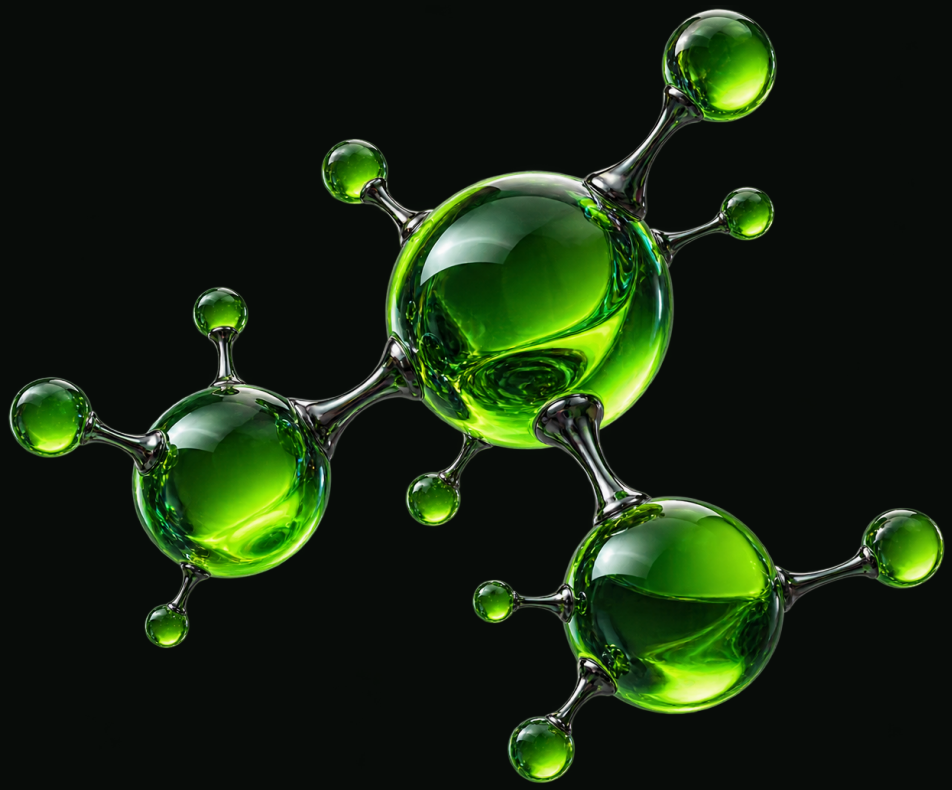
1,000 molecules. Measured assay evidence.

A detailed evaluation of GCN, GINE and GATv2, with matched-partition fingerprint baselines and deployed-service verification.



1,000 distinct parent structures
883 external-source compounds and 117 recorded held-out compounds
10,623 observed assay labels
3,000 deployed model/molecule comparisons

Retrospective research evaluation. No new laboratory or clinical safety measurements.

Study scope and cohort construction

Published experimental assay labels supply the reference outcomes

Source cohort	Molecules	Role
NIH final evaluation	598	Additional external-source evaluation
NIH leaderboard test	285	Additional external-source evaluation
Recorded held-out test subset	117	Historical benchmark evidence, reported separately
Total	1,000	Unique stereochemistry-free fragment-parent keys

Fixed selection protocol

Retain all 598 eligible NIH final-evaluation compounds from the earlier structure-overlap audit. Parse the NIH leaderboard SDF labels and join them to the published SMILES using the sample/batch identifier. Remove invalid structures and parent structures overlapping any original source record or the selected final-evaluation cohort. This retains 285 leaderboard compounds.

Fill the remaining 117 places using ascending source row IDs from the recorded test partition. Exclude parent structures overlapping the original training or validation partitions, the external cohorts, or an already selected record. The selection does not use predicted scores, label correctness, or assay outcomes.

What the reference labels mean

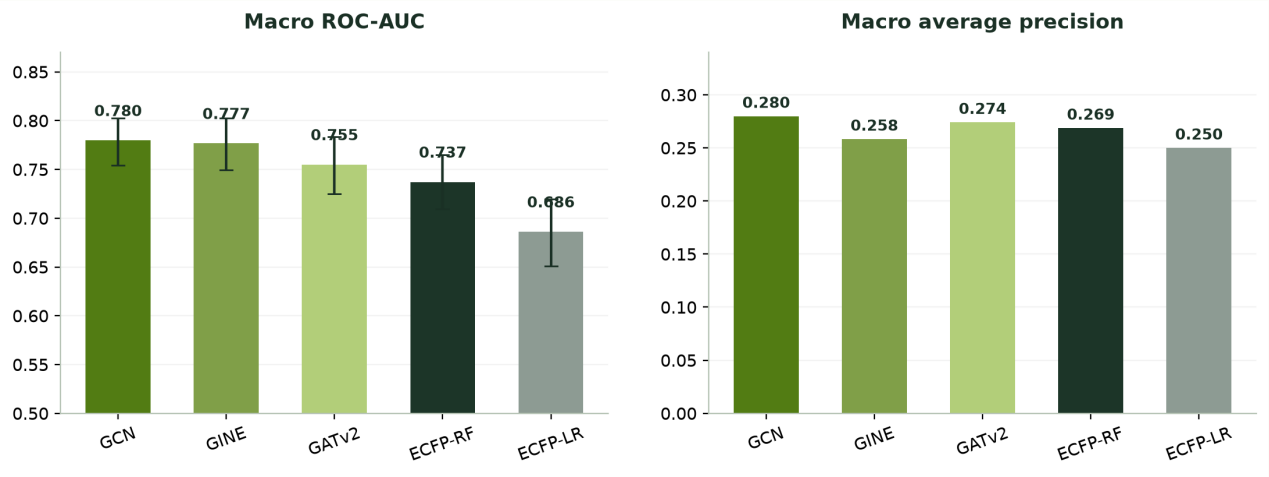
A label of 1 means active and 0 means inactive in a specific NIH/Tox21 assay. Missing labels remain missing and are excluded from each endpoint’s metrics. The reference does not label a molecule as universally safe or toxic.

What this comparison can establish

The study assesses retrospective discrimination and operating-point behavior on existing assay measurements. It does not establish prospective discovery performance, human toxicity, exposure-specific safety, or reduction in laboratory work. The full 1,000-compound cohort mixes sources, so the 883-compound external subset is always reported separately.

Overall model performance

Full mixed cohort of 1,000 compounds, macro averages over twelve assays



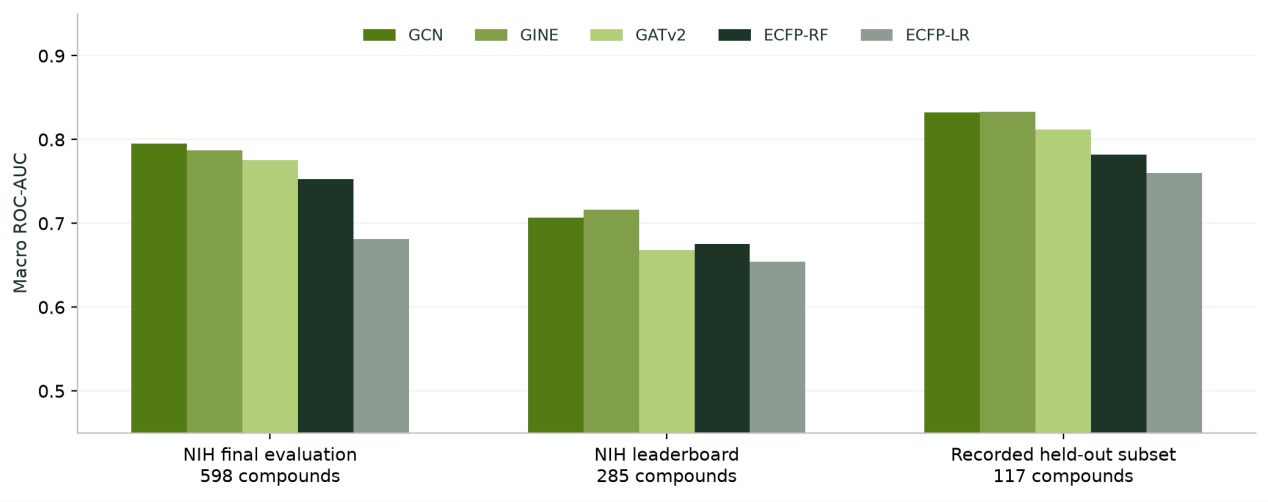
Model	AUC	95% CI	AP	P	R	F1
GCN	0.780	[0.754, 0.803]	0.280	0.180	0.661	0.278
GINE	0.777	[0.749, 0.802]	0.258	0.163	0.720	0.259
GATv2	0.755	[0.725, 0.783]	0.274	0.164	0.655	0.257
ECFP-RF	0.737	[0.709, 0.765]	0.269	0.381	0.244	0.270
ECFP-LR	0.686	[0.651, 0.720]	0.250	0.235	0.393	0.286

GCN has the highest observed macro ROC-AUC at 0.780. GINE scores 0.777 and GATv2 0.755. The GCN-GINE paired interval includes zero, so this evaluation does not establish a reliable separation between those two frozen networks.

The fingerprint random forest has higher macro precision at the fixed 0.5 cutoff than the GNNs, while the GNNs generally have higher recall. Architecture choice therefore depends on the metric and intended operating point, not only AUC.

Source cohorts change the picture

External-source performance is separated from historical benchmark performance



Cohort	GCN	GINE	GATv2	ECFP-RF	ECFP-LR
NIH final evaluation	0.795	0.787	0.776	0.753	0.681
NIH leaderboard test	0.707	0.716	0.668	0.675	0.654
Recorded held-out test subset	0.833	0.833	0.812	0.782	0.760
External only	0.771	0.770	0.747	0.734	0.672

External-only summary

Across the 883 external-source compounds, macro ROC-AUC is GCN 0.771, GINE 0.770, GATv2 0.747, ECFP-RF 0.734, ECFP-LR 0.672.

Cohort metrics use the labels available in each source and give equal weight to endpoints with defined AUC. They do not use a shared fixed positive prevalence. A small held-out subset and different assay coverage can create substantial shifts in observed scores.

These filtered subsets are not the official challenge leaderboard population. Scores should not be compared with published challenge rankings as if the protocols were identical.

Assay-level discrimination

All five fitted models on the 1,000-compound mixed cohort



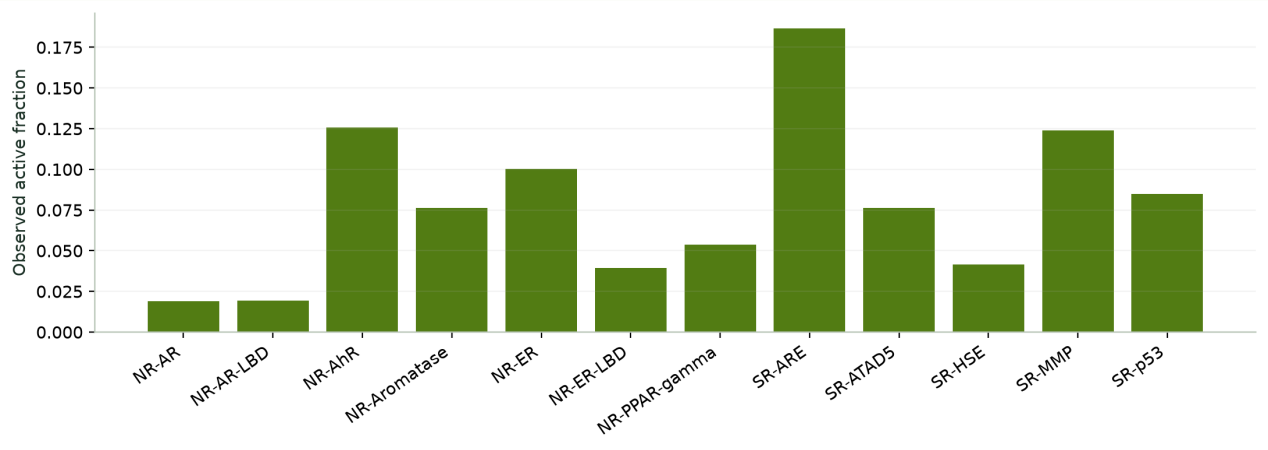
How to read this figure

Each cell shows one model's ROC-AUC for one endpoint after excluding missing labels. AUC 0.5 corresponds to chance ranking under the usual binary interpretation. A higher AUC indicates better ranking on this sample, not better threshold accuracy or a verified biological explanation.

The models do not lead on every assay. Differences in positive support, molecular composition and the learned score distribution can change the preferred model from one endpoint to another.

Label coverage and class imbalance

10,623 observed molecule-assay labels out of 12,000 possible pairs

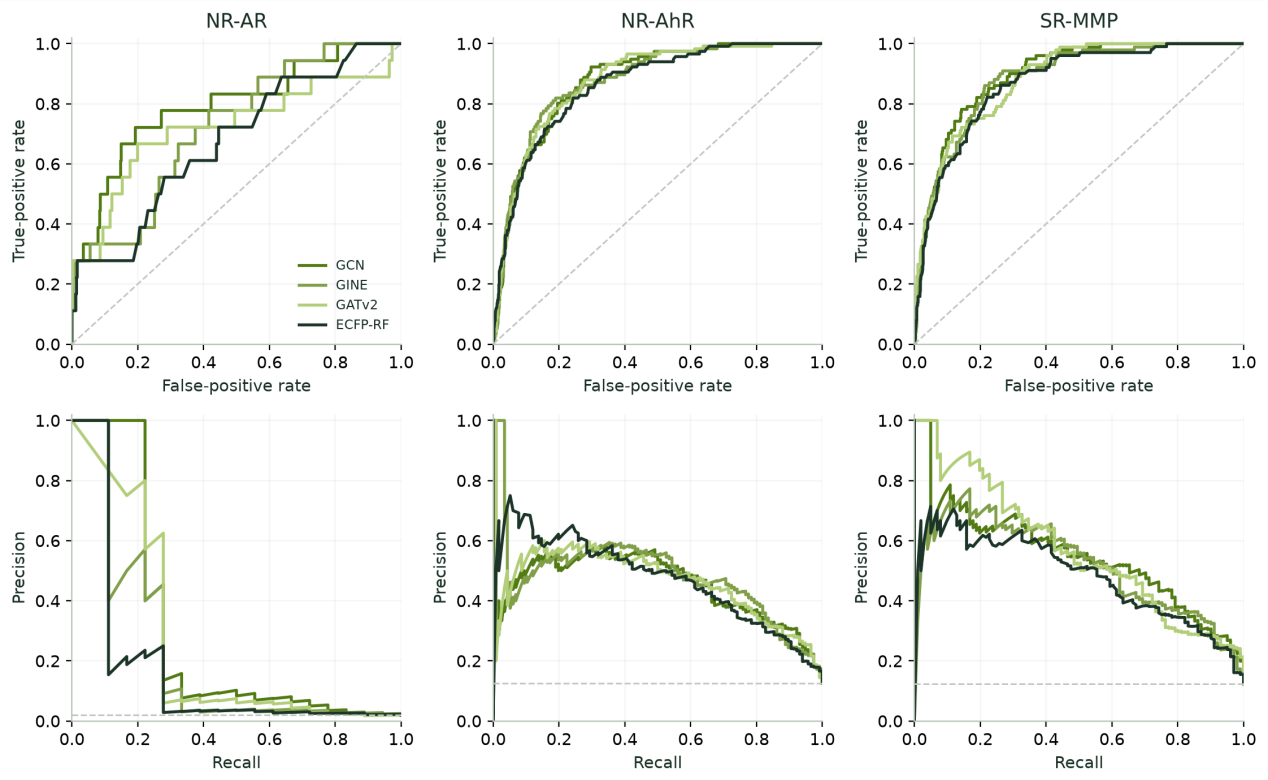


Endpoint	Observed	Active	Inactive	Missing	Active share
NR-AR	940	18	922	60	1.9%
NR-AR-LBD	885	17	868	115	1.9%
NR-AhR	923	116	807	77	12.6%
NR-Aromatase	775	59	716	225	7.6%
NR-ER	828	83	745	172	10.0%
NR-ER-LBD	939	37	902	61	3.9%
NR-PPAR-gamma	910	49	861	90	5.4%
SR-ARE	814	152	662	186	18.7%
SR-ATAD5	945	72	873	55	7.6%
SR-HSE	918	38	880	82	4.1%
SR-MMP	815	101	714	185	12.4%
SR-p53	931	79	852	69	8.5%

Rare active labels make overall classification accuracy easy to misread: an always-inactive classifier can appear accurate while detecting no active compounds. This report prioritizes ROC-AUC, average precision, precision, recall, F1 and confusion counts.

ROC and precision-recall behavior

Illustrative endpoints drawn from the tested measurements



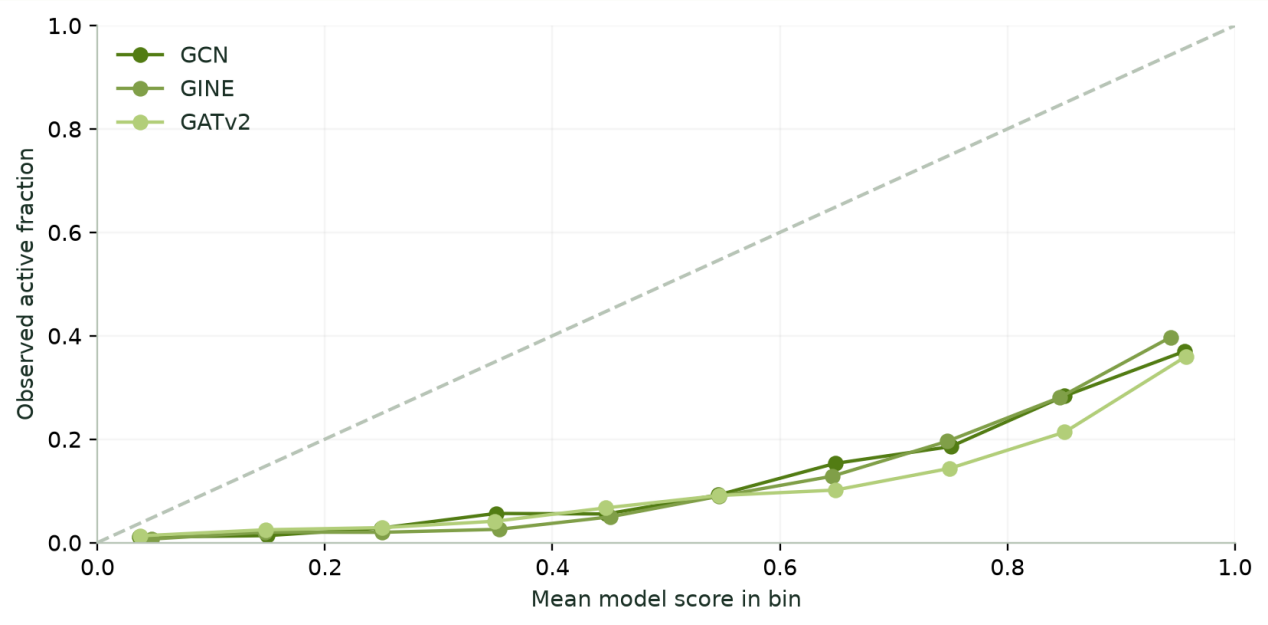
Two views of discrimination

ROC curves compare true-positive and false-positive rates across thresholds. Precision-recall curves show the fraction of predicted actives that are observed active, against the fraction of measured actives recovered. Dashed PR baselines indicate observed activity prevalence.

Average precision uses the standard stepwise definition. It is not presented as trapezoidal PR-AUC. Curves are generated directly from the released prediction arrays; no published chart has been reused.

Operating-point errors and calibration

The 0.5 cutoff audits the implemented display, not a validated safety policy



Model	TP	FP	FN	TN	Macro Brier
GCN	600	2412	221	7390	0.177
GINE	643	3034	178	6768	0.202
GATv2	596	2687	225	7115	0.195
ECFP-RF	254	344	567	9458	0.073
ECFP-LR	378	1029	443	8773	0.107

Counts pool observed molecule-assay pairs and therefore are not independent molecule counts. A molecule can contribute up to twelve outcomes. Brier scores are averaged per endpoint before macro averaging.

The reliability plot pools endpoints as a diagnostic overview. It does not establish calibration for any single assay. Positive-class weighting and historical use of full-dataset prevalence constrain probability interpretation. A lower model score never establishes that a molecule is safe.

Uncertainty and paired model differences

Conditional molecule-level bootstrap, 1,000 replicates with seed 229

Comparison	Full cohort AUC-margin interval	External-only interval
GCN minus GINE	[-0.012, 0.018]	[-0.018, 0.020]
GCN minus GATv2	[0.007, 0.045]	[-0.000, 0.046]
GCN minus ECFP-RF	[0.019, 0.065]	[0.011, 0.061]
GCN minus ECFP-LR	[0.058, 0.127]	[0.055, 0.140]

What is resampled

Each replicate samples molecules with replacement and carries all their observed and missing assay labels together. The same sampled rows are used for every model. Endpoint AUCs are computed independently, then averaged when all twelve endpoints contain both classes.

What the intervals do not include

The intervals hold the fitted weights fixed. They do not include neural training-seed variation, hyperparameter-search uncertainty, scaffold-cluster dependence, assay measurement uncertainty, or adjustments for multiple comparisons.

Interpretation

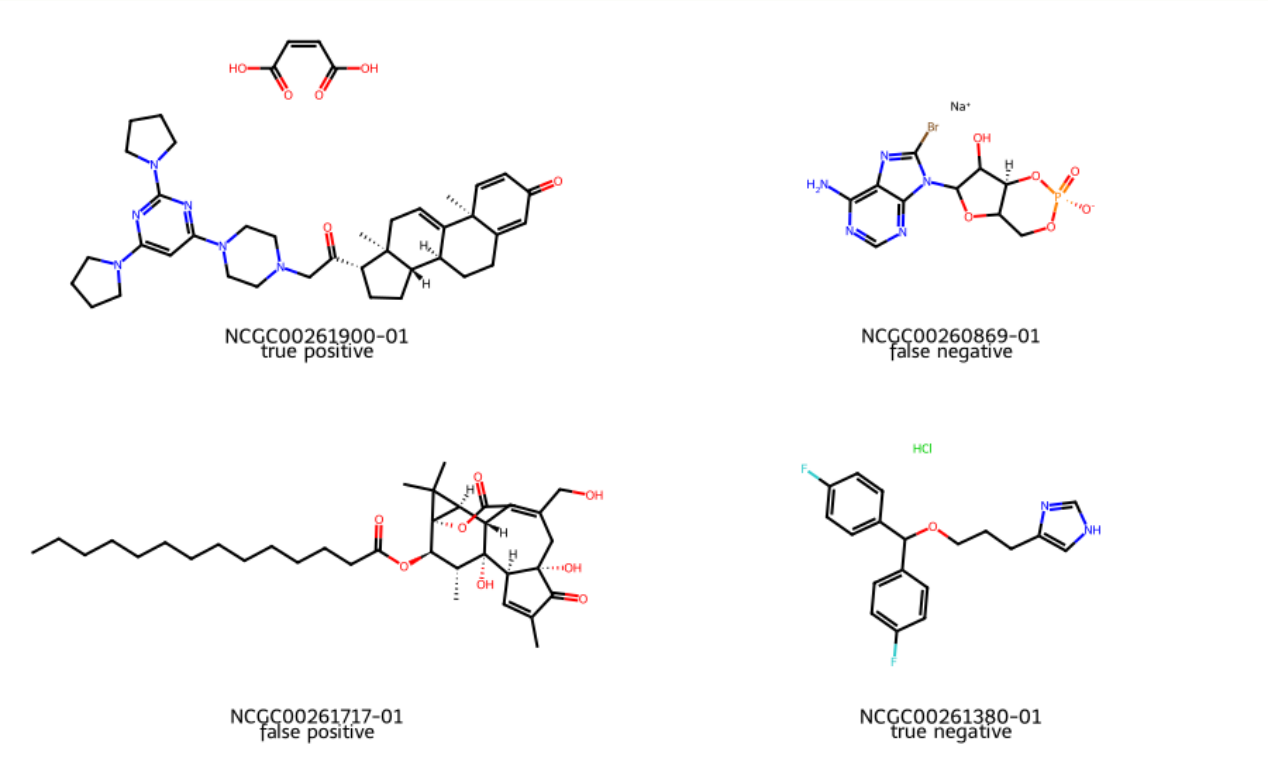
The GCN-GINE full-cohort interval contains zero. The GCN-GATv2 interval is positive in this analysis, but remains conditional on these checkpoints and this selected cohort. Results do not establish general architectural superiority or isolate a chemical mechanism.

The original models have not been retrained across five independent seeds. Bootstrap replicates must not be described as repeated neural training runs.

Predictions against actual assay outcomes

Four deterministic NR-AR examples from the retained measured cohort

Sample ID	GCN outcome	Measured	GCN	GINE	GATv2
NCGC00261900-01	true positive	1	0.937	0.959	0.999
NCGC00260869-01	false negative	1	0.381	0.377	0.042
NCGC00261717-01	false positive	0	0.894	0.660	0.882
NCGC00261380-01	true negative	0	0.087	0.225	0.316

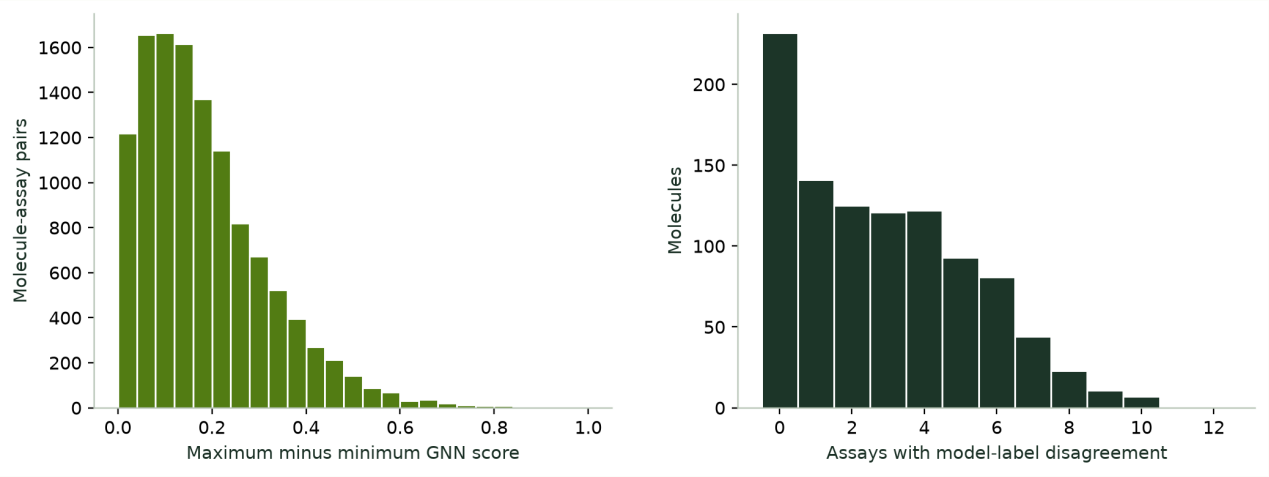


Each example is the first source-order compound satisfying its stated outcome at cutoff 0.5. This deterministic selection illustrates true positives, false negatives, false positives and true negatives. It is not a random sample or a substitute for the full confusion matrix.

Verify each row by joining its sample identifier to the NIH label source and selecting the NR-AR column. The downloadable case file includes the original SMILES, source cohort and all five model scores.

When the models disagree

Prediction agreement is not a validated uncertainty estimate



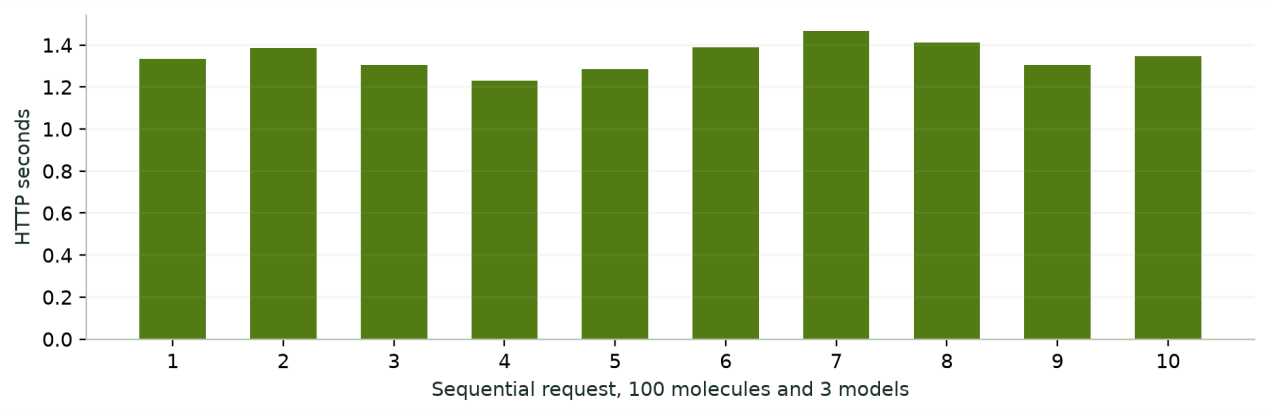
Diagnostic	Measured result
Molecule-assay pairs with label disagreement	2,854 / 12,000
Molecules with at least one disagreement	768 / 1,000
Median score spread	0.156
Mean score spread	0.185

Score spread is the largest minus the smallest GNN score for a given molecule and endpoint. Label disagreement occurs when the three GNNs do not all fall on the same side of 0.5. This diagnostic includes predicted pairs without observed labels because it measures model behavior rather than correctness.

The studio displays each model separately and exposes these differences. It does not collapse all three models into an unvalidated ensemble probability or call agreement “confidence”.

Testing the deployed application

The same 1,000 molecules also pass through the redesigned Flask service



Verification	Result
Molecules sent through the HTTP prediction API	1,000
GNN model/molecule comparisons	3,000
Probability values compared	36,000
Comparisons within absolute error 3e-6	3,000 / 3,000
Largest absolute score error	0.000000298
Service behavior checks	36 / 36 passed
Ten sequential batch requests	13.46 seconds total

Each HTTP request contains 100 compounds and selects all three GNNs. Responses are matched by source SMILES and model name to the direct inference arrays. The checks also cover malformed input, CSV limits, structure descriptors, branded routes and result-specific export isolation.

The timing includes loopback HTTP, CSV parsing, graph conversion, per-molecule model execution and JSON serialization. CSV batches do not generate 2D/3D depictions. This is a local sequential benchmark, not production concurrency, GPU throughput, or a usability study.

Feature contract and model provenance

Saved neural weights are unchanged; the serving pipeline is corrected

Component	Verified implementation
Atom inputs	PyG nine-field categorical SMILES encoding, model-internal clamp
Bond inputs	Three categorical fields: type, stereo, conjugation
Normalization	No inference z-scoring for these frozen checkpoints
GCN	4 layers, width 256, 336,396 parameters, topology only
GINE	4 layers, width 256, 602,672 parameters, bond-aware messages
GATv2	4 layers, width 256, 538,380 parameters, four heads
Readout	Concatenated mean and sum pooling, 512-to-256-to-12 head
Fingerprint RF	300 trees, balanced weights, sqrt features, seed 229
Fingerprint LR	L2, C=1, balanced weights, liblinear, seed 229

The two fingerprint baselines use radius-two, 2,048-bit Morgan fingerprints and the recorded training partition. Each task excludes missing training labels and invalid structures. No external-set hyperparameter search or threshold tuning is performed.

Environment

Python 3.14.5, PyTorch 2.15.0+cpu, PyG 2.8.1, RDKit 2026.09.2. Neural inference uses CPU with four PyTorch threads. Starting repository commit: e705e8f7e02a6fc3d88c8f9c04913cbd5b7956bf.

Limits and the next experiment

Retrospective assay evidence supports qualified screening claims

Historical training limitations

The recorded fingerprint grouping does not enforce Bemis-Murcko scaffold separation. The original notebook derives positive-class weights from all source labels, including held-out prevalence. Original independent neural training seeds are unavailable. The audit retains these limitations rather than silently claiming that frozen-checkpoint re-evaluation fixes training design.

External data limitations

Parent-structure filtering does not establish scaffold disjointness, remove every tautomeric or assay-related dependency, or model exposure and metabolism. Public challenge datasets are retrospective and may have been studied before. The 117 historical test records are explicitly separated from the 883 external-source records.

Clinical and experimental limits

These outcomes concern named assay activity. There are no new wet-lab measurements in this report and no demonstration of human safety, clinical efficacy, laboratory cost reduction, or validated atom-level explanations.

Priority next experiment

Retrain every model on a released scaffold-disjoint split with training-only preprocessing and class weights, matched tuning budgets and at least five declared seeds. Lock calibration and operating thresholds on validation data. Then evaluate on fresh external compounds that have not been used for model selection.

Use in the redesigned studio

Use scores to prioritize research questions. Show separate assay meanings and model outputs. Verify important predictions using measured assay evidence and appropriate laboratory validation. Do not interpret a lower signal as a general safety statement.

GCN: complete endpoint metrics

All 1,000 compounds; missing labels excluded independently per assay

Endpoint	N	Active	AUC	AP	P	R	F1
NR-AR	940	18	0.794	0.299	0.100	0.500	0.167
NR-AR-LBD	885	17	0.750	0.201	0.074	0.353	0.122
NR-AhR	923	116	0.877	0.460	0.308	0.922	0.462
NR-Aromatase	775	59	0.785	0.227	0.187	0.644	0.290
NR-ER	828	83	0.728	0.321	0.195	0.627	0.297
NR-ER-LBD	939	37	0.722	0.190	0.091	0.541	0.156
NR-PPAR-gamma	910	49	0.722	0.131	0.115	0.551	0.191
SR-ARE	814	152	0.788	0.428	0.330	0.757	0.459
SR-ATAD5	945	72	0.757	0.193	0.163	0.694	0.264
SR-HSE	918	38	0.747	0.129	0.086	0.684	0.152
SR-MMP	815	101	0.897	0.553	0.325	0.861	0.472
SR-p53	931	79	0.791	0.225	0.184	0.797	0.299

Confusion counts at the exploratory 0.5 threshold

Endpoint	TP	FP	FN	TN
NR-AR	9	81	9	841
NR-AR-LBD	6	75	11	793
NR-AhR	107	240	9	567
NR-Aromatase	38	165	21	551
NR-ER	52	215	31	530
NR-ER-LBD	20	200	17	702
NR-PPAR-gamma	27	207	22	654
SR-ARE	115	234	37	428
SR-ATAD5	50	257	22	616
SR-HSE	26	277	12	603
SR-MMP	87	181	14	533
SR-p53	63	280	16	572

N counts observed labels for that endpoint. AUC and AP use continuous scores. P, R and F1 use the fixed 0.5 display threshold. No calibration or external-set threshold optimization is applied.

GINE: complete endpoint metrics

All 1,000 compounds; missing labels excluded independently per assay

Endpoint	N	Active	AUC	AP	P	R	F1
NR-AR	940	18	0.723	0.224	0.107	0.333	0.162
NR-AR-LBD	885	17	0.826	0.141	0.071	0.588	0.127
NR-AhR	923	116	0.875	0.484	0.251	0.897	0.392
NR-Aromatase	775	59	0.763	0.195	0.147	0.729	0.245
NR-ER	828	83	0.750	0.315	0.206	0.687	0.317
NR-ER-LBD	939	37	0.727	0.140	0.087	0.595	0.152
NR-PPAR-gamma	910	49	0.717	0.109	0.100	0.776	0.178
SR-ARE	814	152	0.790	0.458	0.319	0.829	0.461
SR-ATAD5	945	72	0.718	0.159	0.131	0.681	0.220
SR-HSE	918	38	0.760	0.135	0.078	0.763	0.142
SR-MMP	815	101	0.885	0.520	0.297	0.911	0.448
SR-p53	931	79	0.784	0.217	0.161	0.848	0.271

Confusion counts at the exploratory 0.5 threshold

Endpoint	TP	FP	FN	TN
NR-AR	6	50	12	872
NR-AR-LBD	10	131	7	737
NR-AhR	104	310	12	497
NR-Aromatase	43	249	16	467
NR-ER	57	220	26	525
NR-ER-LBD	22	230	15	672
NR-PPAR-gamma	38	341	11	520
SR-ARE	126	269	26	393
SR-ATAD5	49	325	23	548
SR-HSE	29	342	9	538
SR-MMP	92	218	9	496
SR-p53	67	349	12	503

N counts observed labels for that endpoint. AUC and AP use continuous scores. P, R and F1 use the fixed 0.5 display threshold. No calibration or external-set threshold optimization is applied.

GATv2: complete endpoint metrics

All 1,000 compounds; missing labels excluded independently per assay

Endpoint	N	Active	AUC	AP	P	R	F1
NR-AR	940	18	0.720	0.240	0.071	0.500	0.125
NR-AR-LBD	885	17	0.707	0.196	0.084	0.471	0.143
NR-AhR	923	116	0.876	0.466	0.258	0.940	0.405
NR-Aromatase	775	59	0.783	0.253	0.176	0.644	0.276
NR-ER	828	83	0.718	0.331	0.178	0.590	0.274
NR-ER-LBD	939	37	0.693	0.147	0.083	0.568	0.144
NR-PPAR-gamma	910	49	0.639	0.101	0.099	0.408	0.159
SR-ARE	814	152	0.784	0.433	0.342	0.816	0.482
SR-ATAD5	945	72	0.726	0.181	0.134	0.694	0.224
SR-HSE	918	38	0.756	0.152	0.084	0.632	0.148
SR-MMP	815	101	0.884	0.573	0.309	0.792	0.444
SR-p53	931	79	0.773	0.214	0.156	0.810	0.262

Confusion counts at the exploratory 0.5 threshold

Endpoint	TP	FP	FN	TN
NR-AR	9	117	9	805
NR-AR-LBD	8	87	9	781
NR-AhR	109	313	7	494
NR-Aromatase	38	178	21	538
NR-ER	49	226	34	519
NR-ER-LBD	21	233	16	669
NR-PPAR-gamma	20	183	29	678
SR-ARE	124	239	28	423
SR-ATAD5	50	324	22	549
SR-HSE	24	263	14	617
SR-MMP	80	179	21	535
SR-p53	64	345	15	507

N counts observed labels for that endpoint. AUC and AP use continuous scores. P, R and F1 use the fixed 0.5 display threshold. No calibration or external-set threshold optimization is applied.

Fingerprint baselines: endpoint detail

Fixed baseline hyperparameters, same recorded training partition

ECFP-RF

Endpoint	AUC	AP	P	R	F1
NR-AR	0.674	0.172	0.235	0.222	0.229
NR-AR-LBD	0.649	0.172	0.333	0.176	0.231
NR-AhR	0.861	0.475	0.467	0.603	0.526
NR-Aromatase	0.762	0.278	0.636	0.119	0.200
NR-ER	0.750	0.329	0.420	0.349	0.382
NR-ER-LBD	0.657	0.180	0.273	0.162	0.203
NR-PPAR-gamma	0.677	0.132	0.222	0.041	0.069
SR-ARE	0.738	0.380	0.431	0.368	0.397
SR-ATAD5	0.743	0.269	0.583	0.097	0.167
SR-HSE	0.740	0.166	0.211	0.105	0.140
SR-MMP	0.869	0.483	0.505	0.554	0.528
SR-p53	0.725	0.187	0.256	0.127	0.169

ECFP-LR

Endpoint	AUC	AP	P	R	F1
NR-AR	0.631	0.211	0.095	0.222	0.133
NR-AR-LBD	0.603	0.179	0.138	0.235	0.174
NR-AhR	0.822	0.491	0.350	0.664	0.458
NR-Aromatase	0.660	0.229	0.264	0.237	0.250
NR-ER	0.679	0.246	0.215	0.530	0.306
NR-ER-LBD	0.673	0.180	0.195	0.405	0.263
NR-PPAR-gamma	0.579	0.108	0.205	0.163	0.182
SR-ARE	0.704	0.352	0.350	0.520	0.418
SR-ATAD5	0.744	0.270	0.330	0.403	0.362
SR-HSE	0.662	0.104	0.124	0.368	0.185
SR-MMP	0.820	0.428	0.348	0.624	0.447
SR-p53	0.656	0.200	0.201	0.342	0.254

The complete CSV additionally contains source-cohort metrics, confusion counts and Brier scores for both baselines.

Sources, files and conference context

Checked on 21 November 2026

NIH Tox21 challenge data
https://tripod.nih.gov/tox21/challenge/data.jsp
Final-evaluation SMILES and labels, leaderboard SMILES and labeled SDF.

MoleculeNet
https://doi.org/10.1039/C7SC02664A
Benchmark context and original Tox21 source dataset.

PyG SMILES feature source
https://pytorch-geometric.readthedocs.io/en/latest/_modules/torch_geometric/utils/smiles.html
Categorical graph-encoding contract.

RDKit scaffold documentation
https://www.rdkit.org/docs/source/rdkit.Chem.Scaffolds.MurckoScaffold.html
Molecular scaffold interpretation.

Project code and checkpoints
https://github.com/eliot-99/Graph-ML-Enabled-Molecular-Design-Assistant-using-Graph-Neural-Networks
Repository starting commit and unchanged neural checkpoints.

RAAISA program
https://raaisa.uem.edu.in/program
Program currently says it will be uploaded shortly. No PPT format or speaking duration is published here.

RAAISA author guidelines
https://raaisa.uem.edu.in/authors-guidlines
Manuscript templates are linked. No presentation template was found in the checked public guidance.

Reproducible deliverables

The data package contains all 1,000 identifiers and source SMILES, parent keys, measured labels, missing-label masks, all model scores, source-cohort metrics, paired-bootstrap summaries, selected observed examples, deployed prediction parity and batch timings. The validation scripts and source input files accompany the package.

Presentation format

The conference presentation uses 16:9 slides, consistent margins, editable evidence charts and source notes. This is a presentation design choice because no mandatory public PPT template was found. It is not represented as an organizer-approved slide template.