Graph ML-Enabled Molecular Design Assistant

Using Graph Neural Networks

Saptarshi Ghosh, Uday Sankar De, Mangaldip Dhua,
Arnab Subhra Ghosh, Sumit Chaira and Kishan Singh

RAAISA 2026 · Paper 229
Institute of Engineering & Management, Kolkata



Molecular assay screening

A structure can support several biological questions

ToxVision estimates activity in 12 Tox21 assays from a molecular graph.



The implemented system predicts assay activity. It does not generate molecules or certify safety.

Research contribution

A common implementation

Compare GCN, GINE and GATv2 using the saved project checkpoints.

A reproducibility audit

Verify features, split membership, deployment outputs and assay-level performance.

Dataset and graph representation

7,831

original Tox21 source rows

9 + 3

atom and bond feature fields

12

assay endpoints

Seven receptor-related endpoints and five cellular stress endpoints

Missing labels are masked separately for each assay. These are two-dimensional graph features.

Three graph architectures

GCN

Normalized neighbor aggregation
Connectivity only

336,396

trainable parameters

GINE

Edge-aware message aggregation
Bond attributes consumed

602,672

trainable parameters

GATv2

Dynamic attention
Four concatenated heads

538,380

trainable parameters

Training and historical split

6,264 training

783 validation

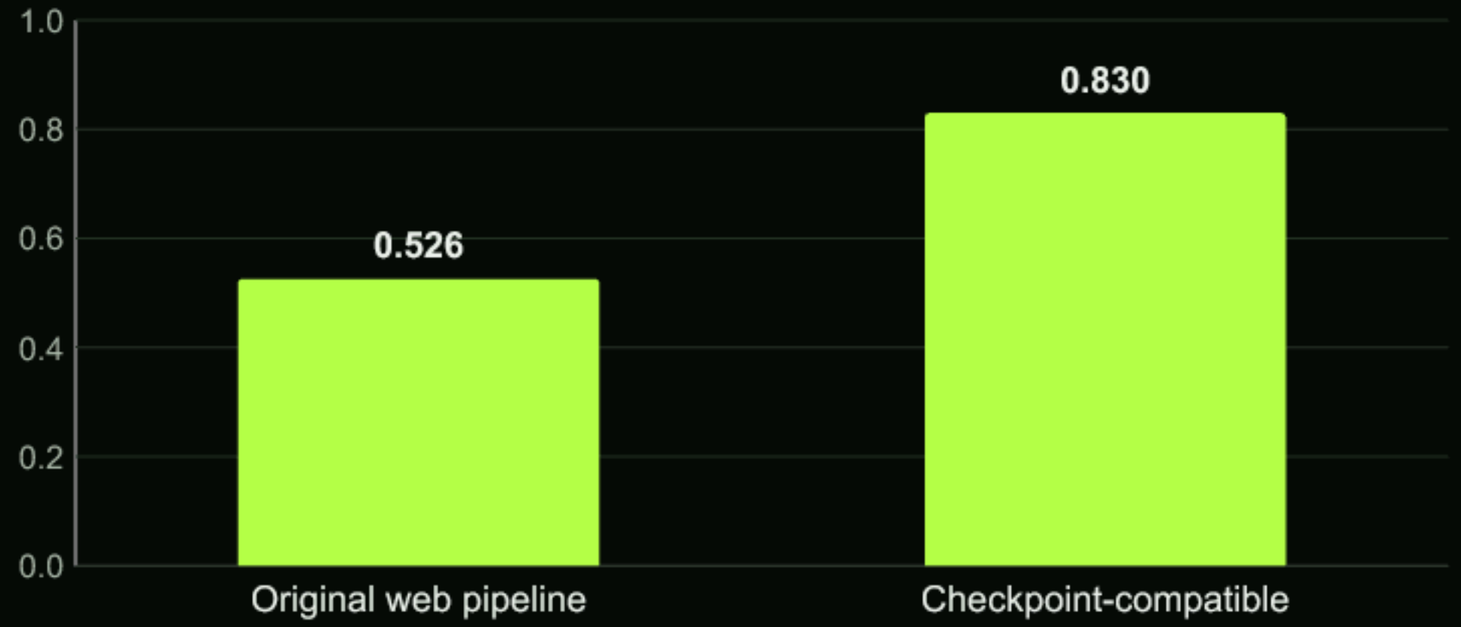
784 test

Fingerprint grouping allows shared Murcko scaffolds across partitions.

171 shared training/test scaffold keys
Frozen checkpoints, no five-seed retraining

The deployment feature mismatch

Same frozen GCN weights and the same recorded test compounds

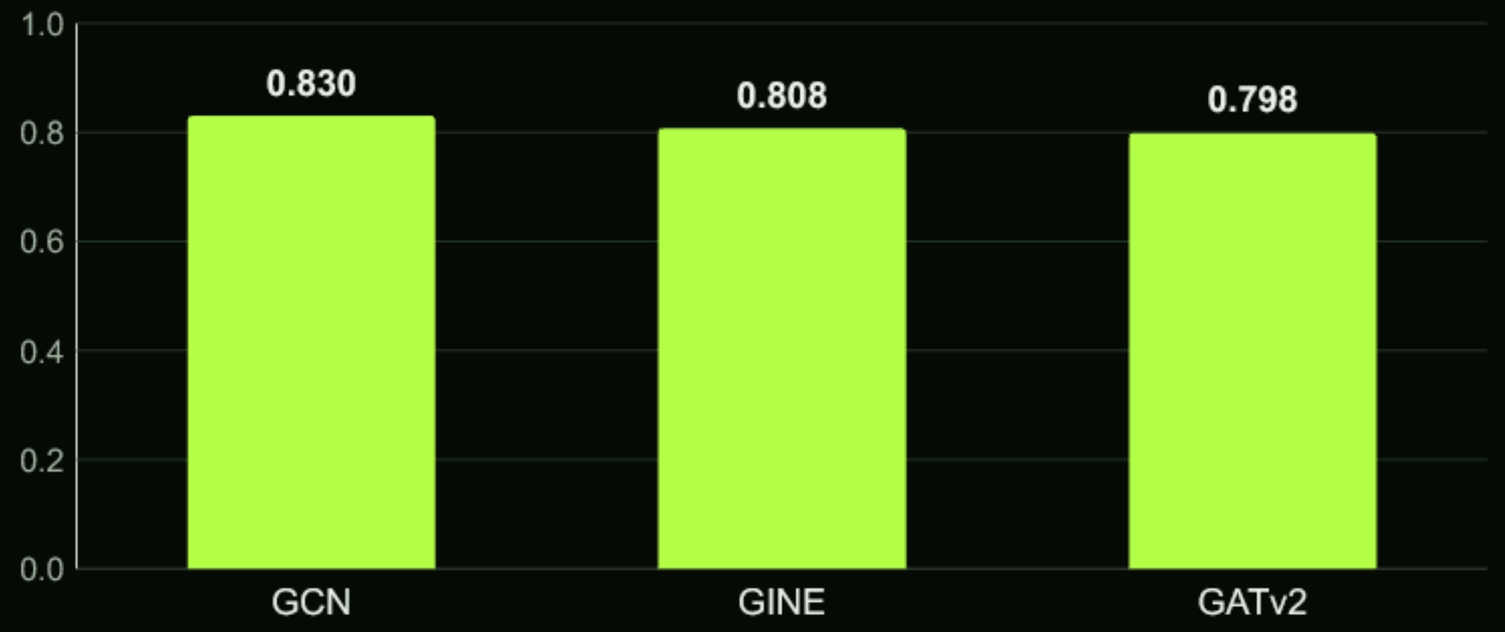


0.526 vs 0.830

Correct PyG encoding
Remove unsupported z-scoring
Verify service prediction parity

Original recorded test results

784 compounds, with missing labels excluded per endpoint



GCN 0.830

95% bootstrap interval
0.807 to 0.854

Ranking is conditional on these checkpoints and this partition.

An expanded 1,000-molecule evaluation

598

NIH final-evaluation compounds

285

NIH leaderboard compounds

117

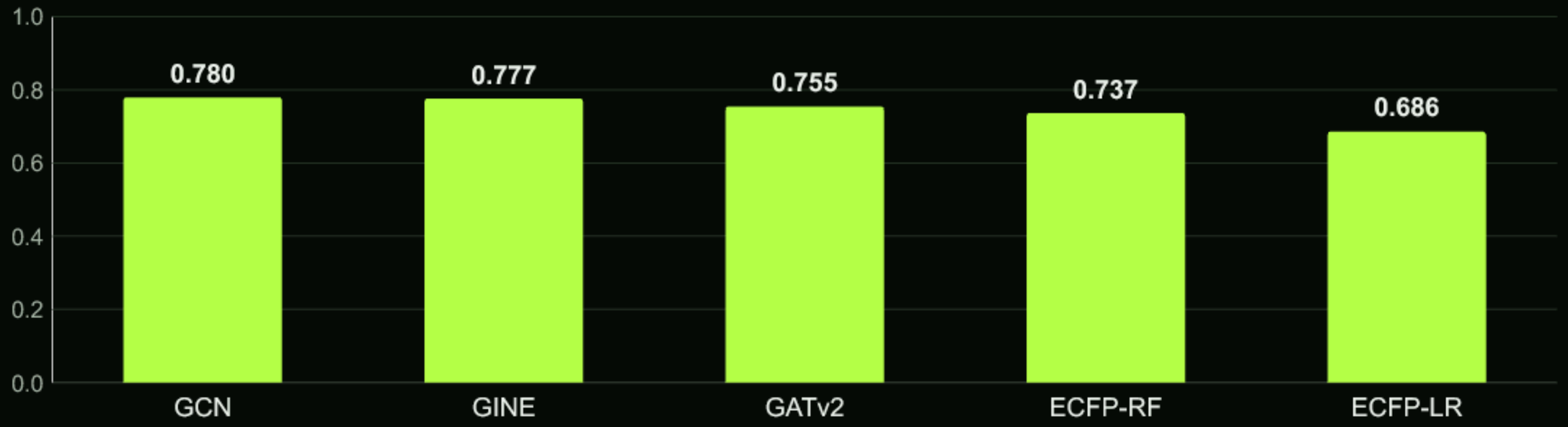
recorded held-out compounds

883 external-source compounds reported separately

1,000 distinct parent structures and 10,623 observed assay labels

Expanded model comparison

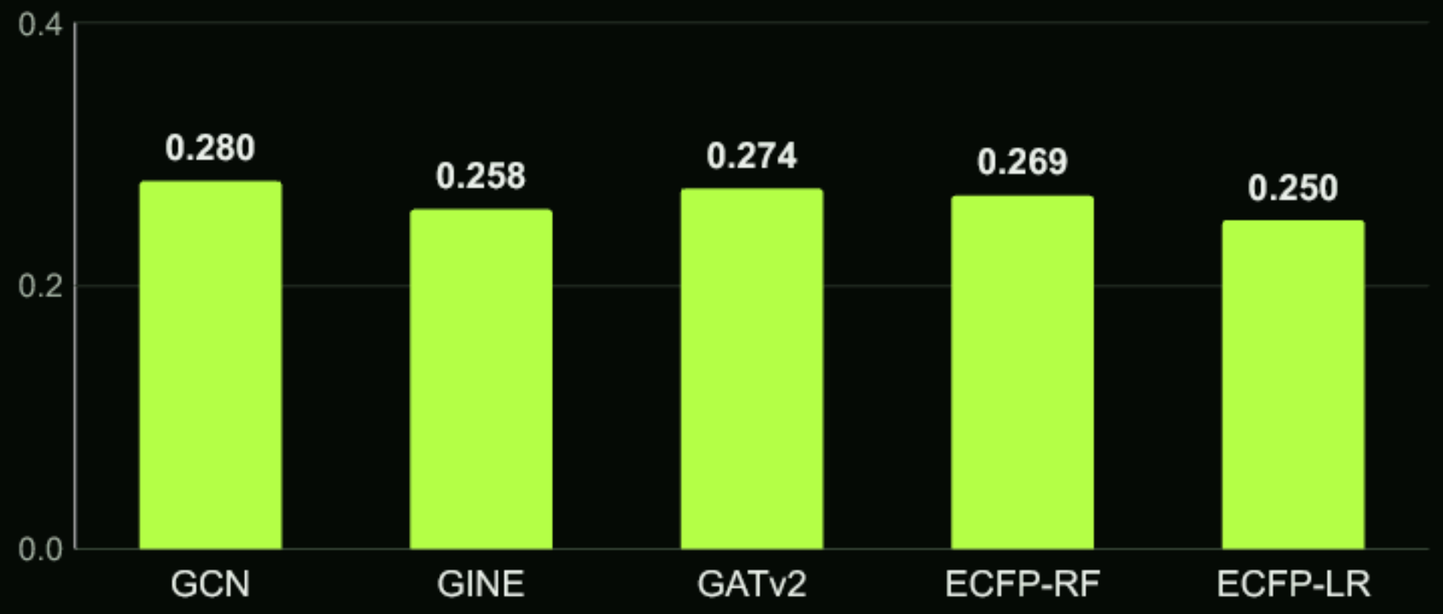
Macro ROC-AUC on the mixed 1,000-compound cohort



GCN 0.780 · GINE 0.777 · GATv2 0.755

Ranking and operating-point errors

Macro average precision on the mixed 1,000-compound cohort

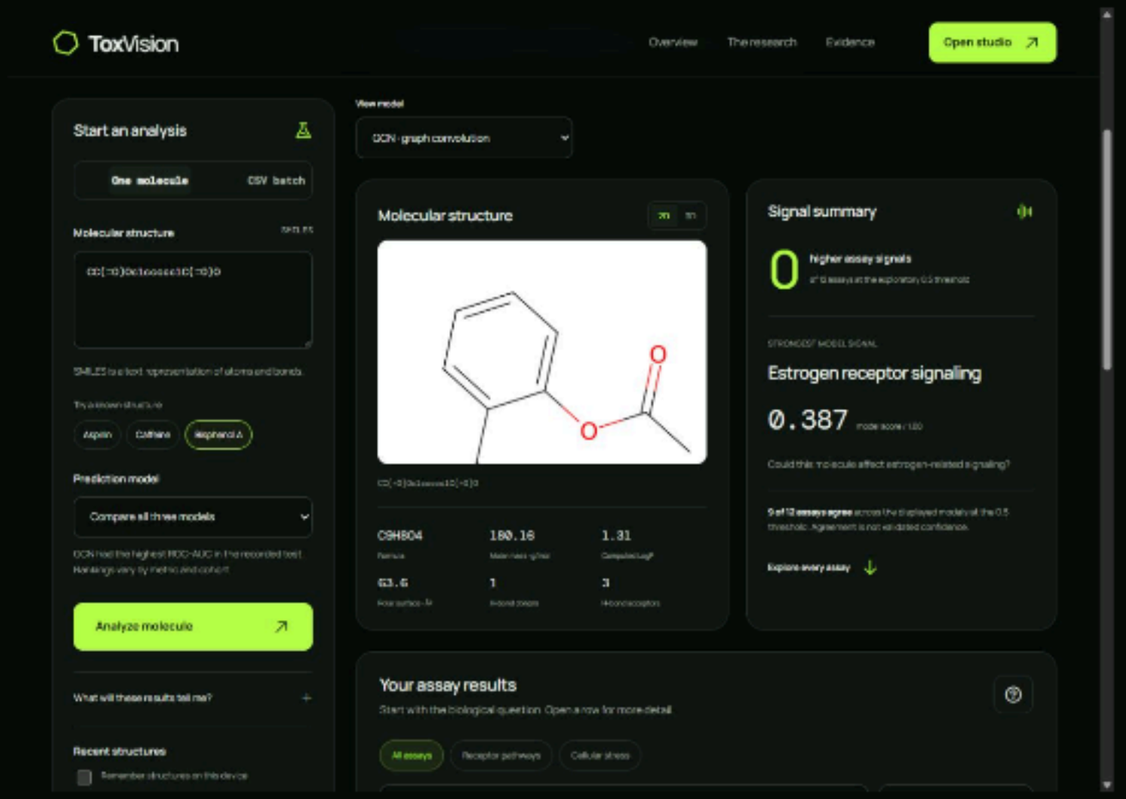


Class-weighted scores are uncalibrated.

AUC is not accuracy

GCN macro precision: 0.180
GCN macro recall: 0.661
Exploratory threshold: 0.5

A clearer molecular analysis studio



Biological questions first

Plain-language assay names
Structure and descriptors
Separate model comparisons
CSV and printable results

Checking the deployed service

1,000

molecules through the HTTP API

3,000

model/molecule parity comparisons

36 / 36

service checks passed

Maximum absolute score difference: 0.000000298

Ten sequential batches of 100 compounds, comparing all three GNNs

Limits and next experiments

Scaffold-disjoint training	Keep complete Murcko groups in one partition.
Independent training seeds	Repeat all architectures under matched tuning budgets.
Calibration and thresholds	Lock each assay’s operating point using validation data.
Prospective validation	Test important findings with new experimental measurements.

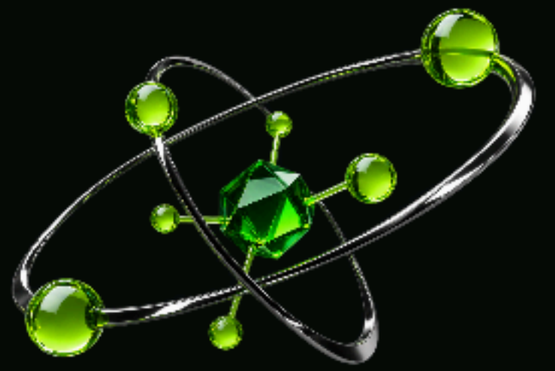
Conclusion

Reliable screening needs a reliable feature contract.

The saved models reproduce the original benchmark result.
The repaired studio matches checkpoint inference.
Measured assay labels support a broader, qualified evaluation.

Questions and research materials

Thank you



Saptarshi Ghosh and co-authors
Institute of Engineering & Management, Kolkata

Paper 229
Graph ML-Enabled Molecular Design Assistant Using Graph Neural Networks