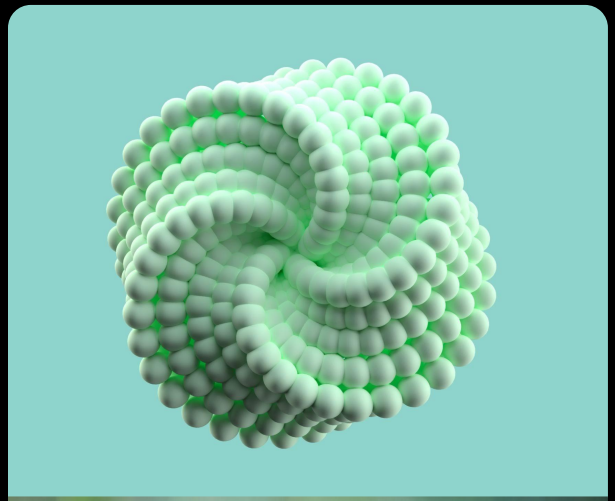
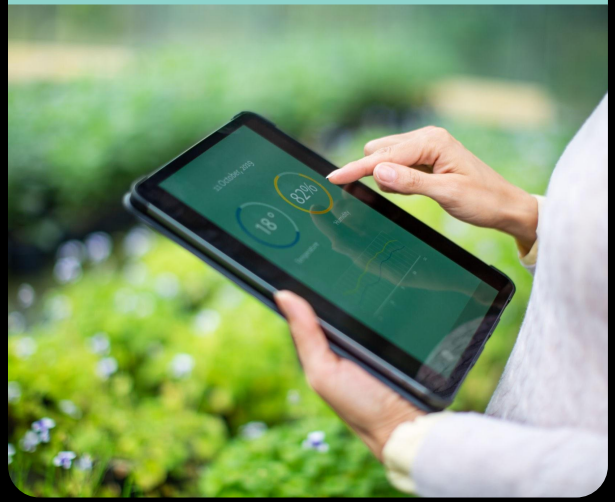




CADD PROJECT

TOXICITY PREDICTION

Contributors
NIKHIL KUMAR (2022322)
ADITYA KUMAR SINHA (2022034)
NUTAN KUMARI (2022341)
HARSH VISHWAKARMA (2022205)

OUTLINE

- Background & Motivation
- ADMET and Toxicity Prediction Models
- Chemical Representation Learning (Paper Summary)
- Methods: Dataset & Model Architecture
- Experimental Setup & Results
- Analysis of Model Performance (Figures)
- Proposed Modifications for Reliability
- Conclusions & Future Work
- References

BACKGROUND AND MOTIVATION

Toxicity is one of the leading causes of drug development failure, responsible for over 30% of clinical trial terminations. Early toxicity screening is therefore essential to avoid late-stage attrition and reduce development costs. ADMET properties—Absorption, Distribution, Metabolism, Excretion, and Toxicity—are fundamental to determining whether a drug candidate is viable. However, traditional in vitro and in vivo assays used for toxicity assessment are time-consuming, expensive, and often limited in scalability and predictive power.

Moreover, conventional approaches like avoiding known toxicophores have shown only limited success, especially in complex therapeutic areas such as oncology, where clinical success rates remain low. With rising complexity in biological systems and increasing pressure to bring safer drugs to market faster, the drug discovery process demands more robust and predictive tools.

Recent advances in artificial intelligence (AI) and machine learning (ML) offer a promising alternative. These technologies enable the development of scalable, data-driven models that can learn complex molecular patterns and predict toxicity outcomes more efficiently. This study focuses on leveraging AI-driven chemical language models to improve the reliability of ADMET property prediction, addressing key limitations of traditional methods and helping accelerate the path to safer drug development.

ADMET and Toxicity Prediction Models

ADMET

Absorption: How well a drug enters the bloodstream

Distribution: How it spreads through the body

Metabolism: How it's broken down

Excretion: How it exits the body

Toxicity: Harmful effects on biological systems

ADMET refers to the pharmacokinetic and toxicological properties of a drug that determine its behavior inside the human body.

- **Absorption**

Refers to how efficiently a drug is taken up into the bloodstream after administration (e.g., oral, intravenous).
Affects bioavailability, i.e., the proportion of the drug that reaches systemic circulation.

- **Distribution**

Describes how the drug spreads from the bloodstream to tissues and organs.
Influenced by plasma protein binding, lipophilicity, and blood–tissue barriers (e.g., blood–brain barrier).

- **Metabolism**

How the body chemically transforms the drug, primarily in the liver via enzyme systems like cytochrome P450.
Converts drugs into active or inactive metabolites.

- **Excretion**

The process of removing drug residues from the body, mainly via urine (renal) or feces (hepatic).
Impacts drug half-life and dosage requirements.

- **Toxicity**

Measures the drug's potential to cause adverse effects or damage to organs.
Can be organ-specific (e.g., hepatotoxicity) or systemic, and is often the primary reason for late-stage drug failure.

TOXICITY PREDICTION MODELS

PREDICTION MODELS:

- **QSAR models**: Use molecular descriptors/fingerprints (e.g., ECFP4)
- **Graph Neural Networks (GNNs)**: Model molecules as graphs of atoms and bonds
- **String-based Models**: Use SMILES/SELFIES encodings (e.g., CLM from paper)
- **Attention + Convolution**: Learn directly from chemical sequences with interpretability

KEY CHALLENGES:

- **Data scarcity**: Many datasets are small and incomplete
- **Class imbalance**: Toxic compounds are often underrepresented
- **Model interpretability**: Deep learning models are often black boxes
- **Generalization**: Models must work on unseen, real-world compounds

CHEMICAL REPRESENTATION LEARNING

Chemical representation learning refers to the process of **automatically learning molecular embeddings** directly from raw chemical structures (e.g., SMILES), rather than relying on handcrafted features or descriptors.

Paper Focus

The study emphasizes **end-to-end learning** of molecular features using neural networks, contrasting it with traditional approaches that use manually engineered descriptors like fingerprints or graph kernels.

Benefits

- **Captures meaningful substructure patterns** (e.g., toxicophores) without explicit rules
- **Improves generalization** across multiple toxicity prediction tasks
- **Highly transferable**—learned embeddings can be reused across datasets or related applications
- **Reduces feature engineering overhead**, enabling rapid model development

RESEARCH PAPER SUMMARY

Study Goals & Key Contributions

Toxicity is a major reason why many drugs fail late in development, increasing cost and risk.
This study introduces an **interpretable chemical language model (CLM)** using SMILES and attention mechanisms to better predict toxicity.

Main Contributions:

- The authors compared various molecular representations, including fingerprints, SMILES variants, and graph-based methods, and found that **SMILES with augmentation** performed best.
- They introduced a **simple yet powerful deep learning model** that outperformed more complex baselines across multiple toxicity prediction tasks.
- To enhance reliability, the model included **uncertainty estimation** using Monte-Carlo dropout and test-time augmentation.
- The model was also validated on a large proprietary dataset, where it maintained strong performance and effectively highlighted known **toxic substructures**.

RESEARCH PAPER SUMMARY

Model & Method Overview

- **Input:** The model uses randomized SMILES strings as input. These SMILES are generated through atom-order randomization to create multiple valid representations of the same molecule. This data augmentation improves training diversity and model generalization.
- **Architecture:**
    1. **Multiscale convolutions** to detect patterns at different levels (like looking at both atoms and groups of atoms).
    2. **Attention layers** that learn which parts of the molecule matter most—especially useful for identifying toxic substructures.
- **Training:** The model is trained using binary cross-entropy loss for toxicity classification tasks. Class balancing techniques are applied to handle label imbalance. SMILES augmentation is performed dynamically during training to increase variability and reducef overfitting.
- **Uncertainty handling:**
    1. **Monte-Carlo Dropout** to simulate an ensemble by introducing randomness at inference.
    2. **Test-Time Augmentation (TTA)** where multiple SMILES variations of the same molecule are averaged to make a more confident prediction.

RESEARCH PAPER SUMMARY

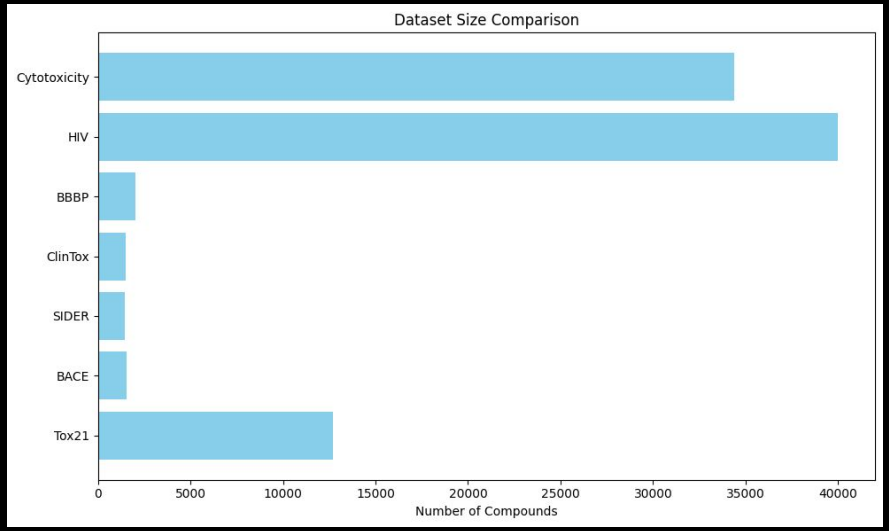
Results & Takeaways

- **Top performance**: SMILES + augmentation achieved the highest ROC-AUC scores.
- **Beats baselines** across various toxicity tasks (e.g., liver, mutagenicity).
- **Attention maps** highlight toxic regions—no extra supervision needed.
- **Uncertainty estimation** improves reliability, especially on unusual molecules.
- **Strong results** even on a 10K+ proprietary dataset—shows real-world potential.

Representation	ROC-AUC
Raw SMILES	0.832* ± 0.005
Canonical SMILES	0.830* ± 0.008
Kekulized SMILES	0.830* ± 0.006
Augmented SMILES	**0.853 ± 0.003**
SMILES without bond direction	0.834* ± 0.006
SMILES without chirality	0.834* ± 0.004
SMILES w/o bond direction & chirality	0.835* ± 0.006
Kekulized w/o bond direction & chirality	0.831* ± 0.004
SMILES with explicit bonds	0.834* ± 0.003
SMILES with explicit hydrogen	0.829* ± 0.007
SELFIES	0.827* ± 0.007
Augmented SELFIES	*0.852 ± 0.004*
Shuffled SMILES	0.830* ± 0.003
SMILES pair encoding	0.776* ± 0.01
Augmented SMILES pair encoding	0.825* ± 0.005

DATASET

DATASET USED: Seven benchmark datasets including Tox21, BACE, SIDER, ClinTox, BBBP, HIV, and Cytotoxicity, covering toxicity, permeability, side effects, and bioactivity with binary or multi-label tasks.



Dataset	Compounds	Tasks	Type	Key Features	Recommended Split
Tox21	12707	12	Toxic / Non-toxic	Hormone/stress pathways; p53 (drug resistance)	Fixed
BACE	1522	1	Binary (IC50 inhibition)	β-secretase 1 inhibitors	Scaffold
SIDER	1427	27	Adverse Drug Reactions (Multi-label)	Side effects grouped by organ systems	Scaffold
ClinTox	1491	2	FDA Approval / Toxicity	Approved vs failed drugs (toxicity)	Scaffold
BBBP	2039	1	Blood-Brain Barrier Penetration	Penetration across blood-brain barrier	Scaffold
HIV	40000	1	Active / Inactive for HIV	HIV replication inhibition	Scaffold
Cytotoxicity	34366	1	Cytotoxicity (HEK293 & HepG2)	Growth inhibition in kidney/liver cells	Stratified CV

PREPROCESSING STEP

PREPROCESSING STEP:

- SMILE Standardization: Cleanup of molecular strings and removal of invalid or incomplete entries.
- Train/Validation/Test Split:
  80% training
  10% validation
  10% testing
  Ensures balanced distribution across splits

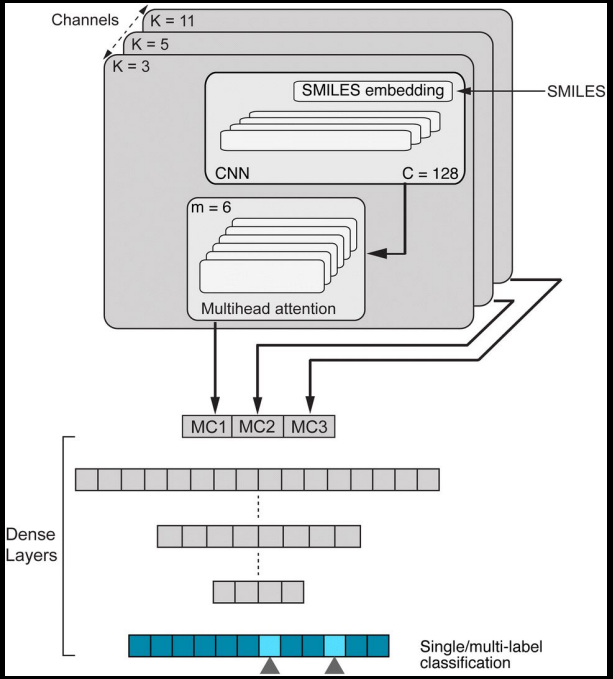
DATA AUGMENTATION:

1. Generates multiple valid SMILES per molecule by randomizing atom traversal.
2. Enhances model robustness.
3. Acts like data augmentation in image processing.

Model Architecture: MCA Model

The MCA model (Multiscale Convolutional Attention) is a novel architecture designed to predict molecular properties directly from SMILES strings. Below are its components:-

- Input
  SMILES strings are first converted into sequence embeddings (numerical vector representations).
- Multiscale Convolutions
  The model uses three parallel 1D convolutional layers with different kernel sizes (K = 3, 5, 11) to capture patterns at multiple scales. Each convolution uses 128 filters.
- Attention Mechanism
  Each convolutional output is passed through a multi-head self-attention block to learn which substructures are most important.
- Concatenation and Dense Layers
  Outputs from all convolution-attention blocks (MC1, MC2, MC3) are concatenated and passed through fully connected layers for feature fusion.
- Output
  The final dense layer gives one or multiple outputs, depending on whether the task is single- or multi-label classification.



Above Figure shows the sequential architecture of the MCA Model comprising of CNN layers then Attention Mechanism then finally dense and output layers.

Experiment Setup

To ensure fair benchmarking and reproducibility, a standardized training pipeline was employed for all models:

- Training Configuration
  Each model was trained for 200 epochs with a batch size of 64, balancing memory efficiency and learning stability.
  A constant learning rate of 1e-3 was used, ensuring gradual and stable model optimization throughout training.
- Hardware
  Experiments were performed on an NVIDIA GPU: A400, providing high parallel processing power.
  This significantly reduced training time and allowed deeper models to be trained effectively.
- Reproducibility
  To ensure consistent and repeatable results, a fixed random seed (12345) was set across all stages.
- Evaluation Metrics
  Performance was assessed using ROC-AUC (how well the model separates classes) and Accuracy (overall correct predictions).

Reproduced Results

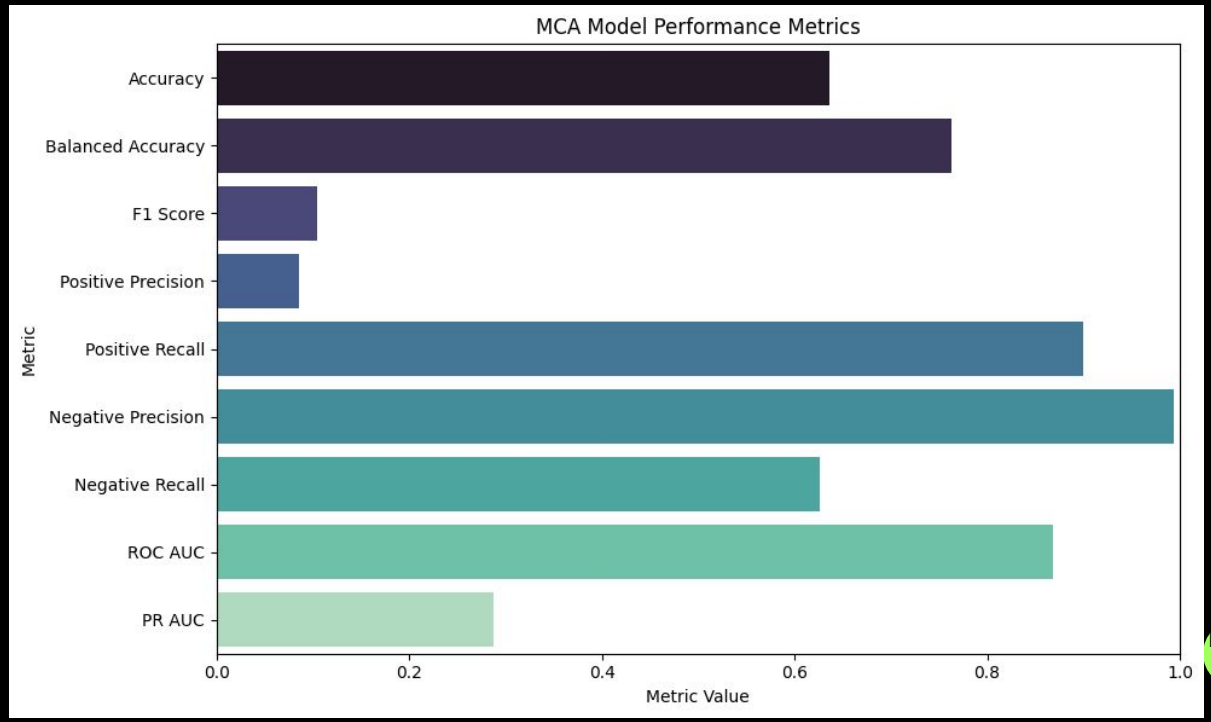
Comparison of the ROC-AUC scores for the three models (MCA, fine-tuned, and RNN) based on testing on the Tox21 dataset.

Type of Model	Original Paper Result	Our Results
MCA	0.853	0.8400
Fine Tuned Model	0.858	0.83976
RNN	0.781	0.8145

Performance Analysis

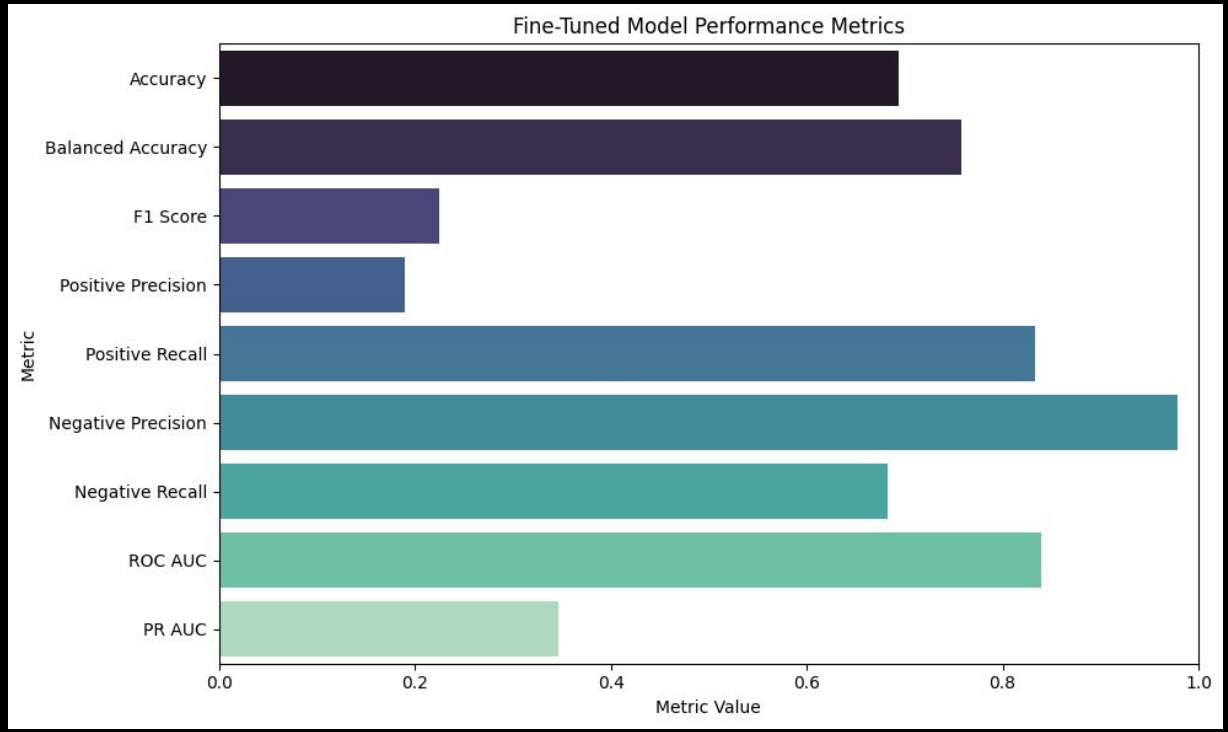
A comparative analysis of the performance metrics for the different models: MCA, fine-tuned, and RNN, based on the Tox21 dataset. The metrics, including ROC-AUC, accuracy, precision, recall, and F1-score, offer a detailed view of each model's capability in predicting toxicity-related endpoints.

MCA Model



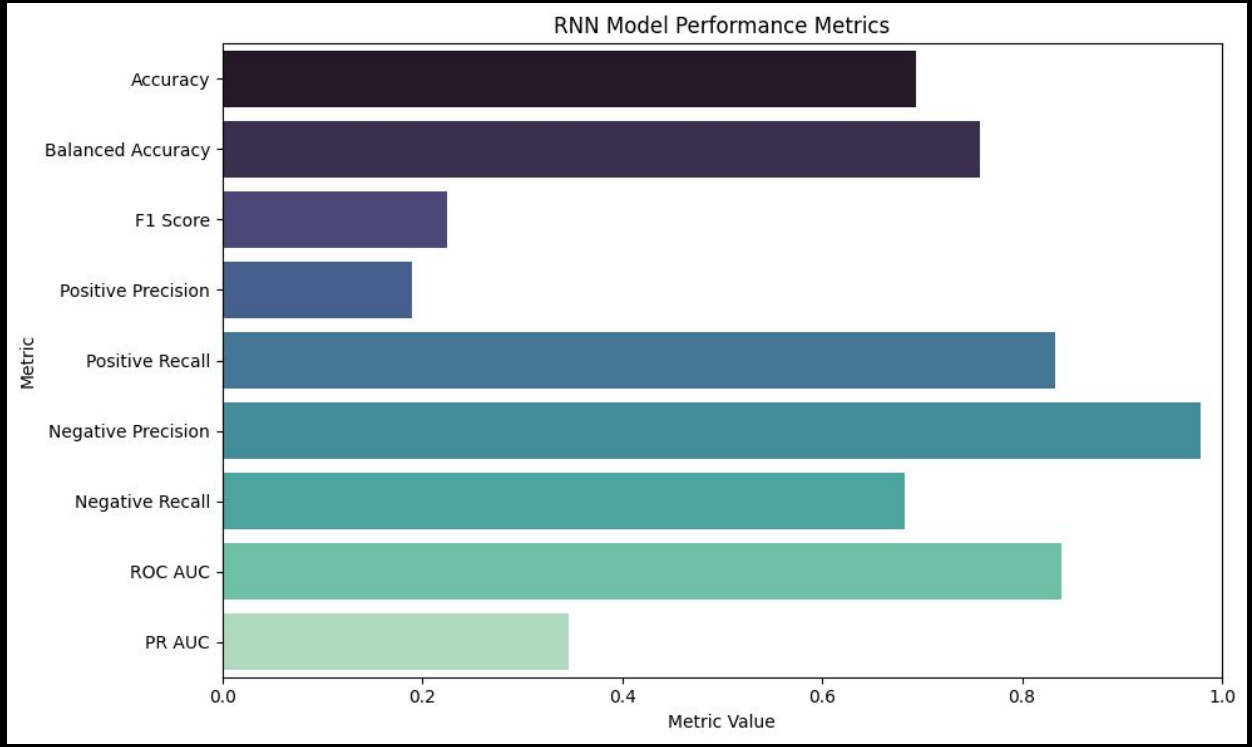
Performance Analysis

Fine-Tuned Model



Performance Analysis

RNN Model

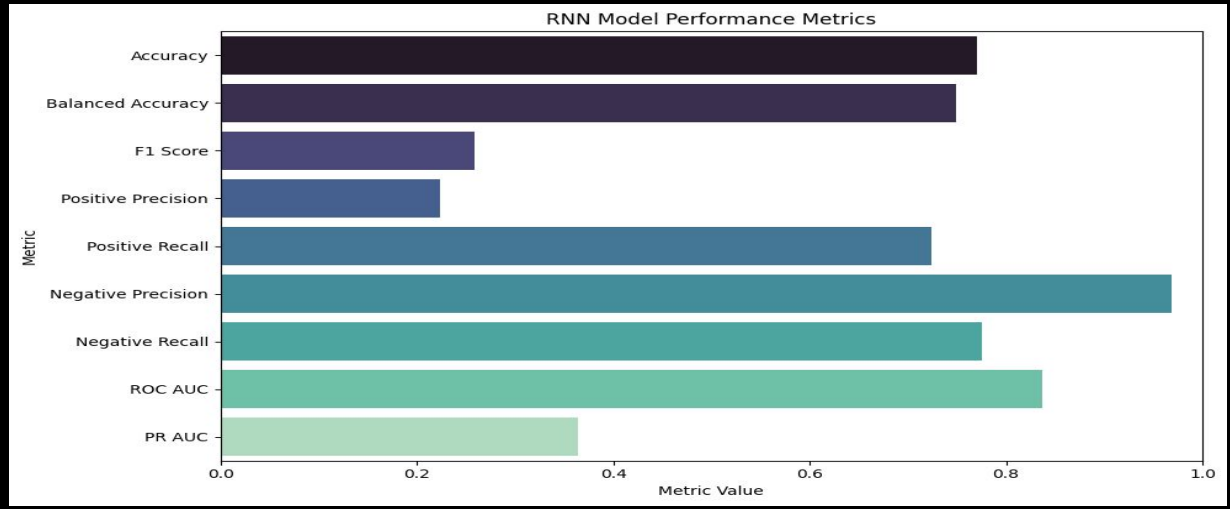


Proposed Modifications

To improve the performance of toxicity prediction on the Tox21 dataset, we fine-tuned the RNN model using several advanced techniques:

Roc_auc	F1	Positive Precision	Positive recall	Negative Precision	Negative Recall	Accuracy	Balanced Accuracy	Precision Recall-Score
0.86770	0.10465	0.08571	0.90000	0.99383	0.62646	0.63670	0.76323	0.28782

These improvements notably enhanced the ROC AUC and balanced accuracy, indicating better overall model performance in handling both toxic and non-toxic classes, despite class imbalance challenges.
This is the figure attached here.



Proposed Modifications

Below are the Parameters we used to tune the RNN based Model

Name of the Parameter	Value of the Parameter Used
Activation Function	Relu
Learning Rate	0.00001
Loss Function	binary_cross_entropy_ignore_nan_and_sum
Number of Layers Tuned	10
Dropout	0.5

Proposed Modifications

Limitations of Our Approach:-

- High Computational Requirements
  The fine-tuned RNN model demands significant computational power for training. In our setup, we relied on A400 GPUs due to the model's complexity and size. This limits scalability and may not be feasible for all research environments or deployment scenarios.
- Training Time and Resource Intensity
  Model training is time-consuming, especially when working with large molecular datasets and complex architectures. Hyperparameter tuning and multiple runs for robustness further add to the computational burden.
- Limited Dataset Generalization
  The model was evaluated primarily on the Tox21 dataset. While results show improvement, generalization to other datasets with different molecular properties remains unverified. Broader testing on additional benchmarks is essential to ensure the model’s reliability in real-world applications.
- Lack of Explainability
  The current model does not offer interpretability regarding which molecular features most influence toxicity predictions. This limits trust and usability in critical domains like drug discovery, where understanding feature impact is crucial for decision-making.

Conclusion & Future Work

The AI-driven toxicity prediction model shows promise in predicting ADMET properties, and with further enhancements, it can become a reliable tool for drug discovery.

- Summary
  The reproduced AI-driven toxicity prediction model aligns well with the original results, effectively using ADMET properties, with this SMILES-based representations and data augmentation were key to enhancing robustness.
- Hardware
  Experiments were performed on an NVIDIA GPU: A400s, providing high parallel processing power.
  This significantly reduced training time and allowed deeper models to be trained effectively.
- Limitations:
  The model's high computational cost and dependence on large unlabeled datasets may limit its accessibility. Additionally, the lack of interpretability can hinder its use in regulatory settings.
- Future Work
  Benchmarking with additional toxicity endpoints and integrating interpretability methods will expand the model's applicability. Model optimization and real-world application in drug discovery pipelines are the next steps for improving efficiency and impact.

REFERENCES

Zhang, et al

“Chemical representation learning for toxicity prediction," Chem. Data. Des., 2023

Wu, et al

“MoleculeNet: a benchmark for molecular machine learning," *Chem. Sci.*, 2018

Chen, et al

"Graph Contrastive Learning in Chemistry," *J. Chem. Inf. Model.*, 2022.

Chollet, F., et al

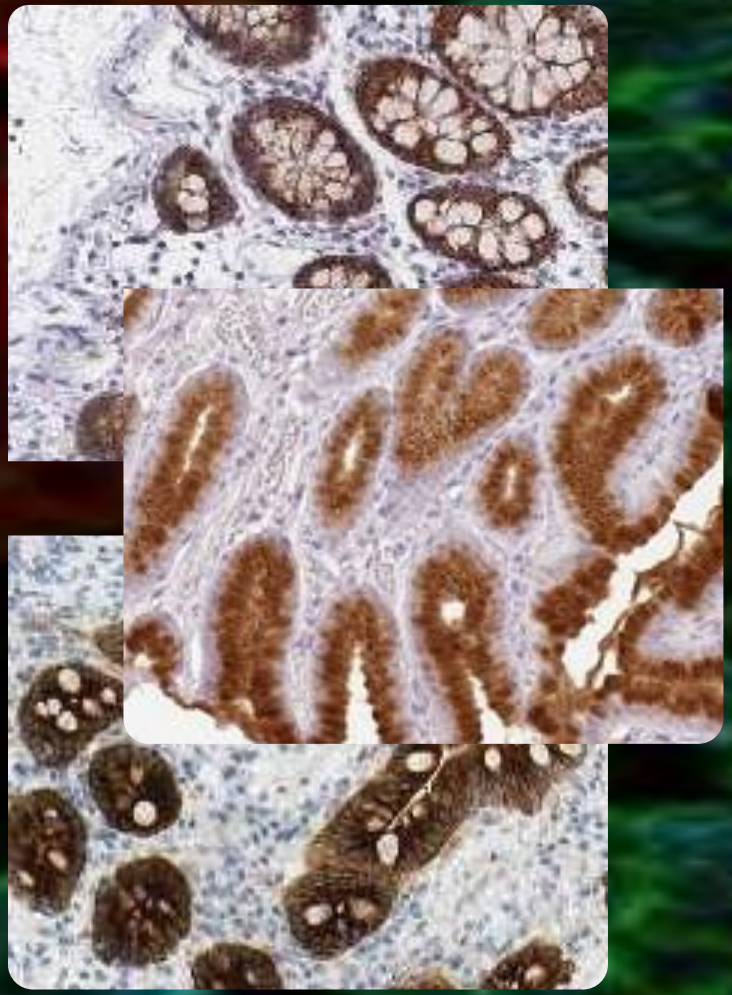
"Keras: The Python Deep Learning Library," *GitHub Repository*, 2015.

Williams, P., et al

"RDKit: Open-Source Cheminformatics," *Journal of Chemical Information and Modeling*, 2013.

Bland, S., et al

"DeepChem: A Library for Deep Learning in Chemistry," *Journal of Chemical Information and Modeling*, 2017.



THANK YOU