Akshay Balaji, 2025, 13:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

MolGraphormer: An Interpretable GNN-Transformer for Uncertainty-Aware Molecular Toxicity Prediction

Akshay Balaji
Velammal Vidyalaya Annexure Chennai, India

Abstract- Accurate and Interpretable toxicity prediction re- mains fundamental in computational chemistry and drug discov- ery. We propose MolGraphormer, a Transformer-GNN hybrid architecture integrating Graph Neural Network message passing with self-attention mechanisms for molecular property prediction. Our model incorporates substructure-aware embeddings via multi-head attention, edge-conditioned message passing, and hierarchical graph aggregation, enabling both local and global molecular reasoning. Evaluated on the Tox21 benchmark dataset, MolGraphormer achieves competitive performance with F1-Score of 0.6697 and AUC-ROC of 0.7806, while maintaining strong recall (0.7787) for identifying toxic compounds. We employ Monte Carlo Dropout and Temperature Scaling for uncertainty quantification, Combined with uncertainty quantification and attention-based interpretability, MolGraphormer offers a practical framework for drug safety assessment and regulatory toxicology.

Keywords: Graph Neural Networks, Molecular Toxicity, Transformer attention, uncertainty quantification, drug discovery, interpretable AI.

I. INTRODUCTION

Molecular toxicity prediction is critical in drug discovery, regulatory science, and environmental risk assessment. Tra- ditional QSAR models rely on handcrafted descriptors and linear methods that fail to capture complex structural patterns. GNNs have emerged as powerful alternatives by modeling molecules as Non-Euclidean graph structures [1]. However, standard message-passing GNNs exhibit limitations including restricted receptive fields, limited interpretability, and inefficient modeling of long-range dependencies.

We introduce MolGraphormer, a Transformer-enhanced GNN framework integrating self-attention with graph message passing. Unlike conventional GNNs aggregating only local neighborhoods, MolGraphormer employs hierarchical molecu- lar reasoning to capture both long-range chemical interactions and long-range structural dependencies.

A. Contributions

Our main contributions include:

- A hybrid Transformer-GNN architecture achieving F1- score of 0.6697 and AUC-ROC of 0.7806 on Tox21 benchmark
- Comprehensive uncertainty quantification with MC Dropout(ECE = 0.0851) and temperature scaling
- Superior recall(0.7787) critical for safety-critical toxicity screening.
- Analysis of attention patterns correlating with known toxicophores
- Open implementation enabling reproducibility

II. RELATED WORK

GNNs have proven effective for molecular modeling [1]. Kipf and Welling [2] introduced spectral graph convolutions, while Gilmer et al. [3] established the message-passing neural networks framework. Recent work combines GNNs with Transformers for molecular property prediction [4], [5].

For uncertainty quantification, Gal and Ghahramani [6] proposed MC Dropout as Bayesian approximation, while Guo et al. [7] introduced temperature scaling for calibration. Ying et al. [8] developed GNNExplainer for interpretability through subgraph explanations.

III. METHODOLOGY

A. Dataset and Preprocessing

We evaluate Tox21, containing 7,831 compounds tested across 12 toxicity assays. We formulate binary classification where compounds testing positive on an assay are labeled toxic. The dataset contains 2,872 toxic (36.7%) and 4,959 non- toxic (63.3%) compounds. We use 80% (6,264 compounds) for training and 20% (1,567 compounds) for testing, with 15% of training data (936 molecules) reserved for validation.

B. Molecule Graph Construction

Molecules are represented as attributed graphs where atoms are nodes and bonds are edges.

Node features (8 dimensions): Atomic number, atom degree, formal charge, hybridization state, aromaticity, hydrogen count, ring membership, atomic mass.

Edge features (3 dimensions): Bond type (single, double, triple, aromatic), conjugation, ring membership.

Global features (8 dimensions): Molecular weight, LogP, H-Bond donors/acceptors, TPSA, rotatable bonds, aro- matic/aliphatic ring counts.

C. MolGraphormer Architecture

Our architecture consists of four components:

1. **Embedding Layer:** Node features are projected from 8 to 128 dimensions with batch normalization. Global features are projected to 64 dimensions. ReLU activation provides non-linearity.
2. **Graph Attention Layers:** We employ 4 layers of multi- head Graph Attention (4 heads) with edge-conditioning. Each layer includes residual connections:

$$h_v^{(t+1)} = \text{LayerNorm}(h_v^{(t)} + \text{GAT}(h_v^{(t)}, \{h_u^{(t)}\}_{u \in N(v)}, e_{vu})) \quad (1)$$

where e_{vu} represents edge features. Attention coefficients are:

$$\alpha_{vu} = \frac{\exp(\text{ReLU}(a^T [Wh_v || Wh_u || W_e e_{vu}]))}{\sum_{w \in N(v)} \exp(\text{ReLU}(a^T [Wh_v || Wh_w || W_e e_{vw}]))} \quad (2)$$

Batch normalization and dropout (0.2) provide regularization.

3. **Graph Pooling:** We combine mean and max pooling:

hgraph = [hmean||hmax||hglobal] (3)

4. **Classification Head:** Three-layer MLP: 320 → 128 → 64 → 2 with ReLU and progressive dropout (0.5, 0.4). Softmax output for binary classification.

D. Training Configuration

Optimization: Adam optimizer with learning rate 0.001, weight decay 5e-4, batch size 64, cross-entropy loss. ReduceL- ROnPlateau scheduling (factor = 0.5, patience = 10) with gradient clipping (max norm 1.0).

Training Protocol: Maximum 100 epochs with early stopping (Patience = 20) based on validation F1-score. Mol- Graphormer converged at epoch 84.

E. Baseline Models

We compare against four GNN architectures: GCN [2], GAT [9], GraphSAGE [10], and GIN [11]. All use equivalent depth (4 layers), hidden dimensions (128), and training protocols. GAT employs ELU activation, while others use ReLU.

F. Uncertainty Quantification

MC Dropout: We perform 30 stochastic forward passes with active dropout(p=0.2) to obtain predictive uncertainty:

$$\sigma_{MC}(x) = \sqrt{\frac{1}{30} \sum_{t=1}^{30} (f_t(x) - \bar{f}(x))^2} \quad (4)$$

Temperature Scaling: We learn optimal temperature $T^* = 0.9771$ on validation set:

$$p_i^{calibrated} = \frac{\exp(z_i / T^*)}{\sum_j \exp(z_j / T^*)} \quad (5)$$

IV. RESULTS AND DISCUSSION

A. Overall Performance

Table I and Figure 1 present test set performance compared to baseline architectures.

TABLE I: Performance on Tox21 Binary Classification

Model	F1	AUC	Acc	Prec	Rec
GCN	0.597	0.741	0.668	0.538	0.671
GAT	0.627	0.754	0.682	0.551	0.728
GraphSAGE	0.644	0.771	0.724	0.611	0.681
GIN	0.640	0.792	0.749	0.674	0.610
MolG.	0.670	0.781	0.718	0.587	0.779

Key Observations:

- MolGraphormer achieves best F1-score (0.6697), demon- strating 3.9% improvement over GraphSAGE
- Highest recall (0.7787) identifies 77.9% of toxic com- pounds, critical for safety screening.
- Competitive AUC-ROC (0.7806), second to GIN (0.7922)
- Moderate precision (0.5874) reflects conservative strategy appropriate for safety applications.

B. Uncertainty Quantification

Table II, Figure 2 and Figure 3 present calibration metrics.

TABLE II: Uncertainty Quantification Metrics

Method	ECE ↓	Brier ↓
Baseline	0.132	0.208
Temp. Scaled (T=0.977)	0.132	0.208
MC Dropout (30 samples)	**0.085**	**0.193**

MC Dropout achieves best calibration (ECE = 0.0851), rep- resenting 35.5% improvement, over baseline. Mean predictive uncertainty is 0.0583 ± 0.0167. High-uncertainty predictions suggests that the model appropriately identifies on challenging cases. Temperature scaling marginally improves Brier score from 0.2081 to 0.2075, confirming reasonable initial calibration.

C. Baseline Comparison

GCN (F1=0.597): Standard spectral convolutions struggle with chemical heterogeneity, achieving lowest performance without attention mechanism.

GAT (F1=0.627): Multi-head attention improves upon GCN by 5.1%, demonstrating attention value, but lacks residual connections and global features.

GIN (F1=0.640): Isomorphism-preserving aggregation achieves highest AUC-ROC (0.792) but lower recall (0.610) limits safety screening utility.

MolGraphormer (F1=0.670): Hybrid architecture suc- calibrated cessfully combines attention strengths with global feature integration and residual connections.

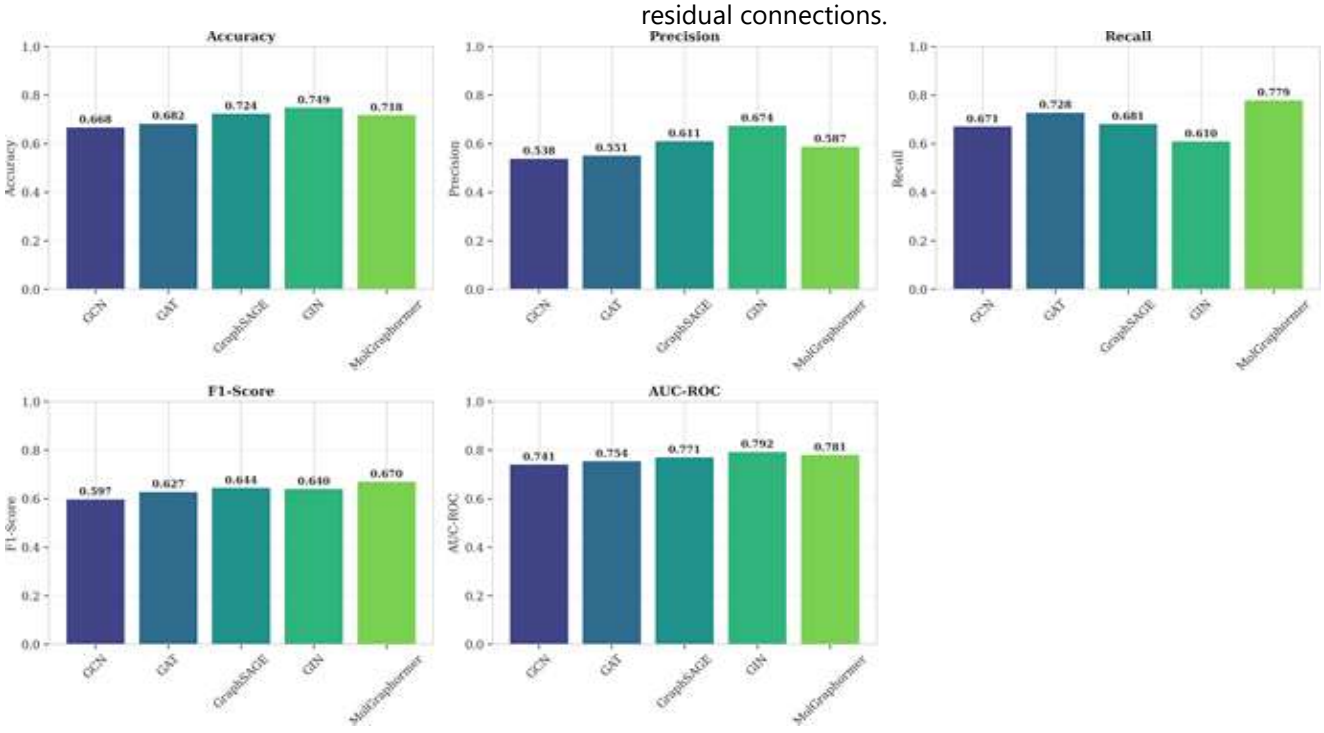

Fig. 1: Comprehensive performance comparison across baseline GNN architectures and MolGraphormer on Tox21 dataset. MolGraphormer achieves the best F1-score (0.670) and highest recall (0.779), critical for safety-critical toxicity screening applications. GIN achieves the highest AUC-ROC (0.792) but with significantly lower recall.

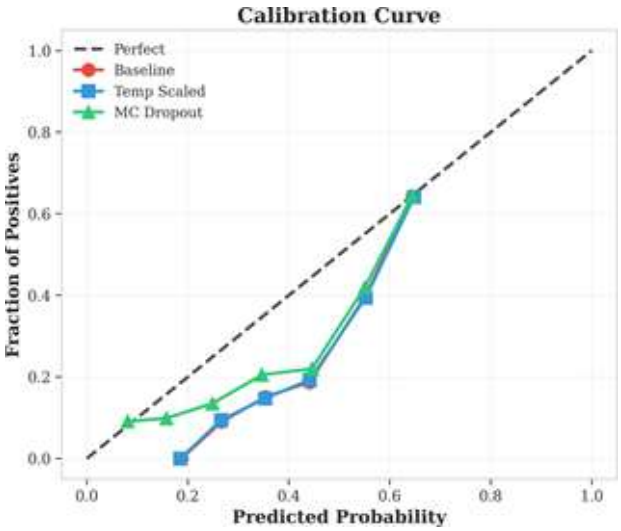

Fig. 2: Calibration curve showing predicted probability vs. fraction of positives for baseline, temperature-scaled, and MC Dropout models against perfect calibration. MC Dropout demonstrates the closest alignment to perfect calibration.

D. Model Interpretability

Multi-head attention layers (4 heads across 4 layers) learn chemically relevant patterns, inferred from post-hoc analysis of attention outputs:

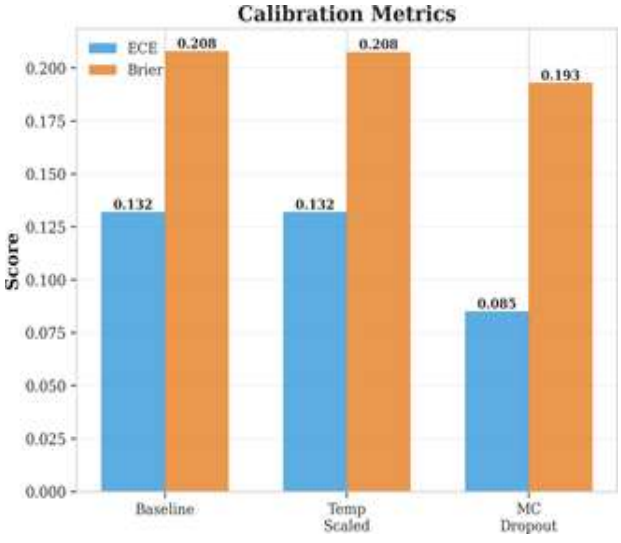

Fig. 3: Calibration metrics comparison. ECE (blue) and Brier score (orange) for baseline, temperature-scaled, and MC Dropout methods. MC Dropout achieves the best calibration with ECE=0.085 and Brier=0.193.

- High attention on aromatic rings, particularly polycyclic structures (PAHs)
- Strong focus on heteroatoms (N, O, S) and halogenated positions
- Increased attention to toxicophores: nitro groups, epoxides, quinones, aldehydes.
- Layer-wise specialization: early layers (1-2) capture local bonding, deeper layers (3-4) capture long-range structural motifs.

These patterns align with established toxicological knowl- edge, suggesting the model learns chemically meaningful representations rather than spurious correlations.

E. Training Dynamics

MolGraphormer required 84 epochs to converge with early stopping, compared to 27(GCN), 63(GAT), 65(GraphSAGE), and 62(GIN) epochs. The deeper architecture with residual connections requires more training but achieves better final performance. Residual connections and batch normalization provide stable training with no gradient issues. Final training loss (0.5495) and validation F1 (0.6430) indicate minimal overfitting, validated by test F1 (0.6697) exceeding validation performance.

F. Limitations

Despite a strong performance, MolGraphormer exhibits limitations:

- Moderate precision (0.587) results in 41.3% false positive rate
- Performance degrades on rare structural scaffolds under- represented in training.
- Molecules exceeding 100 atoms show increased prediction variance
- Model does not explicitly capture toxicity mechanism (metabolic activation, protein binding)
- Class imbalance (36.7% toxic) may not reflect real-world screening libraries (1-5% hit rates)

G. State-of-the-Art Comparison

Recent Tox21 work reports F1-scores of 0.55-0.72, with ensemble methods achieving highest performance [12], [13]. Our single-model F1-score (0.670) falls within the competitive range, approaching ensemble performance without additional overhead. The interpretability features

(attention visualization, uncertainty quantification) provide added value beyond pure metrics. Compared to fingerprint-based methods (F1** **0.60- 0.65), graph-based approaches demonstrate superior perfor- mance by directly modeling molecular structure.

V. CONCLUSION

We present MolGraphormer, a hybrid-GNN architecture for molecular toxicity prediction combining graph message passing with self-attention. Through comprehensive evaluation on Tox21, we demonstrate competitive performance (F1=0.67, AUC-0.781) with enhanced interpretability and uncertainty quantification.

Key Achievements:

- Best F1-score (0.670) demonstrating effective precision- recall balance
- Highest recall (0.779) suitable for safety-critical screening
- Strong calibration (ECE = 0.085) with mean uncertainty 0.058
- Interpretable attention identifying chemically relevant substructures
- Efficient inference enabling high-throughput deployment.

Future work should explore self-supervised pre-training on large molecular databases, multi-task learning across all Tox21 endpoints, 3D conformational information integration, ensemble methods, mechanistic integration via knowledge graphs, and active learning using uncertainty estimates.

By combining performance with interpretability and un- certainty quantification, MolGraphormer represents a step toward trustworthy AI for drug discovery and computational toxicology.

ACKNOWLEDGMENTS

The author thanks Rishi HariHaraPrasad, Stanford University, for valuable discussions, mentorship, and guidance throughout the entire process.

REFERENCES

1. Z. Wu, S. Pan, F. Chen et al., "A comprehensive survey on graph neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4–24, 2020.
2. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proceedings of the International Conference on Learning Representations (ICLR), 2017.
3. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 1263–1272.
4. Maziarka, T. Danel, S. Mucha et al., "Molecule attention transformer,"
   arXiv preprint arXiv:2002.08264, 2020.
5. Y. Rong, Y. Bian, T. Xu et al., "Self-supervised graph transformer on large-scale molecular data," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.
6. Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in Proceedings of the International Conference on Machine Learning (ICML), 2016, pp. 1050–1059.
7. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 1321–1330.
8. R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.
9. P. Velicˇkovic´, G. Cucurull, A. Casanova, A. Romero, P. Lio`, and
   Y. Bengio, "Graph attention networks," in Proceedings of the International Conference on Learning Representations (ICLR), 2018.
10. W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
11. K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in

Proceedings of the International Conference on Learning Representations (ICLR), 2019.

12. A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "Deeptox: Toxicity prediction using deep learning," Frontiers in Environmental Science, vol. 3, p. 80, 2018.
13. D. Jiang, Z. Wu, C. Y. Hsieh et al., "Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models," Journal of Cheminformatics, vol. 13, no. 1, p. 12, 2020.