A Task-Hardness Driven Hybrid Modeling Framework for Molecular Toxicity Prediction

Sayyam Palrecha[1]

***Abstract*— Accurate molecular toxicity prediction is essential in early-stage drug discovery, yet existing models struggle with imbalanced datasets, noisy labels, and heterogeneous task difficulty. This work develops a unified benchmarking pipeline on the Tox21 dataset to evaluate baseline machine learning models, optimized ML classifiers, and graph-based deep learning models. Through the use of SMOTE balancing, Optuna hyperparameter tuning, and GraphConv/GAT architectures, the study observes complementary performance across different toxicity tasks. To characterize these variations, the analysis computes INT-CHEM (intrinsic task difficulty) and EXT-CHEM (cross-task similarity using Optimal Transport Dataset Distance), revealing two distinct groups of tasks favoring either single-task models or knowledge-sharing approaches. Guided by these findings, the work introduces a hybrid model that concatenates GNN-derived graph embeddings with ECFP fingerprints and trains a unified Random Forest classifier. This hybrid representation consistently improves PR-AUC across tasks and outperforms both standalone GNNs and classical models. Overall, the study combines task-hardness analysis with model benchmarking to arrive at a simple, effective, and interpretable hybrid architecture for toxicity prediction.**

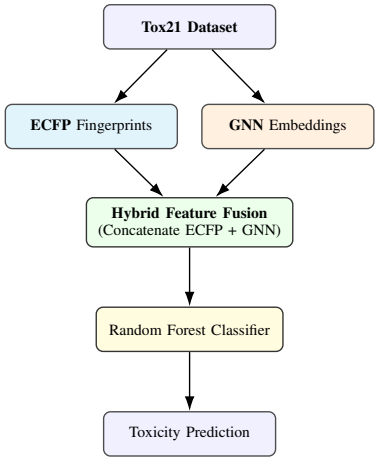


Fig. 1: Overview of the hybrid modeling pipeline combining GNN embeddings and ECFP fingerprints.

I. Introduction

Molecular toxicity prediction is a key component of modern drug discovery, enabling early identification of harmful compounds before expensive laboratory testing. Large screening campaigns such as Tox21 provide rich data across multiple toxicity pathways but exhibit challenges including severe class imbalance, heterogeneous assay noise, and variable task difficulty. These factors complicate the development of robust machine learning models [1]. Classical machine learning (ML) methods such as Random Forests, Support Vector Machines, and Gradient Boosting typically rely on engineered molecular fingerprints like ECFP. While effective in many settings, these models struggle when the underlying structure-toxicity relationships are highly nonlinear. Graph Neural Networks (GNNs), such as Graph Convolutional Networks (GraphConv) [2], [3] and more advanced architectures, offer an alternative by learning representations directly from molecular graphs, often providing more expressive structural insight [4]. However, neither ML models nor GNNs perform uniformly well across all Tox21 tasks. To understand these inconsistencies, the work quantifies task difficulty using two descriptors: INT-CHEM, reflecting intrinsic task hardness, and EXT-CHEM, measuring task relatedness via Optimal Transport Dataset Distance (OTDD) [5]. This analysis reveals two distinct groups of toxicity tasks: those favoring single-task ML models and those benefiting from knowledge sharing in GNNs. Motivated by these findings, the work develops a hybrid framework that joins GNN-derived embeddings with ECFP fingerprints and trains a Random Forest on the combined representation. This simple hybrid approach consistently outperforms standalone ML and GNN models, demonstrating that handcrafted and learned features capture complementary chemical information. The outcome is a unified, task-aware modeling strategy that leverages insights from task-hardness analysis to guide representation learning and predictive modeling.

[1] Undergraduate Student, Electrical Engineering, IIT Hyderabad ee22btech11047@iith.ac.in

II. Literature Survey

Molecular toxicity prediction has been widely explored using both traditional cheminformatics methods and modern deep learning architectures. Prior work broadly falls into three categories: classical machine learning, graph neural networks, and hybrid or multi-view approaches that integrate complementary molecular representations.

A. Classical ML for Toxicity Prediction

Early computational toxicology methods relied on engineered chemical descriptors such as Extended Connectivity Fingerprints (ECFP) and other fragment-based fingerprints[6]. Models like Random Forests (RF), Support Vector Machines (SVM), and Gradient Boosted Decision Trees (XGBoost) achieved strong performance in many QSAR tasks, including those in the Tox21 challenge. These models capture local substructures effectively, but their reliance on handcrafted representations limits their ability to

encode higher-order structural and electronic interactions. Moreover, classical models often struggle on datasets with severe class imbalance, assay noise, and nonlinear structure-toxicity relationships, all of which are characteristic of Tox21.

B. GNNs for Molecular Representation Learning

Graph Neural Networks (GNNs) represent a major shift toward learning molecular features directly from atom-bond graphs. The seminal MoleculeNet benchmark suite by Wu et al. [7] evaluated multiple graph architectures including Graph Convolutional Networks (GraphConv) and Message Passing Neural Networks (MPNNs) across molecular datasets such as Tox21, HIV, and QM9. The attention-based GNN architecture, Graph Attention Networks (GAT), introduced by Veličković et al. [8], further demonstrated the value of adaptive edge-weight learning for capturing chemically relevant interactions. The **MoleculeNet paper strongly influenced the first stage of my work**. After reviewing their benchmark design and model comparisons, a similar methodological philosophy was adopted:

- evaluate classical ML models,
- evaluate GNN models, and
- apply optimization strategies (hyperparameter tuning, data balancing, early stopping)

to construct a unified benchmark for the Tox21 dataset. Although GNNs often outperform classical ML models on complex toxicity mechanisms, their performance varies significantly across tasks. This motivates deeper investigation into task difficulty, task similarity, and the potential for **knowledge transfer across tasks**.

C. Hybrid and Multi-View Learning Approaches

Recent research increasingly focuses on hybrid, multi-representation models that integrate diverse molecular views including fingerprints, graph embeddings, SMILES/SELFIES embeddings, and transformer-based chemical language models. Such models leverage complementary strengths of different feature spaces to achieve stronger predictive performance. Parker et al. [9] demonstrated that combining graph-based molecular featurization with heterogeneous ensemble models substantially improves property prediction accuracy. Their findings underscore the potential of integrating learned graph representations with classical ML methods. Mixture-of-Experts (MoE) architectures and multimodal fusion networks have shown that combining handcrafted descriptors with learned embeddings can outperform single-view models, particularly for toxicity endpoints with heterogeneous mechanisms. Hybrid models are especially powerful when:

- GNNs capture global graph-level structure,
- Fingerprints capture local substructure motifs, and
- Chemical language models (e.g., ChemBERTa) capture sequence-based semantics.

These insights motivate the hybrid approach developed in this work, where GNN-derived embeddings are fused with ECFP fingerprints and used as input to a Random Forest classifier. This strategy aligns with prior findings that multi-view integration can resolve limitations of standalone ML models and GNNs especially for difficult toxicity tasks with complex decision boundaries.

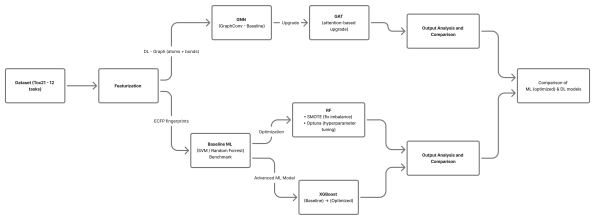


Fig. 2: Approach: Benchmarking workflow for ML and GNN models.

III. Materials and Methods

A. Data Description

The Tox21 dataset, part of the Tox21 Data Challenge and later incorporated into MoleculeNet[7], contains **12 binary toxicity prediction tasks** covering nuclear receptor (NR) signaling pathways and stress response (SR) pathways. The dataset includes approximately **8,000 compounds**, each annotated with SMILES strings and assay-level toxicity labels. As is typical in computational toxicology, the dataset exhibits:

- **severe class imbalance** (positive class often $< 10\%$),
- **missing labels** for certain assay-compound pairs,
- **heterogeneous difficulty** across toxicity mechanisms.

To ensure consistency across all experiments, DeepChem's standardized train/validation/test scaffold split was used.

B. Benchmarking Classical ML Models

Baseline performance using classical machine learning techniques on ECFP fingerprints was first established. ECFP4 fingerprints (radius 2, 1024 bits) were generated using DeepChem. Three ML models were benchmarked:

1) Support-Vector Machine (SVM),
2) Random Forest (RF),
3) XGBoost.

C. Model Optimization: SMOTE and Optuna

To strengthen the classical ML baselines, the following techniques were applied:

- **SMOTE oversampling** to balance the minority class,
- **Optuna hyperparameter tuning** to optimize RF and XGBoost.

The following search space was explored via Optuna:

- RF: (max_depth, n_estimators, min_samples_split, min_samples_leaf),
- XGBoost: (learning_rate, max_depth, subsample, colsample_bytree).

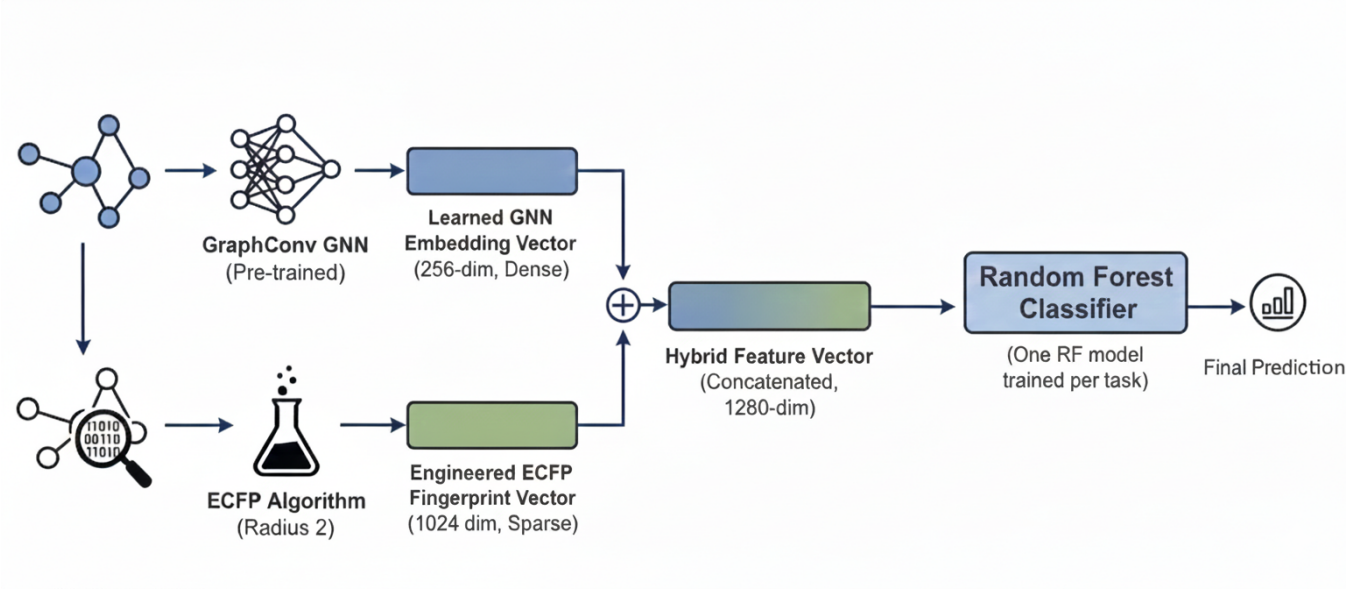


Fig. 3: Hybrid GNN-RF architecture used for final modeling.

D. Benchmarking GNN Models

To investigate deep-learned molecular representations, two GNN architectures were evaluated: GraphConv and GAT.

1) GraphConv Model: The Graph Convolutional Network was implemented using DeepChem's `GraphConvModel`. Molecules were represented as molecular graphs with atoms as nodes and bonds as edges. The model consisted of two graph convolutional layers with hidden dimensions of 64 and 128 units, followed by a dense layer of 128 units with ReLU activation. A global mean pooling layer aggregated node embeddings into a fixed-length molecular representation, which was then passed to a sigmoid output layer for multi-task classification over the 12 Tox21 targets. The model was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 64, dropout of 0.2, and for 30 epochs.

2) Graph Attention Network (GAT): The Graph Attention Network (GAT) was built using DeepChem's `GATModel`, employing two attention-based graph layers with 64 hidden units and four attention heads per layer to capture differential importance among atomic neighbors. The attention-weighted node embeddings were globally pooled to form a molecular-level representation, followed by a dense sigmoid output for multi-task classification across the 12 Tox21 assays. The model used the MolGraphConvFeaturizer with edge features to incorporate both atomic and bond-level information. Training was performed with the Adam optimizer using a learning rate of 0.001, batch size of 64, dropout of 0.2, and 50 epochs.

E. Task Hardness: INT-CHEM and EXT-CHEM

To understand the variability in performance across tasks, task difficulty was quantified following the framework proposed by Fooladi et al. [10].

1) INT-CHEM: Intrinsic Task Hardness: INT-CHEM measures the internal difficulty of a task and is defined as:

$$\text{INT-CHEM} = 1 - \text{PR-AUC}_{\text{few-shot}}$$

Few-shot single-task models (RF and kNN) were trained on small subsets (512, 1024, 2048 samples) and averaged performance. Higher INT-CHEM $\Rightarrow$ harder task.

2) EXT-CHEM: External Hardness (Task Relatedness): EXT-CHEM measures similarity to other tasks via the **Optimal Transport Dataset Distance (OTDD)[5]**. For each pair of tasks, OTDD was computed between:

- ChemBERTa molecular embeddings (feature space),
- Binary labels (label space).

F. Hybrid GNN-RF Model

To capitalize on the distinct strengths of learned graph representations and engineered chemical fingerprints, a two-stage hybrid model was constructed. The model's architecture is designed to first extract features from both paradigms and then use a robust classical model for prediction.

1) Stage 1: Hybrid Feature Generation: A single, high-dimensional feature vector was created for each molecule by concatenating two feature sets.

- **Learned Graph Embeddings:** A `GraphConvModel` from the `DeepChem` library was trained on the complete multi-task Tox21 training set. a validation callback was employed monitoring PR-AUC to save the model's best-performing parameters. This trained GNN was then used as a feature extractor, applying its `predict_embedding` method to generate a dense, N-dimensional graph embedding for each molecule.
- **Engineered Fingerprints:** A 1024-dimensional Extended-Connectivity Fingerprint (ECFP, radius 2) was computed for each molecule using the `CircularFingerprint` featurizer.

Molecules that failed featurization by either method were removed from all datasets to ensure perfect alignment. The resulting feature sets were horizontally concatenated, form-

ing a single hybrid feature vector of size $N(256) + 1024$ for each successfully featurized molecule.

2) Stage 2: Task-Specific Random Forest Classifiers: This hybrid feature matrix was used as the input for the final predictive model. Given the multi-label nature of the Tox21 dataset, a separate, task-specific Random Forest (RF) classifier was trained for each of the 12 assay endpoints. Each RF model was built. All subsequent performance evaluations were conducted using this ensemble of 12 independent hybrid RF models.

G. Model Evaluation and Interpretation

All models were evaluated on the **validation set** using:

- PR-AUC (primary metric),
- ROC-AUC,

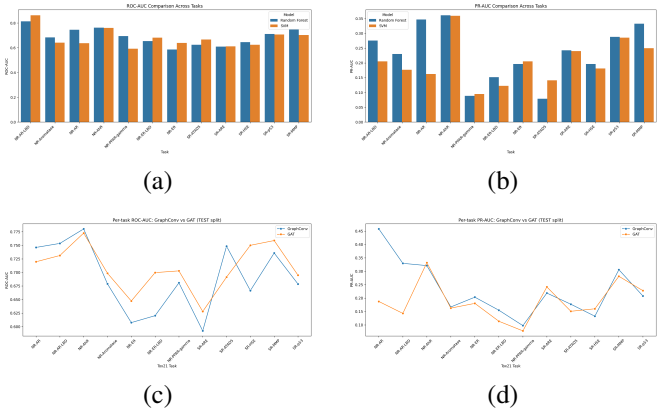
For evaluating model performance on the Tox21 dataset, primarily focus was on the PR-AUC. This choice is motivated by the inherent class imbalance prevalent across many Tox21 assays, where the positive class (toxic compounds) is significantly rarer than the negative class (non-toxic compounds). Unlike the ROC-AUC, which can present an overly optimistic view in the presence of severe imbalance, PR-AUC offers a more reliable assessment of a model's ability to correctly identify the minority class while minimizing false positives. Task-hardness descriptors were analyzed against model performance to interpret why certain tasks favored GNNs, ML models, or hybrid features. The hybrid model showed strong improvements particularly for **high INT-CHEM tasks**, validating the hypothesis that combining global graph features with local fingerprints resolves representational gaps present in GNN-only or fingerprint-only models.

IV. Results and Discussion

This section presents the benchmarking performance of classical machine learning models, graph neural networks, and the proposed hybrid RF-GNN model. Results are reported using ROC-AUC and PR-AUC across all 12 Tox21 tasks.

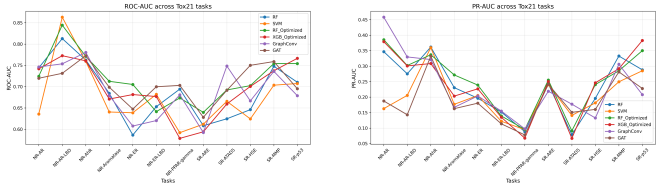
Classical ML models were benchmarked first using ECFP fingerprints. Fig. 4(a,b) compares Random Forest (RF) and Support Vector Machine (SVM) performance across all tasks. RF consistently outperformed SVM, achieving higher PR-AUC on 9 out of 12 tasks, confirming that tree-based methods are well-suited for high-dimensional sparse fingerprints. To establish a strong baseline prior to introducing graph-based and hybrid architectures, systematic optimization was performed of classical machine learning models, focusing on Random Forest (RF) and Extreme Gradient Boosting (XGBoost). The goal was to evaluate how much performance gain can be achieved purely through better configuration of these traditional methods compared to their baseline counterparts. optimization of RF and XGBoost was done using a two-stage procedure:

1) **Data-level balancing via SMOTE**: Synthetic oversampling was used to mitigate class imbalance in each task, improving the representation of minority classes and stabilizing model training.
2) **Hyperparameter tuning via Optuna**: A search space covering tree depth, number of estimators, learning rate (for XGBoost), and regularization terms was explored using a PR-AUC objective to reflect performance under class imbalance.



(a) (b) (c) (d)

Fig. 4: Model comparisons: (a) ROC-AUC and (b) PR-AUC between RF and SVM; (c) ROC-AUC and (d) PR-AUC between GraphConv and GAT across Tox21 tasks.



(a) (b)

Fig. 5: Optimized models comparison with baseline: (a) ROC-AUC and (b) PR-AUC performance across Tox21 tasks.

This approach allowed each optimized model to adapt more effectively to the characteristics of Tox21 tasks, with particular improvements observed for tasks having high INT-CHEM scores. Optimized RF and XGBoost models outperform their baseline versions across nearly all Tox21 tasks. However, it was observed that even well-tuned classical models struggled on tasks with complex structural dependencies, where graph-based or hybrid approaches perform markedly better. These findings align with our task hardness analysis:

- Tasks with **high INT-CHEM** often require more expressive molecular representations than fingerprints alone.
- Tasks with **low EXT-CHEM** benefit less from transfer-style inductive biases found in GNNs.

Next, the GraphConv neural network trained using DeepChem was evaluated. Per-task ROC-AUC and PR-AUC are shown in Fig. 4(c,d). Although GNNs performed well on certain structure-driven tasks such as NR-AR, NR-AhR, and SR-ARE, they struggled on tasks with high intrinsic hardness (INT-CHEM), consistent with previous findings in molecular property prediction. To compare all models jointly, results were aggregated for baseline RF, optimized

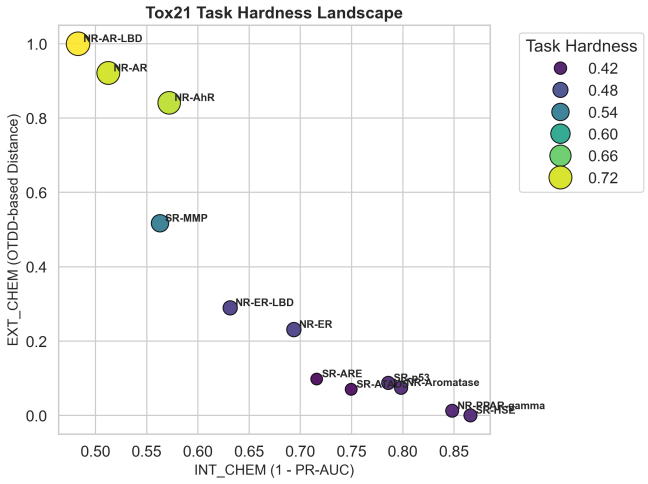


Fig. 6: Task hardness landscape using INT-CHEM and EXT-CHEM descriptors, showing two distinct groups of tasks.

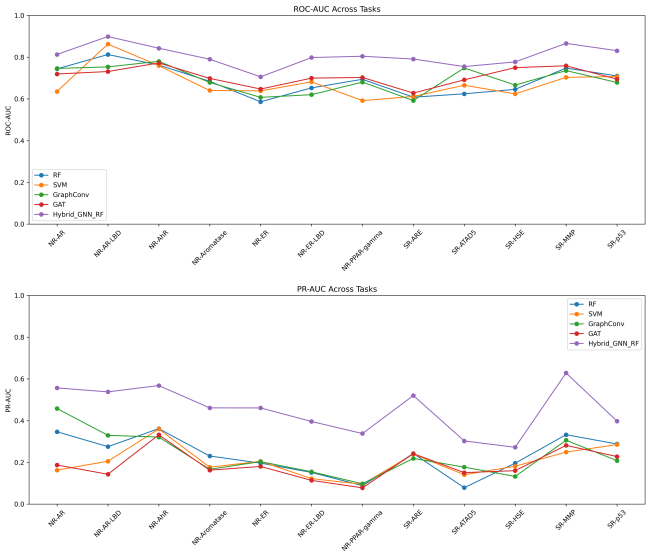


Fig. 7: Hybrid model comparison across all Tox21 tasks using (a) ROC-AUC and (b) PR-AUC.

RF, optimized XGBoost, GraphConv, and GAT (Fig. 5). GNNs performed competitively with optimized tree-based models but did not consistently surpass them. This reinforced the hypothesis that different models capture complementary chemical information. **To delve deeper into the nature of these observed performance differences, the analysis leverages established task-hardness descriptors, specifically INT-CHEM and EXT-CHEM.** These metrics provide a crucial lens through which to interpret model behavior, revealing underlying characteristics of the toxicity tasks themselves [10]. Two clear groups of toxicity tasks are observed(also shown in Fig. 6):

- **High INT-CHEM, Low EXT-CHEM:** intrinsically difficult tasks but does have similar tasks nearby in chemical space (externally easy) where multi-task GNNs should theoretically excel.
- **Low INT-CHEM, High EXT-CHEM:** intrinsically easier and externally difficult tasks well-handled by single-task RF models.

Surprisingly, GraphConv did not outperform RF on the high INT-CHEM tasks, suggesting that the GNN learned incomplete structural representations. This insight motivated the hybrid model. The hybrid model concatenates GNN embeddings with ECFP fingerprints and trains a single Random Forest on the joint representation. Figures 7(a,b) show that the hybrid model significantly improves PR-AUC across nearly all tasks, particularly those previously identified as high-hardness. Tables I and II together highlight the overall performance gains, showing that while optimized classical models provide modest improvements, the proposed Hybrid_GNN_RF model consistently delivers the largest increase in predictive accuracy across Tox21 tasks.

TABLE I: Average ROC-AUC and PR-AUC scores across Tox21 tasks scores for all models.

Model	PR_AUC	ROC_AUC
RF	0.232358	0.689495
SVM	0.201932	0.676948
RF(Optimized)	0.249211	0.717878
XGB(Optimized)	0.238184	0.694999
GraphConv	0.231218	0.690771
GAT	0.188265	0.707952
Hybrid_GNN_RF	**0.453431**	**0.806352**

TABLE II: Percentage PR-AUC improvement over RF baseline for all models across Tox21 tasks.

Task	RF(Opt.)	XGB(Opt.)	Hybrid_GNN_RF
NR-AR	11.13	9.29	**60.64**
NR-AR-LBD	9.70	9.26	**95.42**
NR-AhR	-6.82	-14.71	**57.14**
NR-Aromatase	17.84	-11.78	**100.17**
NR-ER	21.60	15.40	**134.70**
NR-ER-LBD	-8.29	-10.99	**161.05**
NR-PPAR-gamma	8.08	-23.66	**282.91**
SR-ARE	5.13	3.19	**114.60**
SR-ATAD5	15.77	-15.10	**285.09**
SR-HSE	22.63	25.59	**38.85**
SR-MMP	-14.36	-12.32	**88.97**
SR-p53	21.55	32.91	**38.08**

V. Conclusion

In this study, a comprehensive study of molecular toxicity prediction was conducted using the Tox21 dataset, benchmarking a wide spectrum of classical machine learning models, graph neural networks, and hybrid architectures. The first objective was to establish a rigorous baseline. Classical models such as Random Forests, SVMs, and XGBoost, were evaluated and the results showed that careful optimization through SMOTE balancing and Optuna hyperparameter tuning significantly improves their performance. However, these models remained limited by their reliance on fixed molecular fingerprints. Graph-based deep learning models, specifically GraphConv and GAT, offered more expressive structural modeling but still exhibited inconsis-

tent task-wise performance. To understand the source of this variability, the analysis incorporated INT-CHEM and EXT-CHEM descriptors to quantify task hardness and inter-task relatedness. This characterization revealed two distinct groups of Tox21 tasks: those that benefit from single-task learning and those that benefit from knowledge sharing. These findings provided insight into why neither classical ML nor GNNs consistently achieved superior performance across all endpoints. Motivated by these insights, the study introduced a hybrid modeling strategy that combines learned GNN embeddings with handcrafted ECFP fingerprints, using a Random Forest classifier as the final predictor. This hybrid approach successfully integrates complementary molecular views: GNN embeddings capture structural and relational information, while ECFP fingerprints provide precise substructural cues. Empirically, the hybrid model outperformed both optimized classical models and standalone GNNs across most toxicity tasks, especially on those with high intrinsic hardness. Overall, our study demonstrates that no single representation or modeling paradigm is universally optimal for Tox21. Instead, the most effective strategy is one that integrates multiple molecular views informed by task hardness analysis. The work highlights the value of task-aware model design and provides a generalizable framework for combining classical cheminformatics features with deep learned representations. Future work may include quantifying feature attribution (SHAP, gradients) to disentangle ECFP vs. GNN contributions and task-wise gains (with INT/EXT-CHEM comparisons), extend models via mixture-of-experts routing informed by hardness descriptors and transfer/meta-learning for few-shot adaptation, and integrate domain constraints by adding toxicophore alerts and PAINS filters as features/penalties, reporting rule-model concordance to ensure mechanistic plausibility and reduce spurious correlations.

REFERENCES

[1] R. Huang, M. Xia, D.-T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S. 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," *Frontiers in Environmental Science*, vol. 3, 01 2016.

[2] T. Siameh, "Semi-supervised classification with graph convolutional networks," 12 2023.

[3] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. Adams, "Convolutional networks on graphs for learning molecular fingerprints," *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, 09 2015.

[4] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1263–1272, JMLR.org, 2017.

[5] K. Nguyen, H. Nguyen, T. Pham, and N. Ho, "Lightspeed geometric dataset distance via sliced optimal transport," 05 2025.

[6] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[7] Z. Wu, B. Ramsundar, E. Feinberg, J. Gomes, *et al.*, "Moleculenet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018.

[8] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[9] M. L. Parker, S. Mahmoud, B. Montefiore, M. Ören, H. Tandon, C. Wharrick, and M. D. Segall, "Improving predictions of molecular properties with graph featurization and heterogeneous ensemble models," *Journal of Chemical Information and Modeling*, vol. 65, no. 21, pp. 11644–11655, 2025.

[10] H. Fooladi, S. Hirte, and J. Kirchmair, "Quantifying the hardness of bioactivity prediction tasks for transfer learning," *Journal of Chemical Information and Modeling*, vol. 64, no. 10, pp. 4031–4046, 2024.