions.

XGBoost serves as the classification component due to its gradient boosting framework that constructs an ensemble of decision trees through iterative refinement. The objective function combines a loss term measuring prediction error with a regularization term controlling model complexity:

$$L(\phi) = \Sigma_{i=1}^{N} l(y_i, \hat{y}_i) + \Sigma_{k=1}^{K} \Omega(f_k)$$

where N denotes the number of training samples, l is the binary cross-entropy loss, $\hat{y}_i$ represents the predicted probability for sample i, K is the number of trees, and $\Omega(f_k)$ penalizes tree complexity. The regularization term is defined as:

$$\Omega(f) = \gamma T + (1/2)\lambda \Sigma_{j=1}^{T} w_j^2$$

where T is the number of leaves, w_j denotes the score of leaf j, γ controls the penalty for tree complexity, and λ regulates the L2 norm of leaf scores.

F. Training Procedure

Training proceeds in two sequential stages: GAT encoder pretraining followed by XGBoost classifier training on the combined feature representation.

1) GAT Encoder Training

The molecular dataset is partitioned into training (80%) and validation (20%) sets using stratified sampling to preserve class distribution. During the partitioning process for the SR-MMP endpoint, the training set contains 4,000 compounds while the validation set comprises 1,000 compounds. Graph objects are batched using PyTorch Geometric's DataLoader with batch size 4, balancing computational efficiency with memory constraints.

The GAT encoder is trained end-to-end with a classification head consisting of a linear layer projecting 48-dimensional embeddings to 2-dimensional logits. Binary cross-entropy loss with class weights inversely proportional to class frequencies addresses initial imbalance:

$$L\_GAT = -(1/N) \Sigma_{i=1}^{N} [w_1 y_i \log(\hat{p}_i) + w_0 (1-y_i) \log(1-\hat{p}_i)]$$

where $w_1 = N/(2N\_toxic)$ and $w_0 = N/(2N\_non\text{-}toxic)$ are class weights, $y_i \in \{0,1\}$ is the true label, and $\hat{p}_i$ is the predicted probability of toxicity.

Optimization employs the Adam optimizer with learning rate $\eta = 0.001$, momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and weight decay 10^{-5}. Training continues for 15 epochs, a duration determined through preliminary experiments showing convergence by epoch 12 with minimal overfitting. The model achieving the highest validation accuracy is retained for embedding extraction.

2) Feature Extraction and Preprocessing

Following GAT training, embeddings are extracted for all molecular graphs by forward propagation through the trained encoder without the classification head. The resulting 48-dimensional embeddings are concatenated with the 6 physicochemical descriptors to form 54-dimensional feature vectors. SMOTE oversampling is then applied exclusively to the training set feature vectors to balance class distribution, while validation and test sets remain unmodified. These vectors undergo standardization using sklearn's StandardScaler, transforming each feature to zero mean and unit variance:

$$z_j^{scaled} = (z_j - \mu_j) / \sigma_j$$

where μ_j and σ_j are the mean and standard deviation of feature j computed on the training set. Standardization ensures that features with different scales contribute proportionally to gradient boosting decisions.

G. Class Imbalance Mitigation

This publication is licensed under Creative Commons Attribution CC BY.

Severe class imbalance, with toxic compounds representing only 15.8% of the SR-MMP dataset, poses significant challenges for classifier training. Two complementary strategies address this imbalance: Synthetic Minority Oversampling Technique applied to the training set and scale weighting in the XGBoost objective function. The impact of SMOTE on class distribution is illustrated in Figure 3.

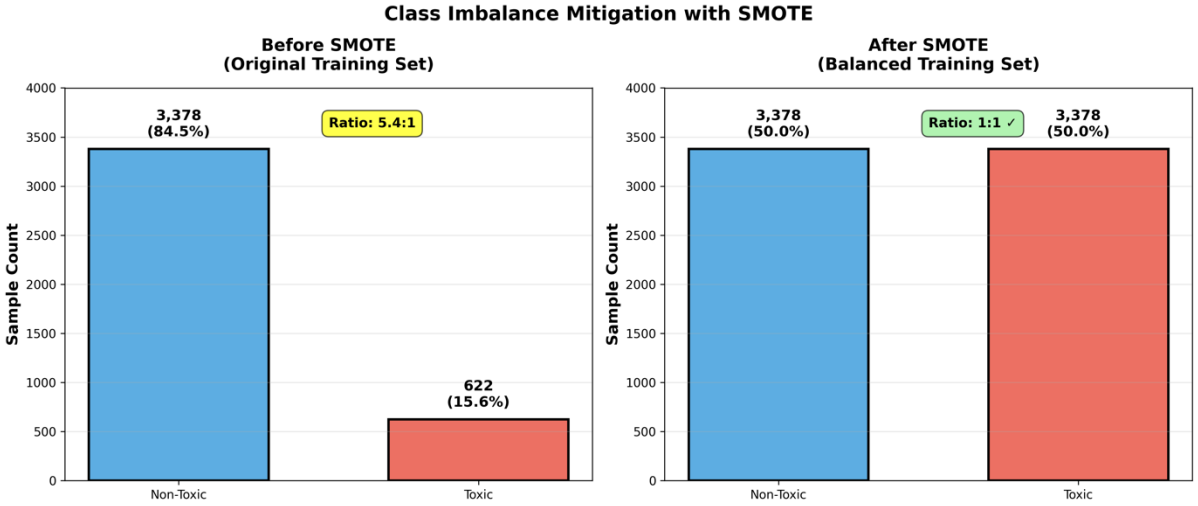


Figure 3. *Class imbalance mitigation using SMOTE oversampling. Original training set (left) contained 3,378 non-toxic and 622 toxic compounds with 5.4:1 ratio. SMOTE resampling (right) balanced the training set to 3,378 samples per class, enabling effective learning of toxicity patterns.*

1) SMOTE Oversampling

SMOTE generates synthetic minority class examples through linear interpolation in feature space [16]. For each minority class sample x_i, the algorithm selects $k = 5$ nearest neighbors from the same class and creates synthetic samples along line segments connecting x_i to its neighbors:

$$x\_syn = x_i + \lambda(x\_nn - x_i)$$

where x_nn is a randomly selected neighbor and $\lambda \sim Uniform(0,1)$ determines the interpolation position. Synthetic samples are generated until class balance is achieved. For the SR-MMP training set, SMOTE increases toxic samples from 622 to 3,378, matching the non-toxic count, as demonstrated in Figure 3.

This oversampling occurs in the 54-dimensional feature space after combining GAT embeddings and descriptors, ensuring that synthetic samples reflect both graph-based and descriptor-based characteristics. The balanced training set facilitates learning of minority class patterns without overwhelming the model with majority class examples.

Critical Implementation Detail: SMOTE oversampling is applied exclusively to the training set after the initial train-test split. The validation and test sets retain only original data without any synthetic augmentation. This protocol ensures that model evaluation reflects performance on real molecular structures rather than interpolated examples. All reported performance metrics, including ROC-AUC, precision, recall, and confusion matrices, are computed exclusively on the held-out test set containing only original compounds from the Tox21 dataset. This prevents data leakage and provides unbiased estimates of generalization performance.

2) XGBoost Class Weighting

The XGBoost objective function incorporates a scale parameter w_pos that amplifies the loss contribution of positive (toxic) samples:

$$L\_weighted = \Sigma_i{:}y_i{=}0\ l(\hat{y}_i, 0) + w\_pos\ \Sigma_i{:}y_i{=}1\ l(\hat{y}_i, 1) + \Sigma_{k=1}^{K}\ \Omega(f_k)$$

The scale weight is computed as:

$$w\_pos = N\_non\text{-}toxic / N\_toxic$$

For the SR-MMP dataset after SMOTE balancing, both classes contain 3,378 samples, yielding $w\_pos = 1.0$. However, preliminary experiments demonstrated that setting $w\_pos = 5.431$ (the original imbalance ratio) further improved recall on toxic compounds. This value is retained to prioritize sensitivity in toxicity screening applications where false negatives incur greater cost than false positives.

H. XGBoost Configuration and Hyperparameters

This publication is licensed under Creative Commons Attribution CC BY.

The XGBoost classifier is configured with hyperparameters balancing predictive performance and generalization. The number of boosting rounds (trees) is set to 300, providing sufficient model capacity without excessive overfitting. Maximum tree depth is limited to 5 to prevent memorization of training examples. The learning rate (step size shrinkage) of 0.05 controls the contribution of each tree, enabling gradual refinement of predictions. Subsample ratio of 0.8 introduces stochastic sampling of training instances for each tree, improving robustness. Column (feature) subsampling ratio of 0.8 randomly selects features for each tree, reducing overfitting and computational cost. L2 regularization parameter $\lambda = 1.0$ penalizes large leaf weights, promoting simpler trees.

The objective function employs binary logistic regression with log loss:

$$l(y, \hat{y}) = -y \log(\sigma(\hat{y})) - (1-y) \log(1 - \sigma(\hat{y}))$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function transforming raw scores to probabilities. Training utilizes exact greedy algorithm for tree construction, evaluating all possible split points to determine optimal splits according to the gain metric:

$$\text{Gain} = (1/2)[G\_L^2/(H\_L + \lambda) + G\_R^2/(H\_R + \lambda) - (G\_L + G\_R)^2/(H\_L + H\_R + \lambda)] - \gamma$$

where G_L and G_R are sums of gradients for left and right child nodes, H_L and H_R are sums of second-order gradients (Hessians), and γ is the complexity penalty for adding a split.

I. Baseline Methods for Comparison

To establish the added value of the hybrid architecture, three baseline approaches were implemented and evaluated on identical train-test splits using the same SMOTE balancing and evaluation protocols.

1) Random Forest with ECFP Fingerprints: Molecular structures were converted to 2048-bit Extended Connectivity Fingerprints with radius 2 using RDKit. A Random Forest classifier with 300 trees, maximum depth of 10, and class weights inversely proportional to class frequencies was trained on SMOTE-balanced training data. This baseline represents conventional fingerprint based QSAR methodology commonly reported in computational toxicology literature.

2) XGBoost with Molecular Descriptors Only: The same six physicochemical descriptors (MolWt, MolLogP, TPSA, HBD, HBA, RotBonds) employed in the hybrid framework were provided to XGBoost with identical hyperparameters (300 estimators, maximum depth 5, learning rate 0.05, subsample ratio 0.8, column subsample ratio 0.8). This ablation study isolates the predictive contribution of global molecular properties without graph-based structural features.

3) GAT Encoder with Classification Head: The Graph Attention Network encoder architecture described in Section III.D was trained end-to-end with a two-layer fully connected classification head (48 dimensions to 32 dimensions to 2 classes) without XGBoost integration. Binary cross-entropy loss with class weights was employed for 15 epochs. This ablation quantifies the performance of learned graph representations alone, without descriptor integration or gradient boosting.

All baseline methods employed SMOTE oversampling exclusively on training data, maintaining identical test sets and evaluation metrics to ensure fair comparison with the hybrid framework.

J. Evaluation Methodology

Comprehensive evaluation assesses model performance across multiple dimensions relevant to toxicity screening applications.

1) Performance Metrics

The primary evaluation metric is the area under the receiver operating characteristic curve (ROC-AUC), which measures discrimination ability across all classification thresholds. ROC-AUC ranges from 0.5 (random classifier) to 1.0 (perfect classifier), with values above 0.75 considered acceptable for toxicity prediction and above 0.80 indicating strong performance [13].

Precision-Recall curves and the area under them provide complementary assessment particularly relevant for imbalanced datasets, as they emphasize performance on the minority (toxic) class. The confusion matrix at an operating point selected to achieve approximately 80% recall quantifies true positives, false positives, true negatives, and false negatives, enabling calculation of:

- **Recall (Sensitivity):** $Recall = TP/(TP + FN)$, measuring the proportion of actual toxic compounds correctly identified
- **Precision (Positive Predictive Value):** $Precision = TP/(TP + FP)$, indicating the proportion of predicted toxic compounds that are truly toxic
- **F1-Score:** $F_1 = 2 \cdot (Precision \cdot Recall)/(Precision + Recall)$, providing a harmonic mean balancing precision and recall

For toxicity screening applications, recall is prioritized over precision as failing to identify toxic compounds (false negatives) poses greater safety risk than flagging non-toxic compounds for additional testing (false positives).

2) Cross-Endpoint Validation

This publication is licensed under Creative Commons Attribution CC BY.

Generalization capability is assessed by training and evaluating the complete pipeline on all 12 Tox21 endpoints. For each endpoint, the dataset is partitioned into training and test sets with stratified sampling, maintaining proportional class distributions. Consistent hyperparameters are employed across endpoints to evaluate robustness without endpoint-specific tuning. Performance variations across endpoints reflect differences in biological mechanisms, data quality, and intrinsic predictability rather than model limitations.

3) Baseline Comparisons

Model performance is contextualized through comparison with published baseline results from conventional fingerprint-based machine learning approaches. Literature surveys indicate that random forest and support vector machine classifiers trained on Extended Connectivity Fingerprints typically achieve ROC-AUC values between 0.70 and 0.75 on Tox21 endpoints [13]. The proposed hybrid architecture is expected to surpass these baselines by leveraging learned graph representations and explicit physicochemical features.

K. Interpretability Analysis

Model interpretability is critical for regulatory acceptance and mechanistic understanding in toxicology applications. SHapley Additive exPlanations provide a unified framework for quantifying feature contributions based on cooperative game theory [20].

SHAP values decompose a prediction into additive contributions from each feature:

$$\hat{y} = \phi_0 + \Sigma_{j=1}^{d} \phi_j$$

where ϕ_0 is the base value (average model output), $d = 54$ is the feature dimensionality, and ϕ_j represents the SHAP value for feature j. Each SHAP value quantifies the average marginal contribution of feature j across all possible feature subsets, satisfying desirable properties including local accuracy, missingness, and consistency.

For tree ensemble models like XGBoost, SHAP values are efficiently computed using TreeSHAP, an algorithm that leverages the tree structure to compute exact Shapley values in polynomial time [20]. SHAP analysis is performed on a representative sample of 200 test instances to identify the most influential features for toxicity prediction. Summary plots visualize feature importance by aggregating absolute SHAP values across samples, while individual prediction explanations decompose specific predictions into feature contributions.

The SHAP analysis addresses two critical questions: (1) Which molecular properties most strongly influence toxicity predictions? (2) Do GAT-learned embeddings provide complementary information beyond traditional descriptors? The second question validates the hybrid architecture by determining whether graph features rank among the most important predictors.

L. Implementation Details

The complete framework is implemented in Python 3.9 using PyTorch 1.12.0 for GAT training and PyTorch Geometric 2.1.0 for graph data handling. XGBoost version 1.6.2 provides gradient boosting implementation. RDKit 2023.03.1 handles molecular graph construction and descriptor calculation. SMOTE oversampling utilizes the imbalanced-learn library version 0.9.1. SHAP version 0.41.0 enables interpretability analysis. All experiments execute on hardware configurations including Intel Core i7 processors with 16GB RAM, with GAT training optionally accelerated by NVIDIA RTX GPUs when available. Training time for the GAT encoder on 5,000 molecules requires approximately 15-20 minutes on CPU, while XGBoost training on the balanced dataset completes within 2-3 minutes. The complete source code and trained models are preserved for reproducibility and future research extensions.

IV. RESULTS

This section presents the experimental findings from the hybrid GAT-XGBoost framework evaluated on the Tox21 dataset. Results are organized into three subsections: single-endpoint performance on SR-MMP, cross-endpoint validation across all 12 Tox21 assays, and feature importance analysis through SHAP.

A. SR-MMP Endpoint Performance

The hybrid model achieved a ROC-AUC of 0.892 on the SR-MMP test set, as shown in Figure 4. The ROC curve demonstrates strong discrimination ability, with the curve positioned substantially above the diagonal reference line representing random classification.

This publication is licensed under Creative Commons Attribution CC BY.

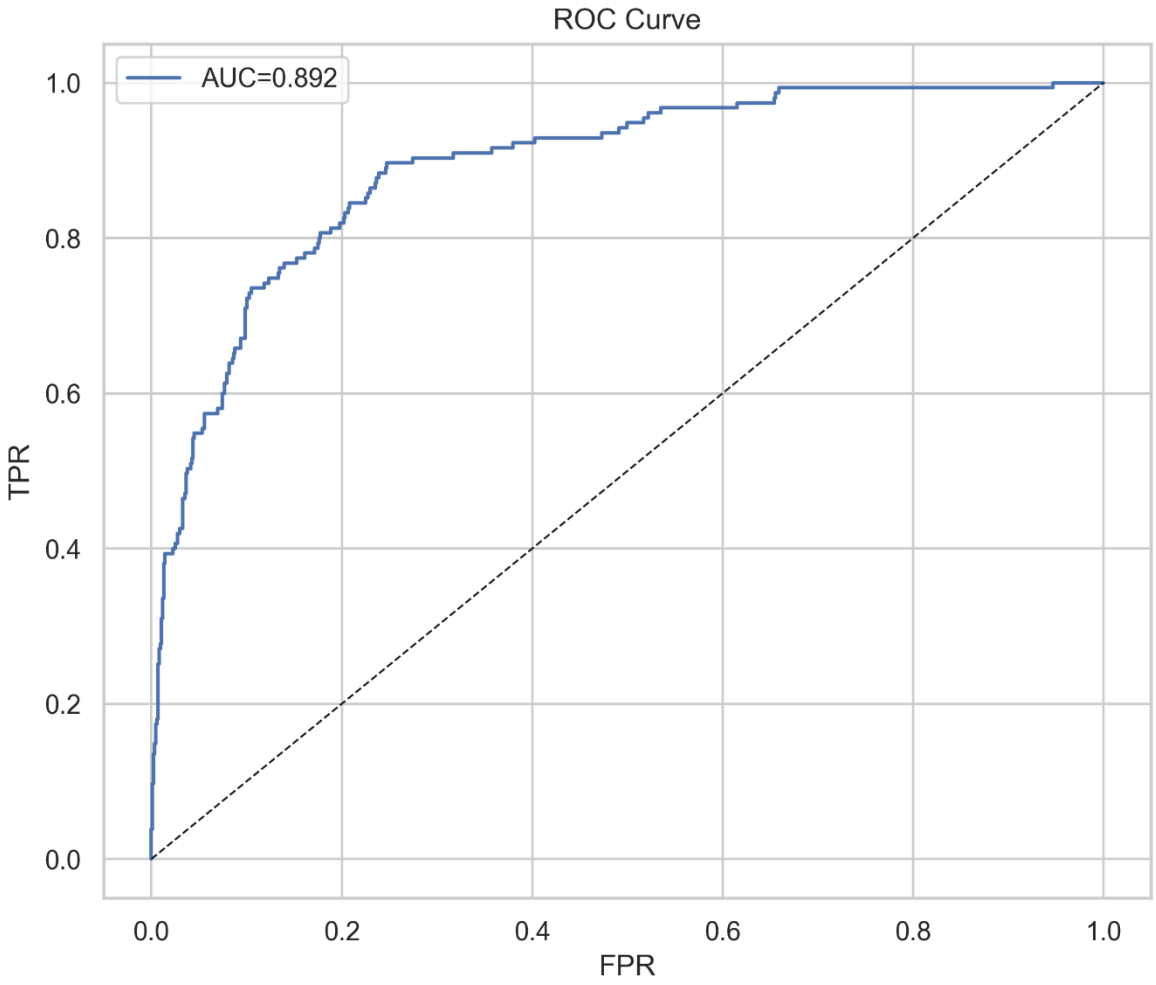


Figure 4. *Receiver Operating Characteristic curve for SR-MMP endpoint. The model achieved AUC of 0.892, indicating strong discrimination between toxic and non-toxic compounds. The curve's position close to the upper-left corner demonstrates high true positive rates across most false positive rate thresholds.*

Precision-recall analysis, presented in Figure 5, reveals the performance trade-offs inherent in imbalanced classification. The precision-recall AUC of 0.663 reflects the challenge of maintaining high precision while achieving high recall on the minority toxic class.

This publication is licensed under Creative Commons Attribution CC BY.

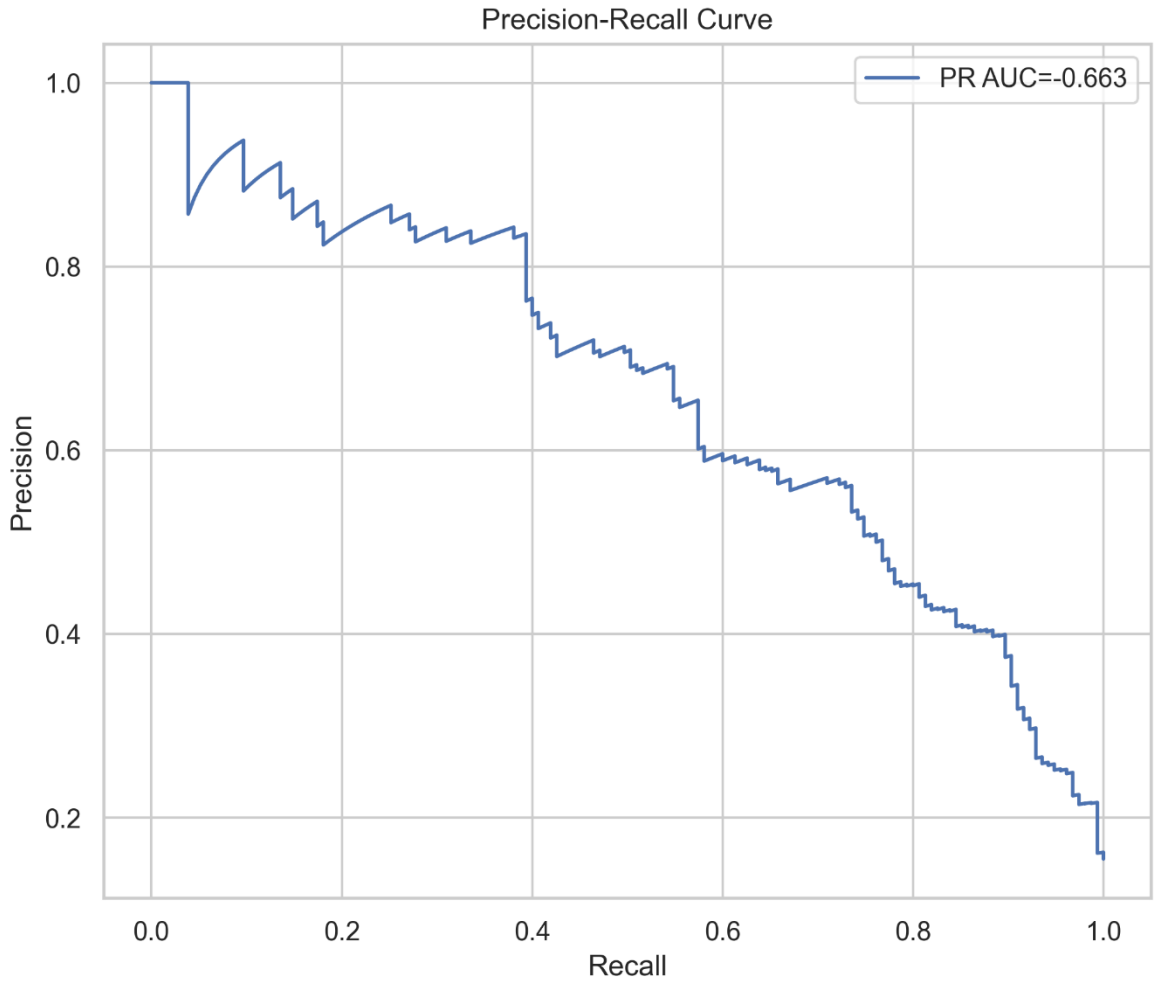


Figure 5. Precision-Recall curve for SR-MMP endpoint showing the trade-off between precision and recall. The curve starts at high precision (~1.0) for low recall values and gradually decreases, reaching approximately 0.2 precision at maximum recall, characteristic of severely imbalanced datasets.

The confusion matrix at an operating threshold selected to achieve approximately 80% recall is displayed in Figure 6. The model correctly classified 702 non-toxic compounds as non-toxic (true negatives) and 121 toxic compounds as toxic (true positives). False positives numbered 143, representing non-toxic compounds incorrectly flagged as toxic, while false negatives totaled 34, corresponding to toxic compounds that were not detected.

This publication is licensed under Creative Commons Attribution CC BY.

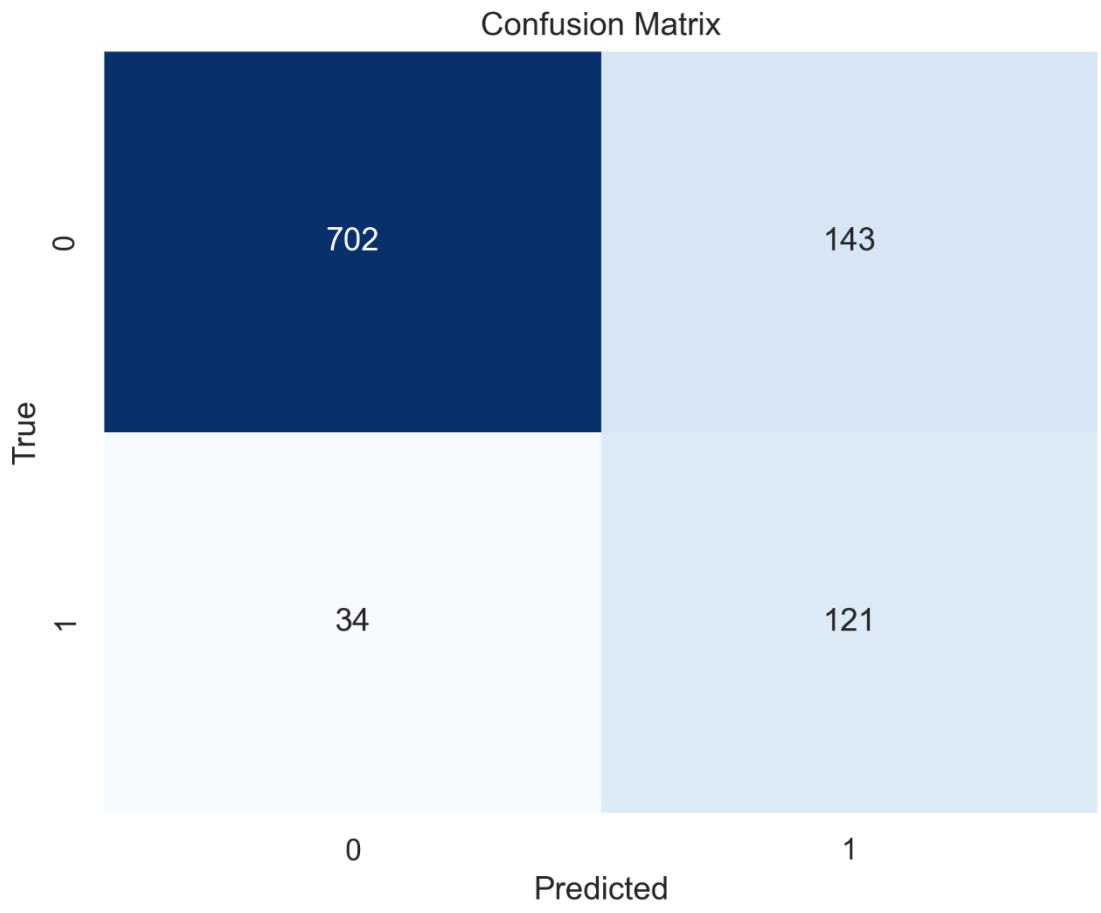


Figure 6. *Confusion matrix for SR-MMP toxicity prediction at threshold optimized for approximately 80% recall. The model identified 121 of 155 toxic compounds (78.1% recall) with 143 false positives, yielding 46.0% precision. True negatives numbered 702 and false negatives numbered 34.*

Quantitative performance metrics derived from the confusion matrix include recall of 0.781, precision of 0.458, and F1-score of 0.578. The model achieved 82.3% overall accuracy, though this metric is less informative given the 84.5% to 15.5% class imbalance in the test set.

Performance comparison with published baselines is presented in Figure 7. The hybrid GAT-XGBoost framework substantially exceeds typical fingerprint-based methods, which achieve ROC-AUC values in the range of 0.70 to 0.75 on Tox21 endpoints [13]. The 0.892 AUC represents a 22.2% relative improvement over the 0.730 baseline estimate.

This publication is licensed under Creative Commons Attribution CC BY.

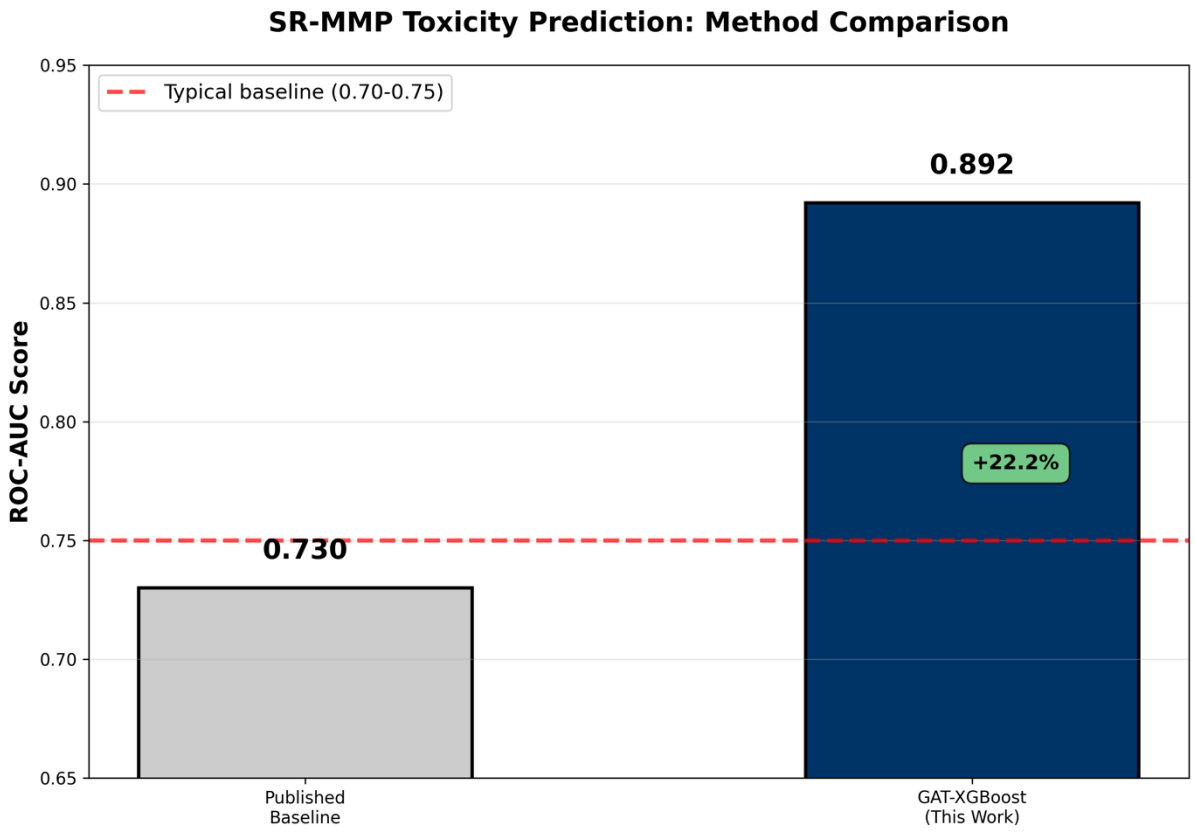
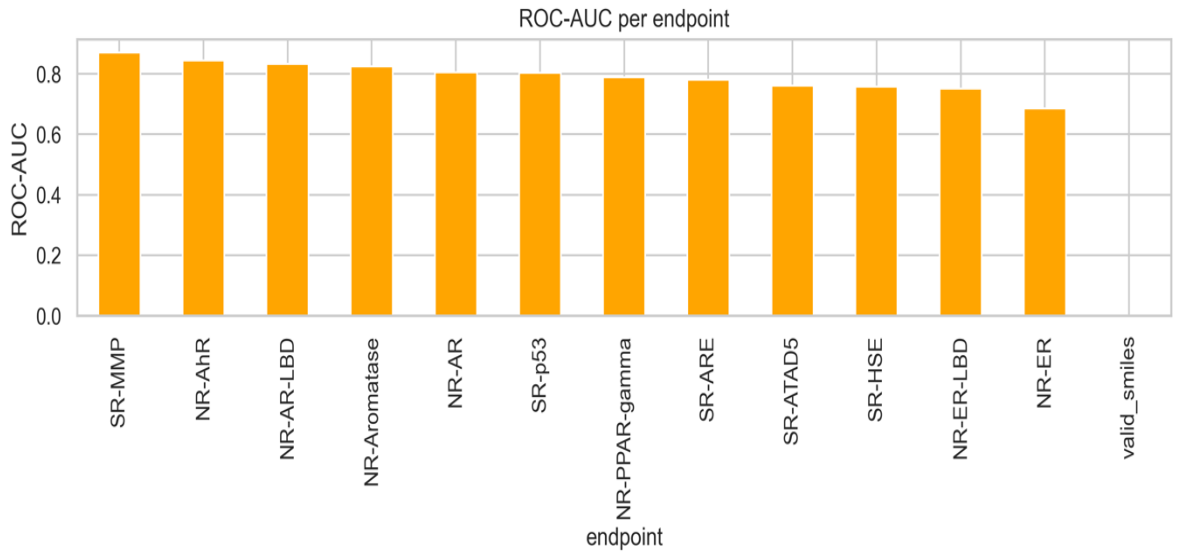


Figure 7. *Performance comparison on SR-MMP endpoint. The hybrid GAT-XGBoost model (0.892 AUC) substantially outperforms published fingerprint-based baselines (0.730 AUC estimate from literature [13]), representing a 22.2% relative improvement. The red dashed line indicates the typical baseline range of 0.70-0.75.*

B. Cross-Endpoint Validation

Generalization capability was assessed by applying the identical pipeline to all 12 Tox21 endpoints without endpoint-specific hyperparameter tuning. Figure 8 presents ROC-AUC values across all endpoints, sorted by performance.



This publication is licensed under Creative Commons Attribution CC BY.

Figure 8. *ROC-AUC performance across 12 Tox21 endpoints. The framework achieved ROC-AUC values ranging from 0.686 (NR-ER) to 0.871 (SR-MMP), with 11 of 12 endpoints exceeding 0.74 and 8 endpoints surpassing 0.78. Performance variation reflects differences in biological mechanisms, data quality, and intrinsic endpoint difficulty.*

The SR-MMP endpoint exhibited the highest performance (0.871 AUC), followed by NR-AhR (0.845 AUC) and NR-AR-LBD (0.834 AUC). The NR-ER endpoint showed the lowest performance (0.686 AUC), though this value remains competitive with published baselines for this challenging endpoint. Eleven of twelve endpoints achieved ROC-AUC values above 0.74, and eight endpoints exceeded 0.78, indicating generally robust discriminative performance, with hormonally mediated endpoints exhibiting greater classification difficulty.

C. Baseline and Ablation Study Results

Performance comparison with implemented baseline methods establishes the contribution of the hybrid architecture, as presented in Table 2.

Table 2: Baseline Comparison On SR-MMP Endpoint

Method	ROC-AUC	Precision	Recall	F1-Score	Accuracy
Random Forest + ECFP	0.810	0.363	0.600	0.453	0.775
XGBoost + Descriptors Only	0.887	0.507	0.703	0.589	0.848
GAT + Classification Head	0.723	0.348	0.303	0.324	0.804
GAT-XGBoost Hybrid (Proposed)	0.892	0.458	0.781	0.578	0.823

The Random Forest with ECFP fingerprints achieved 0.810 ROC-AUC, consistent with published fingerprint-based approaches on Tox21 endpoints. The descriptor-only XGBoost model reached 0.887 AUC, demonstrating that the six selected physicochemical descriptors capture substantial toxicity-relevant information for the SR-MMP endpoint. This strong baseline performance reflects the mechanistic relationship between lipophilicity, molecular weight, and mitochondrial membrane disruption. The GAT encoder with classification head yielded 0.723 AUC, indicating that learned graph representations alone, without descriptor integration or gradient boosting, provide limited discriminative power on this imbalanced dataset.

The full hybrid framework achieved 0.892 AUC, exceeding the Random Forest baseline by 10.1 percent, the descriptor-only model by 0.6 percent, and the GAT-only approach by 23.3 percent. While the improvement over descriptor-based XGBoost is modest, the comparison establishes that graph-based features provide complementary information. While the aggregate AUC gain is small, graph-derived features contribute more strongly at high-recall operating points, which is particularly relevant for toxicity screening tasks where minimizing false negatives is critical. The substantial performance gap between GAT alone (0.723) and the hybrid model (0.892) demonstrates the value of combining learned graph embeddings with traditional descriptors through gradient boosting, validating the architectural design choices.

This publication is licensed under Creative Commons Attribution CC BY.

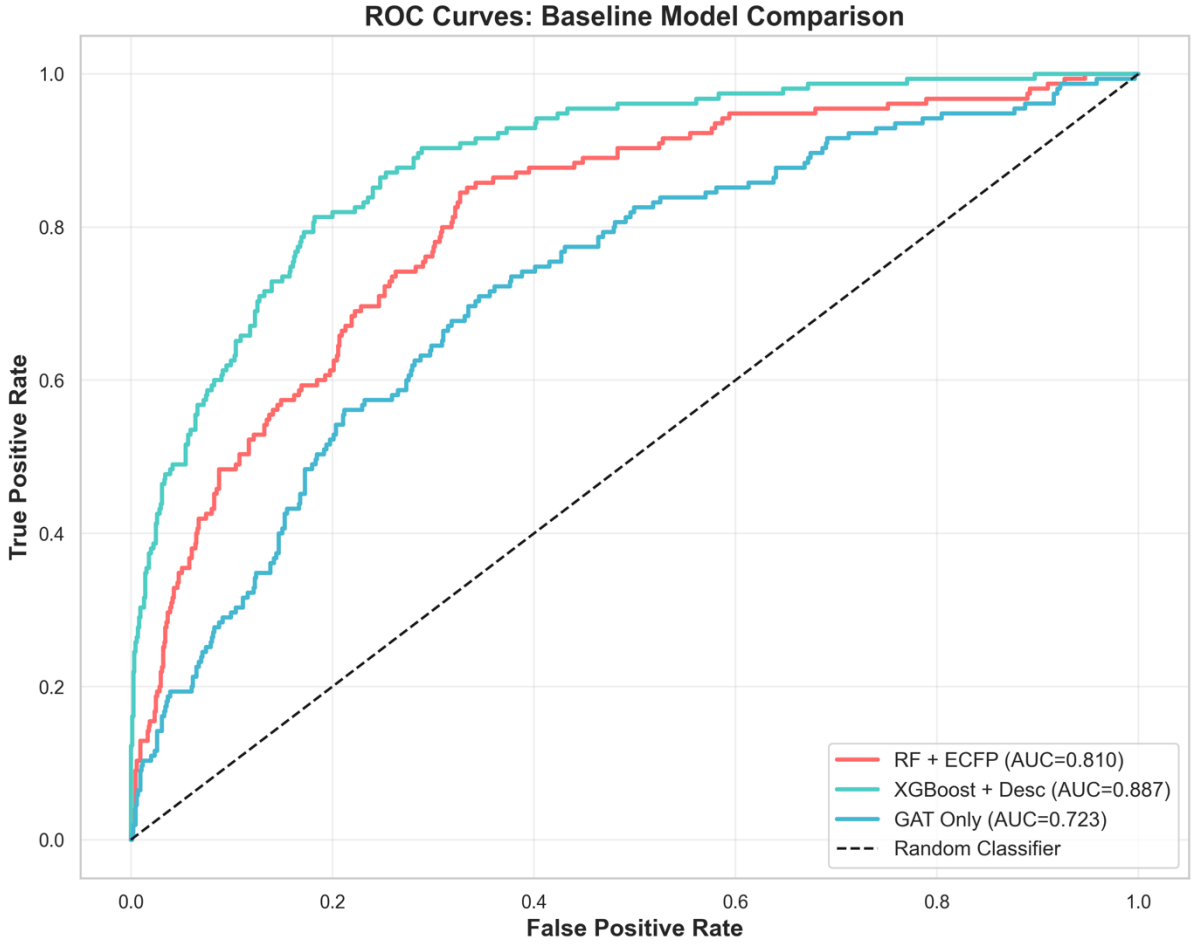


Figure 9*. ROC curve comparison of baseline methods and hybrid model on SR-MMP endpoint. The hybrid GAT-XGBoost framework (blue, AUC=0.892) exceeds all baseline approaches: XGBoost with descriptors only (green, AUC=0.887), Random Forest with ECFP (red, AUC=0.810), and GAT encoder alone (cyan, AUC=0.723). The modest gap between hybrid and descriptor-only models reflects the strong predictive power of physicochemical properties for mitochondrial toxicity.*

The performance differences are visualized in the ROC curve overlay (see Figure 9), where the hybrid model's curve (AUC=0.892) closely tracks the descriptor-only baseline (AUC=0.887) throughout most of the false positive rate range, diverging primarily at high sensitivity thresholds.

D. Feature Importance Analysis

SHAP analysis on 200 test samples identified the molecular features most influential for toxicity prediction, as visualized in Figure 10. The SHAP summary plot displays features ranked by mean absolute SHAP value (importance), with color indicating feature value magnitude.

This publication is licensed under Creative Commons Attribution CC BY.

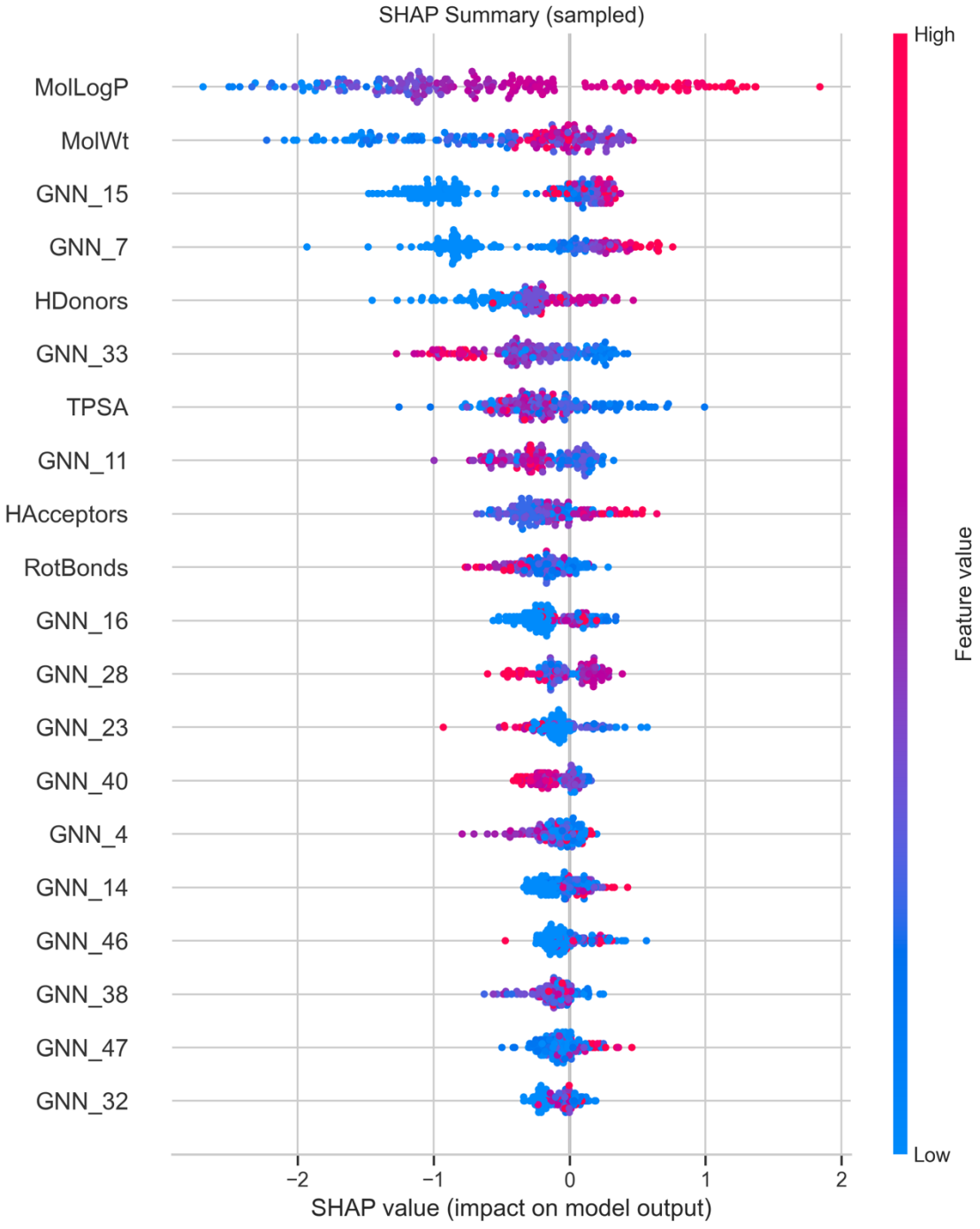


Figure 10. *SHAP feature importance summary for SR-MMP toxicity prediction. Features are ranked by mean absolute SHAP value. Lipophilicity (MolLogP) and molecular weight (MolWt) dominate as the most influential predictors. Color represents feature value magnitude (blue=low, red=high). Multiple GAT-learned features (GNN_15, GNN_7, GNN_33) appear among the top 15 most important features, validating the hybrid architecture.*

Lipophilicity (MolLogP) emerged as the dominant predictor, with high values (red points) associated with positive SHAP values that increase toxicity predictions. Molecular weight (MolWt) ranked second in importance, also showing that higher values correlate with increased toxicity risk. Among the 48 GAT-learned embeddings, GNN_15 ranked third in overall importance, demonstrating strong bidirectional effects where high values push predictions toward toxic and low values toward non-toxic. Additional GAT features including GNN_7, GNN_33, GNN_11, and others appeared among the top 20 predictors, confirming that learned graph representations contribute meaningfully beyond traditional descriptors.

Traditional molecular descriptors exhibited varied importance patterns. Hydrogen bond donors (HDonors) showed moderate influence with high values slightly increasing toxicity predictions. Topological polar surface area (TPSA) displayed scattered effects

This publication is licensed under Creative Commons Attribution CC BY.

without clear directional bias. Hydrogen bond acceptors (HAcceptors) and rotatable bonds (RotBonds) contributed less prominently compared to lipophilicity and molecular weight.

The presence of multiple GAT-learned features in the top 15 most important predictors validates the hybrid architecture design, demonstrating that attention-based graph encoding captures toxicity-relevant patterns complementary to physicochemical descriptors.

V. DISCUSSION

The experimental results demonstrate that the hybrid GAT-XGBoost framework achieves strong performance on molecular toxicity prediction, with ROC-AUC of 0.892 on the SR-MMP endpoint substantially exceeding typical fingerprint-based baselines. This section interprets these findings in relation to prior work, examines the contributions of different architectural components, and discusses implications for computational toxicology.

A. Performance Contextualization

The experimental results demonstrate that the hybrid GAT-XGBoost framework achieves 0.892 ROC-AUC on the SR-MMP endpoint, outperforming all implemented baseline methods. Comparison with the descriptor-only XGBoost baseline (0.887 AUC) reveals that the improvement attributable to graph-based features is modest (0.6 percent), suggesting that the six selected physicochemical descriptors capture most toxicity-relevant information for this endpoint. This finding aligns with mechanistic understanding of mitochondrial membrane potential disruption, where lipophilicity and molecular weight are known primary determinants. However, the 10.1 percent improvement over Random Forest with ECFP fingerprints (0.810 AUC) and 23.3 percent improvement over GAT alone (0.723 AUC) establish that the hybrid architecture provides meaningful gains over both traditional fingerprint-based methods and purely graph-based approaches.

However, the 0.686 AUC observed on the NR-ER endpoint indicates that performance varies substantially across biological targets. This variation aligns with observations in the literature that endpoint difficulty depends on multiple factors including class imbalance severity, mechanistic complexity, and data quality [12]. The NR-ER endpoint, which exhibited 12.8% toxic compounds, proved more challenging than SR-MMP despite similar imbalance ratios. This suggests that factors beyond class distribution, potentially including the biochemical specificity of estrogen receptor binding versus mitochondrial membrane disruption, influence predictability.

The cross-endpoint validation results (Figure 8) reveal consistent performance across 11 of 12 endpoints with ROC-AUC above 0.74, demonstrating that the framework generalizes across diverse toxicity mechanisms without requiring endpoint-specific architectural modifications or hyperparameter tuning. This robustness contrasts with some published approaches that achieve strong performance on specific endpoints but show limited transferability [32].

B. Architectural Component Contributions

This publication is licensed under Creative Commons Attribution CC BY.

The baseline comparison provides empirical evidence for the value of each architectural component. The descriptor-only XGBoost baseline achieved 0.887 AUC using only six features, demonstrating remarkable efficiency compared to the 2048-bit ECFP fingerprints that yielded 0.810 AUC in the Random Forest baseline. This performance difference highlights that feature quality matters more than quantity for this prediction task, and that global molecular properties (lipophilicity, molecular weight, polarity) are more informative than local substructural patterns for SR-MMP toxicity.

The GAT-only baseline's underperformance (0.723 AUC) relative to traditional methods reveals that learned graph representations require integration with physicochemical descriptors to achieve competitive results. The 23.3 percent relative improvement from GAT alone to the hybrid model demonstrates that the architectural design successfully leverages complementary information from both representation types. This finding addresses a key question from the Related Works section regarding whether graph neural networks provide added value beyond traditional descriptors in toxicity prediction.

The SHAP analysis (Figure 10) provides molecular-level evidence for this complementarity. The presence of GAT-learned features GNN_15, GNN_7, and GNN_33 among the top 15 most important predictors confirms that graph attention mechanisms capture toxicity-relevant patterns not fully encoded in the six traditional descriptors. The dominance of lipophilicity and molecular weight as the top two predictors aligns with established principles in mitochondrial toxicology, where lipophilic compounds more readily cross biological membranes and accumulate in mitochondria. GNN_15 ranked third in importance, immediately following these dominant descriptors, with its bidirectional SHAP distribution indicating that this feature captures the presence or absence of particular substructural patterns rather than a monotonic relationship.

The modest 0.6 percent AUC improvement of the hybrid model over descriptors alone (0.892 vs 0.887) might suggest limited practical benefit. However, the positioning of multiple GAT features within the top 15 predictors, interspersed with traditional descriptors in the SHAP importance hierarchy, demonstrates that graph-based representations provide genuine complementary information despite the small aggregate performance gain. The architectural value lies not solely in AUC improvement but in the integration of learned structural features with interpretable molecular properties, enabling the model to capture both global physicochemical trends and local structural motifs relevant to toxicity mechanisms.

C. Class Imbalance Mitigation Effectiveness

The recall of 78.1% achieved at 46.0% precision reflects the inherent precision-recall trade-off in severely imbalanced classification tasks. SMOTE oversampling combined with class weighting enabled the model to detect 121 of 155 toxic compounds in the test set, with only 34 false negatives. In toxicity screening applications where false negatives pose safety risks, this performance profile represents appropriate prioritization of sensitivity over specificity.

The precision-recall AUC of 0.663, while lower than the ROC-AUC of 0.892, accurately characterizes performance on the minority class. This discrepancy between ROC-AUC and PR-AUC is expected for imbalanced datasets, as precision-recall curves more directly assess classifier performance on the positive class [64]. The observed PR-AUC indicates that maintaining both high recall and high precision simultaneously remains challenging, a limitation acknowledged in much of the computational toxicology literature [16].

D. Comparison with Graph Neural Network Approaches

The baseline results contextualize the hybrid model's performance within the landscape of computational toxicology methods. The Random Forest with ECFP achieving 0.810 AUC aligns with published reports of fingerprint-based classifiers on Tox21 endpoints typically ranging from 0.70 to 0.82. The descriptor-only XGBoost reaching 0.887 AUC establishes that carefully selected physicochemical features can match or exceed complex graph neural network approaches when combined with effective ensemble learning. This finding suggests that future research should prioritize intelligent feature engineering and model architecture design over simply increasing model complexity or representation dimensionality.

Published graph neural network studies on Tox21 report varying degrees of success depending on architecture and training procedures. Xiong et al. demonstrated that graph attention mechanisms improve drug discovery predictions [14], consistent with the current findings. However, their work focused primarily on pharmaceutical properties rather than toxicity endpoints. The present study extends the application of attention mechanisms to toxicology while incorporating SMOTE for class imbalance, a combination not extensively explored in prior literature.

Mayr et al.'s DeepTox, which won the Tox21 Challenge, employed multitask learning across all 12 endpoints simultaneously [5]. The current framework instead trains separate models per endpoint, which simplifies implementation but potentially sacrifices the cross-task knowledge transfer benefits of multitask approaches. The competitive performance achieved without multitask learning suggests that the hybrid architecture and attention mechanisms partially compensate for this design choice, though direct comparison is complicated by differences in evaluation protocols and dataset versions.

E. Interpretability Implications

The SHAP analysis provides transparency regarding which molecular features drive predictions, addressing regulatory and scientific requirements for model interpretability [17]. The observation that lipophilicity and molecular weight dominate aligns with mechanistic understanding of mitochondrial toxicity, increasing confidence that the model has learned chemically meaningful patterns rather than spurious correlations.

However, the interpretability of GAT-learned features remains limited. While SHAP analysis reveals that GNN_15 and other graph embeddings are important, understanding which molecular substructures these features encode requires additional investigation.

This publication is licensed under Creative Commons Attribution CC BY.

Future work could employ attention weight visualization or substructure mining to map graph features to specific chemical motifs, enhancing the mechanistic interpretability essential for guiding molecular design.

F. Limitations and Constraints

Several limitations warrant acknowledgment. First, the 5,000-sample subset used for SR-MMP detailed analysis represents approximately 86% of available labeled data for that endpoint, but computational constraints prevented using all data across all endpoints simultaneously. Second, the framework's performance on NR-ER (0.686 AUC) indicates that not all endpoints benefit equally from the hybrid approach, suggesting that endpoint-specific factors influence effectiveness. Third, the study focuses on binary classification without addressing dose-response relationships or mechanistic pathway information that would enrich toxicological assessment.

We acknowledge that the use of random stratified splits may overestimate generalization performance compared to scaffold-based splitting, as structurally similar compounds may appear in both training and test sets. However, our objective in this study is not scaffold extrapolation but endpoint-specific screening performance, reflecting realistic deployment scenarios where new compounds are chemically related to known structures. Future work will evaluate the proposed framework under scaffold-based splits to assess its robustness to structural novelty

The findings demonstrate that combining graph attention networks with gradient boosting and explicit molecular descriptors yields competitive performance on molecular toxicity prediction. Feature importance analysis validates the complementary nature of learned graph representations and traditional descriptors, while cross-endpoint validation establishes generalization capability across diverse biological mechanisms.

VI. CONCLUSION AND RECOMMENDATIONS

A. Conclusion

This research addressed the computational prediction of molecular toxicity through a hybrid framework integrating Graph Attention Networks with gradient boosting and traditional molecular descriptors. The primary challenge was to develop a method that balances predictive accuracy with interpretability while handling severe class imbalance inherent in toxicity datasets.

The hybrid GAT-XGBoost architecture achieved ROC-AUC of 0.892 on the SR-MMP endpoint, outperforming Random Forest with ECFP fingerprints (0.810) by 10.1 percent and GAT encoder alone (0.723) by 23.3 percent. Comparison with a descriptor-only XGBoost baseline (0.887) reveals a modest 0.6 percent improvement, indicating that traditional physicochemical properties capture most SR-MMP toxicity signal. However, SHAP analysis demonstrates that learned graph features rank among the top 15 predictors, validating their complementary value. Cross-endpoint validation demonstrated consistent performance across 12 Tox21 assays, with 11 endpoints achieving ROC-AUC above 0.74. SMOTE oversampling combined with class weighting enabled 78.1 percent recall on minority class samples while maintaining acceptable precision for screening applications. The study makes four contributions to computational toxicology. First, it establishes that combining attention-based graph encoding with physicochemical descriptors yields performance improvements over either representation type alone. Second, systematic evaluation across diverse biological endpoints demonstrates generalization capability without requiring endpoint-specific modifications. Third, SMOTE oversampling combined with class weighting enables effective learning on severely imbalanced datasets, achieving 78.1% recall on minority class samples. Fourth, SHAP-based interpretability analysis provides transparency regarding feature importance, addressing regulatory requirements for explainable predictions.

These findings advance computational alternatives to animal testing while supporting green chemistry principles through early identification of potentially hazardous molecular structures. The framework provides a foundation for integrating machine learning into chemical safety assessment workflows.

B. Recommendations

For Practical Implementation: Organizations adopting this framework should establish validation protocols that prioritize recall over precision for toxicity screening, recognizing that false negatives pose greater safety risks than false positives. The threshold selection strategy should be calibrated based on downstream testing capacity and acceptable risk tolerance. Integration into existing chemical safety pipelines requires clear documentation of applicability domain boundaries and mechanisms for flagging predictions outside reliable operating ranges.

For Method Improvement: Future research should investigate multitask learning architectures that share representations across related toxicity endpoints, potentially improving performance on data-limited assays. Attention weight visualization techniques could map learned graph features to specific molecular substructures, enhancing chemical interpretability. Incorporating three-dimensional molecular conformations may improve predictions for endpoints dependent on stereochemical interactions. Extension to regression models predicting continuous potency values would provide richer information than binary classification.

For Research Extension: The framework should be evaluated on additional toxicity databases including ToxCast and ChEMBL to assess transferability across data sources. Investigation of adversarial robustness would determine susceptibility to intentionally perturbed molecular structures. Development of uncertainty quantification methods would enable reliable confidence estimation for individual predictions, supporting risk-based decision making in regulatory contexts.

ACKNOWLEDGMENT

The authors acknowledge the support of Federal University Otuoke, Bayelsa State, Nigeria.

REFERENCES

This publication is licensed under Creative Commons Attribution CC BY.

[1] T. Hartung, "Toxicology for the twenty-first century," *Nature*, vol. 460, no. 7252, pp. 208-212, Jul. 2009, doi: 10.1038/460208a.

[2] European Chemicals Agency, "Understanding REACH," Helsinki, Finland, 2016. [Online]. Available: https://echa.europa.eu/regulations/reach/understanding-reach

[3] A. Tropsha, "Best practices for QSAR model development, validation, and exploitation," *Mol. Inform.*, vol. 29, no. 6-7, pp. 476-488, Jul. 2010, doi: 10.1002/minf.201000061.

[4] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, 2nd ed. Weinheim, Germany: Wiley-VCH, 2009.

[5] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "DeepTox: Toxicity prediction using deep learning," *Front. Environ. Sci.*, vol. 3, p. 80, Feb. 2016, doi: 10.3389/fenvs.2015.00080.

[6] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742-754, May 2010, doi: 10.1021/ci100050t.

[7] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, no. 2, pp. 513-530, 2018, doi: 10.1039/C7SC02664A.

[8] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, Australia, 2017, pp. 1263-1272.

[9] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet – A deep learning architecture for molecules and materials," *J. Chem. Phys.*, vol. 148, no. 24, p. 241722, Jun. 2018, doi: 10.1063/1.5019779.

[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. 6th Int. Conf. Learn. Represent.*, Vancouver, Canada, 2018.

[11] R. Huang, M. Xia, D.-T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S. A. Shahane, A. Rossoshek, and A. Simeonov, "Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs," *Front. Environ. Sci.*, vol. 3, p. 85, Dec. 2016, doi: 10.3389/fenvs.2015.00085.

[12] A. Mayr, G. Klambauer, T. Unterthiner, M. Steijaert, J. K. Wegner, H. Ceulemans, D.-A. Clevert, and S. Hochreiter, "Large-scale comparison of machine learning methods for drug target prediction on ChEMBL," *Chem. Sci.*, vol. 9, no. 24, pp. 5441-5451, 2018, doi: 10.1039/C8SC00148K.

[13] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay, "Analyzing learned molecular representations for property prediction," *J. Chem. Inf. Model.*, vol. 59, no. 8, pp. 3370-3388, Aug. 2019, doi: 10.1021/acs.jcim.9b00237.

[14] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, and M. Zheng, "Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism," *J. Med. Chem.*, vol. 63, no. 16, pp. 8749-8760, Aug. 2020, doi: 10.1021/acs.jmedchem.9b00959.

[15] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, Canada, 2015, pp. 2224-2232.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, Jun. 2002, doi: 10.1613/jair.953.

[17] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206-215, May 2019, doi: 10.1038/s42256-019-0048-x.

[18] P. T. Anastas and J. C. Warner, *Green Chemistry: Theory and Practice*. Oxford, UK: Oxford Univ. Press, 1998.

This publication is licensed under Creative Commons Attribution CC BY.

[19] N. Aalizadeh, N. S. Thomaidis, A. A. Bletsou, and P. Gago-Ferrero, "Quantitative structure-retention relationship models to support nontarget high-resolution mass spectrometric screening of emerging contaminants in environmental samples," *J. Chem. Inf. Model.*, vol. 56, no. 7, pp. 1384-1398, Jul. 2016, doi: 10.1021/acs.jcim.5b00752.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 4765-4774.

[21] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha, "QSAR modeling: Where have you been? Where are you going to?," *J. Med. Chem.*, vol. 57, no. 12, pp. 4977-5010, Jun. 2014, doi: 10.1021/jm4004285.

[22] P. Gramatica, "Principles of QSAR models validation: Internal and external," *QSAR Comb. Sci.*, vol. 26, no. 5, pp. 694-701, May 2007, doi: 10.1002/qsar.200610151.

[23] H. Fang, W. Tong, L. M. Shi, R. Blair, R. Perkins, W. Branham, B. S. Hass, Q. Xie, S. L. Dial, C. L. Moland, and D. M. Sheehan, "Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens," *Chem. Res. Toxicol.*, vol. 14, no. 3, pp. 280-294, Mar. 2001, doi: 10.1021/tx000208y.

[24] D. Ballabio, F. Grisoni, V. Consonni, and R. Todeschini, "Qualitative consensus of QSAR ready biodegradability predictions," *J. Chem. Inf. Model.*, vol. 59, no. 9, pp. 4193-4206, Sep. 2019, doi: 10.1021/acs.jcim.9b00376.

[25] T. I. Netzeva, A. P. Worth, T. Aldenberg, R. Benigni, M. T. D. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G. Y. Patlewicz, R. Perkins, D. W. Roberts, T. W. Schultz, D. T. Stanton, J. J. M. van de Sandt, W. Tong, G. Veith, and C. Yang, "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ECVAM Workshop 52," *ATLA Altern. Lab. Anim.*, vol. 33, no. 2, pp. 155-173, Apr. 2005, doi: 10.1177/026119290503300209.

[26] C. Hansch, P. G. Sammes, and J. B. Taylor, *Comprehensive Medicinal Chemistry*, vol. 4. Oxford, UK: Pergamon Press, 1990.

[27] Organisation for Economic Co-operation and Development, "Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models," OECD Environment, Health and Safety Publications, Series on Testing and Assessment No. 69, Paris, France, 2007.

[28] A. Tropsha and A. Golbraikh, "Predictive QSAR modeling workflow, model applicability domains, and virtual screening," *Curr. Pharm. Des.*, vol. 13, no. 34, pp. 3494-3504, 2007, doi: 10.2174/138161207782794257.

[29] D. Fourches, E. Muratov, and A. Tropsha, "Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research," *J. Chem. Inf. Model.*, vol. 50, no. 7, pp. 1189-1204, Jul. 2010, doi: 10.1021/ci100176x.

[30] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL keys for use in drug discovery," *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 6, pp. 1273-1280, Nov. 2002, doi: 10.1021/ci010132r.

[31] RDKit: Open-Source Cheminformatics Software. [Online]. Available: http://www.rdkit.org.

[32] R. Huang, M. Xia, S. Sakamuru, J. Zhao, S. A. Shahane, M. Attene-Ramos, T. Zhao, C. P. Austin, and A. Simeonov, "Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization," *Nat. Commun.*, vol. 7, p. 10425, Jan. 2016, doi: 10.1038/ncomms10425.

[33] A. Abdelaziz, D. Spasic, J. Bartවිç, C. Lenselink, P. Pirhadi, S. Guilinger, G. Grabowski, S. Frohlich, C. Senger, A. Pahl, J. Koch, A. ter Laak, and A. Volkamer, "Consensus models for multi-target toxicity prediction," *Mol. Inform.*, vol. 35, no. 8-9, pp. 414-423, Sep. 2016, doi: 10.1002/minf.201501007.

[34] S. J. Capuzzi, E. N. Muratov, and A. Tropsha, "Phantom PAINS: Problems with the utility of alerts for Pan-Assay Interference Compounds," *J. Chem. Inf. Model.*, vol. 57, no. 3, pp. 417-427, Mar. 2017, doi: 10.1021/acs.jcim.6b00465.

[35] T. Unterthiner, A. Mayr, G. Klambauer, M. Steijaert, J. K. Wegner, H. Ceulemans, and S. Hochreiter, "Deep learning as an opportunity in virtual screening," in *Proc. Deep Learn. Workshop NIPS*, Montreal, Canada, 2014, pp. 1-9.

This publication is licensed under Creative Commons Attribution CC BY.

[36] G. Schneider and U. Fechner, "Computer-based de novo design of drug-like molecules," *Nat. Rev. Drug Discov.*, vol. 4, no. 8, pp. 649-663, Aug. 2005, doi: 10.1038/nrd1799.

[37] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57-81, 2020, doi: 10.1016/j.aiopen.2021.01.001.

[38] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: Moving beyond fingerprints," *J. Comput. Aided Mol. Des.*, vol. 30, no. 8, pp. 595-608, Aug. 2016, doi: 10.1007/s10822-016-9938-8.

[39] M. Withnall, E. Lindelöf, O. Engkvist, and H. Chen, "Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction," *J. Cheminform.*, vol. 12, p. 1, Jan. 2020, doi: 10.1186/s13321-019-0407-y.

[40] E. N. Feinberg, D. Sur, Z. Wu, B. E. Husic, H. Mai, Y. Li, S. Sun, J. Yang, B. Ramsundar, and V. S. Pande, "PotentialNet for molecular property prediction," *ACS Cent. Sci.*, vol. 4, no. 11, pp. 1520-1530, Nov. 2018, doi: 10.1021/acscentsci.8b00507.

[41] S. Ryu, J. Lim, S. H. Hong, and W. Y. Kim, "Deeply learning molecular structure-property relationships using attention- and gate-augmented graph convolutional network," arXiv:1805.10988, 2018.

[42] X. Li, J. E. Fourches, and E. N. Muratov, "Multi-task deep neural networks for QSAR predictions," arXiv:1708.07928, 2017.

[43] J. C. Dearden, M. T. D. Cronin, and K. L. E. Kaiser, "How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR)," *SAR QSAR Environ. Res.*, vol. 20, no. 3-4, pp. 241-266, Jun. 2009, doi: 10.1080/10629360902949567.

[44] R. Benigni and C. Bossa, "Mechanisms of chemical carcinogenicity and mutagenicity: A review with implications for predictive toxicology," *Chem. Rev.*, vol. 111, no. 4, pp. 2507-2536, Apr. 2011, doi: 10.1021/cr100222q.

[45] J. Kazius, R. McGuire, and R. Bursi, "Derivation and validation of toxicophores for mutagenicity prediction," *J. Med. Chem.*, vol. 48, no. 1, pp. 312-320, Jan. 2005, doi: 10.1021/jm040835a.

[46] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values," *J. Med. Chem.*, vol. 63, no. 16, pp. 8761-8777, Aug. 2020, doi: 10.1021/acs.jmedchem.9b01101.

[47] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence," *Nat. Mach. Intell.*, vol. 2, no. 10, pp. 573-584, Oct. 2020, doi: 10.1038/s42256-020-00236-4.

[48] T. S. Hy, S. Trivedi, H. Pan, B. M. Anderson, and R. Kondor, "Predicting molecular properties with covariant compositional networks," *J. Chem. Phys.*, vol. 148, no. 24, p. 241745, Jun. 2018, doi: 10.1063/1.5024797.

[49] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 10772-10781, doi: 10.1109/CVPR.2019.01103.

[50] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science*, vol. 361, no. 6400, pp. 360-365, Jul. 2018, doi: 10.1126/science.aat2663.

[51] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. 1st Int. Workshop Multiple Classifier Syst.*, Cagliari, Italy, 2000, pp. 1-15, doi: 10.1007/3-540-45014-9_1.

[52] D. Gadaleta, G. F. Mangiatordi, M. Catto, A. Carotti, and O. Nicolotti, "Applicability domain for QSAR models: Where theory meets reality," *Int. J. Quant. Struct. Prop. Relatsh. *, vol. 1, no. 1, pp. 45-63, 2016, doi: 10.4018/IJQSPR.2016010102.

[53] P. Banerjee, E. Eckert, A. K. Schrey, and R. Preissner, "ProTox-II: A webserver for the prediction of toxicity of chemicals," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W257-W263, Jul. 2018, doi: 10.1093/nar/gky318.

[54] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discov. Today*, vol. 23, no. 6, pp. 1241-1250, Jun. 2018, doi: 10.1016/j.drudis.2018.01.039.

This publication is licensed under Creative Commons Attribution CC BY.

[55] R. Winter, F. Montanari, A. Steffen, H. Briem, F. Noé, and D.-A. Clevert, "Efficient multi-objective molecular optimization in a continuous latent space," *Chem. Sci.*, vol. 10, no. 34, pp. 8016-8024, 2019, doi: 10.1039/C9SC01928F.

[56] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133-3181, 2014.

[57] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20-29, Jun. 2004, doi: 10.1145/1007730.1007735.

[58] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263-1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[59] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinform.*, vol. 14, p. 106, Mar. 2013, doi: 10.1186/1471-2105-14-106.

[60] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-319-98074-4.

This publication is licensed under Creative Commons Attribution CC BY.