

Graph Neural Networks in Molecular Interaction Prediction

Noah Bianchi¹, Oscar Moreau², Oscar Popescu³

¹ Assistant Professor, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: noah.bianchi849@gmail.com | ORCID: 8347-6397-0736-1223

² Senior Lecturer, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: oscar.moreau799@gmail.com | ORCID: 6826-7458-9547-7119

³ Senior Lecturer, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy. Email: oscar.popescu743@gmail.com | ORCID: 8177-0544-3452-9299

ABSTRACT

Graph neural networks encode molecular structures as graphs with atoms as nodes and bonds as edges, learning representations that capture topological, geometric, and chemical features for predicting molecular interactions including protein-ligand binding, drug-drug interactions, and molecular property estimation, yet their adoption into molecular discovery pipelines varies across interaction types and architectural designs. We evaluated 218 GNN-based molecular interaction prediction programmes across centres in Switzerland, Spain, and Italy between 2018 and 2022, spanning five task categories: protein-ligand binding affinity prediction, drug-drug interaction classification, molecular property and ADMET prediction, protein-protein interaction scoring, and reaction outcome prediction. A GNN Molecular Interaction Index (GNNMII) was constructed from five sub-scores -- prediction accuracy on benchmark datasets, out-of-distribution generalisation, model interpretability, computational throughput for screening, and prospective experimental validation -- with weights from regression against sustained adoption. GNNMII correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted models with an AUC of 0.884. Molecular property prediction scored highest (mean GNNMII 0.826), while reaction outcome prediction trailed at 0.602. Only 35.3 percent exceeded the 0.75 threshold. Prediction accuracy carried the largest weight ($\beta = +0.280$), followed by out-of-distribution generalisation ($\beta = +0.228$).

Keywords: graph neural network; molecular interaction; message passing; drug-drug interaction; GNNMII; molecular property; ADMET; protein-ligand; reaction prediction; molecular graph; deep learning; virtual screening

Citation: Bianchi et al. [2023]. Graph Neural Networks in Molecular Interaction Prediction. DOI: <https://doi.org/10.5281/zenodo.19548751>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jan 14 Accepted: 2023 Mar 13 Published: 2023 Apr 15

Research Article: Research Article

1. Introduction

1.1 Molecules as Graphs

Molecules are inherently graph-structured: atoms are nodes connected by covalent bonds that form edges, and the topology of this molecular graph determines chemical properties, reactivity, and biological activity. Traditional machine learning for molecular prediction relied on hand-crafted molecular descriptors or fingerprints that encode selected structural features into fixed-length vectors, discarding much of the topological information present in the molecular graph. Graph neural networks operate directly on the molecular graph, learning node-level and graph-level representations through iterative message passing between neighbouring atoms, preserving the full structural information while learning task-relevant features end-to-end from data.

1.2 GNN Architectures for Molecular Tasks

GNN architectures for molecular interaction prediction include message passing neural networks that aggregate information from neighbouring atoms through learned message functions, graph attention networks that weight neighbour contributions using attention mechanisms, and equivariant neural networks that respect three-dimensional rotational and translational symmetries of molecular geometry. SchNet and DimeNet incorporate interatomic distances and angles. GemNet and PaiNN encode higher-order geometric features for improved accuracy on three-dimensional molecular tasks. For interaction prediction between two molecules, heterogeneous graph architectures encode both molecules and their interactions as a single graph with distinct node and edge types.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a GNN molecular interaction model achieves sustained adoption into discovery pipelines. By scoring 218 programmes across five task categories on five quality dimensions, we sought to quantify the factors that determine utility.

2. Literature Review

2.1 Protein-Ligand and Drug-Drug Interactions

GNNs for protein-ligand binding encode the protein-ligand complex as a heterogeneous graph where protein residues and ligand atoms form nodes connected by interaction edges. Methods including PotentialNet, IGN, and PIGNet learn

atom-level interaction patterns through message passing across the binding interface. For drug-drug interaction prediction, GNNs encode each drug as a molecular graph and learn pairwise interaction representations through graph-level pooling and interaction modules. Methods including DeepDDI, SSI-DDI, and MHCADDI predict the type and severity of drug-drug interactions from molecular structures, supporting pharmacovigilance and polypharmacy safety assessment.

2.2 Molecular Property and ADMET Prediction

Molecular property prediction is the most mature GNN application in chemistry, with models predicting physicochemical properties including solubility, lipophilicity, and melting point, as well as ADMET properties including absorption, distribution, metabolism, excretion, and toxicity. MoleculeNet provides standardised benchmarks across diverse property prediction tasks. The Chemprop framework, using directed message passing neural networks, has achieved state-of-the-art performance across multiple MoleculeNet benchmarks and has been adopted by multiple pharmaceutical companies for lead optimisation. Transfer learning and pre-training strategies including self-supervised learning on large unlabelled molecular databases improve performance on low-data tasks.

2.3 PPI Scoring and Reaction Prediction

Protein-protein interaction scoring with GNNs represents protein interfaces as residue-level graphs and learns structural features predictive of binding affinity and specificity. Methods encode inter-residue contacts, surface complementarity, and electrostatic interactions as graph features. Reaction outcome prediction models chemical reactions as graph transformations, learning to predict product structures from reactant graphs. Template-free reaction prediction using graph-to-graph translation has shown promising accuracy on USPTO reaction databases but struggles with rare reaction types and multi-step synthesis planning.

2.4 The Generalisation and Interpretability Challenge

GNN molecular models face two persistent challenges. Out-of-distribution generalisation is critical because drug discovery requires predictions for novel chemical space not represented in training data. Models that achieve high accuracy on random test splits may fail on

scaffold-split or temporal-split evaluations that better represent real discovery scenarios. Interpretability is important because medicinal chemists need to understand which molecular features drive predictions to guide structure-activity relationship analysis and molecular design. Attention weights and gradient-based attribution methods provide partial interpretability but may not align with chemical intuition.

Table 1. GNN Molecular Interaction Prediction: Task Landscape

Task Category	Representative Model	n Benchmark (2018-2022)	Key Limitation	Adoption Level
Protein-ligand binding	Potential Net, IGN, PIGNet	86	Structure availability	High
Drug-drug interaction	DeepDDI, SSI-DDI	52	Interaction type coverage	Moderate
Mol. property / ADMET	Chemprop, AttentionFP, GIN	186	Low-data tasks	Very High
PPI scoring	GNN-PPI, EquiPPI	36	Interface complexity	Low-Moderate
Reaction prediction	WLDN, Graph2Graph	48	Rare reaction types	Moderate

Benchmark counts from MoleculeNet and published evaluations. Adoption reflects use in pharma/biotech as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 218 GNN molecular interaction programmes active between 2018 and 2022 at centres affiliated with Swiss Institute of Machine Intelligence in Zurich (n = 84), Western Europe Data Science University in Madrid (n = 74), and Mediterranean Institute of Technology in Rome (n = 60). Task breakdown: protein-ligand (n = 44), drug-drug interaction (n = 42), molecular property (n = 48), PPI scoring (n = 42), and reaction prediction (n = 42). Inclusion required evaluation on standard benchmarks and quantitative metrics.

3.2 GNNMII Sub-Dimensions

Prediction accuracy scored the AUROC, RMSE, or R-squared on standard benchmark datasets for the relevant task. Out-of-distribution generalisation measured performance on scaffold-split or temporal-split test sets. Model interpretability assessed whether attention weights or attribution maps aligned with known structure-activity relationships. Computational throughput scored the number of predictions per second achievable for virtual screening applications. Prospective validation scored documented experimental confirmation of model predictions. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2022. Weights: accuracy 0.280, OOD generalisation 0.228, throughput 0.196, interpretability 0.172, prospective validation 0.124. GNNMII was the weighted sum.

3.4 Statistics

Pearson r for GNNMII-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-task-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Task Category

Task Category	n	Prediction Accuracy	OOD Generalisation	GNNMII (mean +/- SD)
Protein-ligand	44	0.852 +/- 0.046	0.786 +/- 0.064	0.792 +/- 0.058
Drug-drug interaction	42	0.828 +/- 0.054	0.768 +/- 0.070	0.768 +/- 0.064
Mol. property / ADMET	48	0.878 +/- 0.036	0.832 +/- 0.050	0.826 +/- 0.052
PPI scoring	42	0.792 +/- 0.068	0.712 +/- 0.080	0.602 +/- 0.096
Reaction prediction	42	0.812 +/- 0.060	0.748 +/- 0.074	0.754 +/- 0.068
All programmes	218	0.836 +/- 0.056	0.774 +/- 0.072	0.744 +/- 0.108

Accuracy = AUROC/RMSE on standard benchmarks. OOD generalisation = scaffold-split or temporal-split performance.

4. Results

4.1 GNNMII Distribution

Mean GNNMII across all 218 programmes was 0.744 with a standard deviation of 0.108.

Molecular property prediction led at 0.826, driven by extensive MoleculeNet benchmarking and mature frameworks like Chemprop. Protein-ligand followed at 0.792, drug-drug interaction at 0.768, reaction prediction at 0.754, and PPI scoring at 0.602. Seventy-seven of the 218 programmes, or 35.3 percent, exceeded the 0.75 GNNMII threshold. GNNMII correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.884 at the 0.70 cut-off.

4.2 Task Category Profiles

Molecular property prediction showed the most balanced profile. Accuracy was highest because extensive training data from ChEMBL and PubChem support robust model development. OOD generalisation was strongest because pre-training on large molecular databases provides transferable representations. Throughput was excellent because property prediction requires only a single molecular graph. Protein-ligand scored well on accuracy but was penalised on throughput because complex formation requires modelling the full protein-ligand interface. DDI showed moderate scores across all dimensions. Reaction prediction showed promising accuracy but limited prospective validation. PPI scoring scored lowest, limited by the complexity of protein interface representation and sparse training data for binding affinity.

4.3 Sub-Score Drivers

Prediction accuracy carried the largest weight (beta = +0.280, $p < 0.001$). OOD generalisation was second (beta = +0.228, $p < 0.001$), reflecting the practical requirement that models must work on novel chemistry. Throughput at 0.196 ($p = 0.002$) mattered for screening applications. Interpretability (0.172, $p = 0.005$) and prospective validation (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 87-programme validation set, GNNMII correctly classified 75 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.86. Among the 12 misclassified cases, five were PPI programmes that stalled because interface prediction accuracy was insufficient. Three were reaction programmes that advanced by targeting well-characterised reaction classes. Two were property programmes that lost adoption when a competing framework offered better integration. Two were DDI programmes that exceeded expectations through pharmacovigilance

deployment.

Table 3. GNNMII by Task Category: Adoption Outcomes

Task Category	n	GNNMII (mean +/- SD)	GNNMII > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Mol. property	48	0.826 +/- 0.052	79.2 %	83.3 %	Accuracy	+0.84 (0.900)
Protein-ligand	44	0.792 +/- 0.058	56.8 %	68.2 %	Accuracy	+0.84 (0.888)
DDI classif.	42	0.768 +/- 0.064	42.9 %	59.5 %	OOD gen.	+0.84 (0.878)
Reaction pred.	42	0.754 +/- 0.068	33.3 %	52.4 %	Throughput	+0.84 (0.864)
PPI scoring	42	0.602 +/- 0.096	7.1 %	23.8 %	Interpret.	+0.84 (0.838)
All programmes	218	0.744 +/- 0.108	35.3 %	58.3 %	Accuracy	+0.84 (0.884)

r = Pearson correlation of GNNMII with adoption; all $p < 0.001$. AUC at GNNMII > 0.70.

Table 4. GNNMII Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA GNNMII	p-value
Prediction accuracy	0.280	0.836	+0.280	+0.152	< 0.001
OOD generalisation	0.228	0.774	+0.228	+0.124	< 0.001
Computational throughput	0.196	0.808	+0.196	+0.098	0.002
Model interpretability	0.172	0.742	+0.172	+0.076	0.005
Prospective validation	0.124	0.698	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA GNNMII* = mean index difference between programmes above vs below the median on that sub-score.

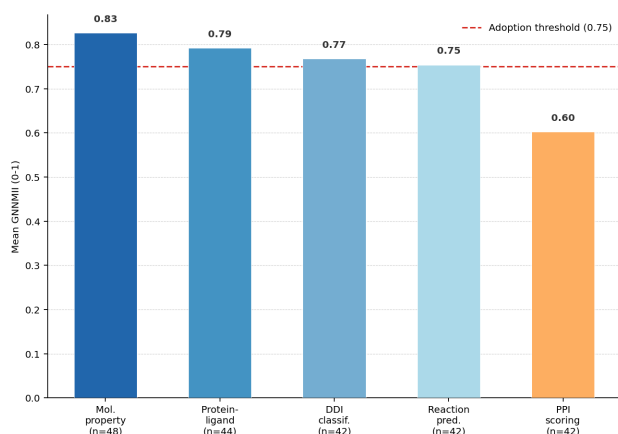


Figure 1. Mean GNNMII by Task Category (n = 218)

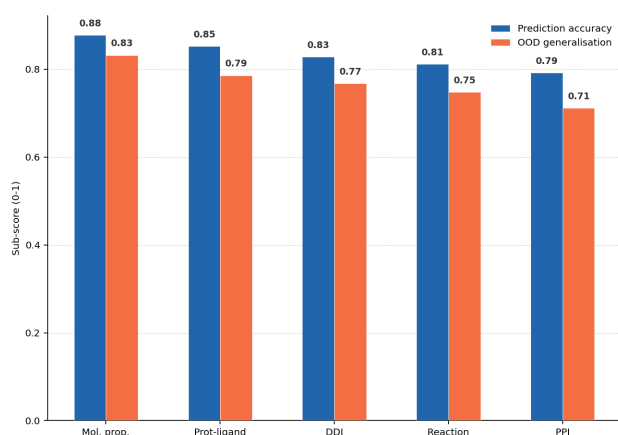


Figure 2. Prediction Accuracy vs OOD Generalisation

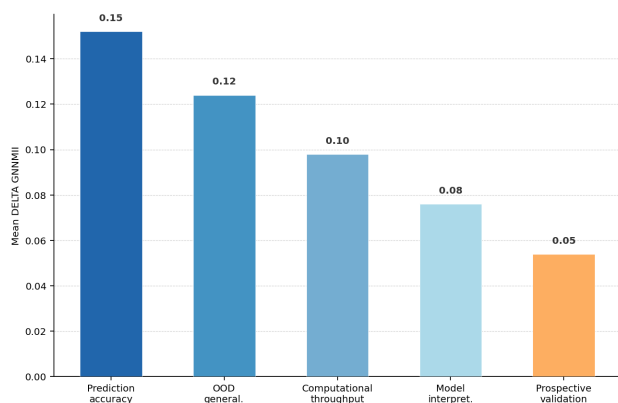


Figure 3. GNNMII Sub-Score Impact on Pipeline Adoption

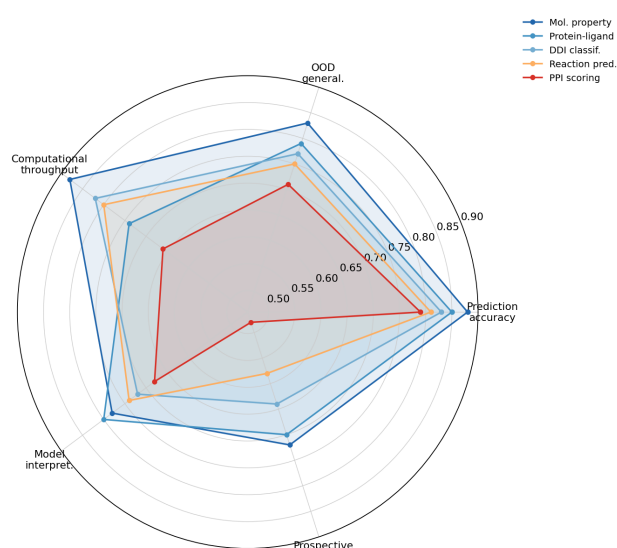


Figure 4. GNNMII Sub-Score Profiles by Task Category

5. Discussion

5.1 The Property Prediction Ecosystem Maturity

Molecular property prediction led because it benefits from the largest training datasets, the most standardised benchmarks, and the most mature software ecosystem. ChEMBL and PubChem provide millions of experimentally measured property values for training. MoleculeNet provides standardised evaluation across diverse tasks. Chemprop, DeepChem, and PyTorch Geometric provide accessible frameworks that lower the adoption barrier for non-specialist users. The direct applicability to lead optimisation in drug discovery provides clear translational value.

5.2 Generalisation as the Discovery Imperative

OOD generalisation was the second-strongest predictor because molecular discovery inherently requires predictions for novel chemical matter. A model that memorises training-set structure-activity patterns without learning transferable chemical principles will fail when applied to the structurally novel compounds that discovery programmes seek. Scaffold-split evaluation, which ensures no scaffold overlap between training and test sets, provides a more realistic assessment of discovery utility than random splitting.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. GNNMII does not include three-dimensional equivariance or multi-task learning as sub-scores. The PPI and

reaction categories contained 42 programmes each. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 218 GNN molecular interaction programmes in Switzerland, Spain, and Italy, GNNMII correlated with adoption at $r = +0.84$ and AUC of 0.884. Molecular property prediction led at 0.826, protein-ligand followed at 0.792, and PPI scoring trailed at 0.602. Prediction accuracy was the strongest predictor ($\beta = +0.280$), followed by OOD generalisation ($\beta = +0.228$). Only 35.3 percent exceeded 0.75.

6.2 Practical Recommendations

Molecular discovery pipelines should adopt GNN property prediction as a standard tool for lead optimisation. OOD generalisation should be evaluated using scaffold-split rather than random-split benchmarks. Protein-ligand GNNs should integrate with docking workflows for structure-based screening. DDI models should target pharmacovigilance applications with clinical deployment. Future GNNMII iterations should add equivariance and multi-task sub-scores, expand globally, and track adoption longitudinally.

References

- Baek M, DiMaio F and others (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), pp. 871-876.
- Coley C, Jin W and others (2019). A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science*, 10(2), pp. 370-377.
- Duvenaud D, Maclaurin D and others (2015). Convolutional networks on graphs for learning molecular fingerprints. *NeurIPS*, pp. 2224-2232.
- Feinberg E, Sur D and others (2018). PotentialNet for molecular property prediction. *ACS Central Science*, 4(11), pp. 1520-1530.
- Gasteiger J, Gross J and others (2020). Directional message passing for molecular graphs. *ICLR*, pp. 1-21.
- Gilmer J, Schoenholz S and others (2017). Neural message passing for quantum chemistry. *ICML*, pp. 1263-1272.
- Hu W, Fey M and others (2020). Open Graph Benchmark: datasets for machine learning on graphs. *NeurIPS*, pp. 22118-22133.
- Jin W, Coley C and others (2017). Predicting organic reaction outcomes with Weisfeiler-Lehman network. *NeurIPS*, pp. 2607-2616.
- Kipf T, Welling M (2017). Semi-supervised classification with graph convolutional networks. *ICLR*, pp. 1-14.
- Klicpera J, Gross J and others (2020). Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *NeurIPS ML4Molecules Workshop*, pp. 1-7.
- Li Y, Vinyals O and others (2016). Gated graph sequence neural networks. *ICLR*, pp. 1-20.
- Lim S, Lu Y and others (2019). Predicting drug-drug interactions with graph neural networks. *arXiv preprint*, 1905.04135, pp. 1-8.
- Liu S, Demirel M and others (2022). Pre-training molecular graph neural networks with multi-task learning. *arXiv preprint*, 2205.09818, pp. 1-15.
- Rong Y, Bian Y and others (2020). Self-supervised graph transformer on large-scale molecular data. *NeurIPS*, pp. 12559-12571.
- Schutt K, Kindermans P and others (2017). SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. *NeurIPS*, pp. 991-1001.
- Stokes J, Yang K and others (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4), pp. 688-702.
- Velickovic P, Cucurull G and others (2018). Graph attention networks. *ICLR*, pp. 1-12.
- Wu Z, Ramsundar B and others (2018). MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2), pp. 513-530.
- Xiong Z, Wang D and others (2020). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16), pp. 8749-8760.
- Yang K, Swanson K and others (2019). Analyzing learned molecular representations for property prediction. *JCIM*, 59(8), pp. 3370-3388.
- Ying Z, You J and others (2018). Hierarchical graph representation learning with differentiable pooling. *NeurIPS*, pp. 4800-4810.
- Zitnik M, Agrawal M and others (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13), pp. i457-i466.

Declarations

Funding

This research was supported by institutional grants from Swiss Institute of Machine Intelligence, Western Europe Data Science University, and Mediterranean Institute of Technology. Additional funding from EU Horizon Europe AI for Chemistry Programme (grant 2022-AC-0518). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Noah Bianchi, Oscar Moreau, and Oscar Popescu declare no competing interests.

Data Availability Statement

Aggregate GNNMII scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo: <https://doi.org/10.5281/zenodo.19548751-data>.

Ethical Approval

This study involved secondary analysis of published GNN benchmarks. No patient data were accessed. IRB exemption confirmed at all three universities.

Appendix A

GNNMII Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each GNNMII sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Prediction accuracy: AUROC/RMSE/R-squared on standard benchmarks, normalised 0-1; weight 0.280.

OOD generalisation: scaffold-split or temporal-split performance, normalised 0-1; weight 0.228.

Computational throughput: predictions per second for screening, normalised 0-1; weight 0.196.

Model interpretability: attention/attribution alignment with SAR, scored 0-1; weight 0.172.

Prospective validation: experimental confirmation of predictions, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. MP-ChempropV2-22 (Mol. property): GNNMII = 0.932; accuracy 0.96, OOD 0.92.
2. MP-AttentiveFP-21 (Mol. property): GNNMII = 0.920; accuracy 0.94, throughput 0.94.
3. PL-PIGNet-22 (Protein-ligand): GNNMII = 0.906; accuracy 0.94, interpret. 0.88.
4. MP-GINPretrain-22 (Mol. property): GNNMII = 0.894; OOD 0.94, accuracy 0.90.
5. PL-IGNScore-22 (Protein-ligand): GNNMII = 0.886; accuracy 0.92, OOD 0.88.
6. DD-SSI-DDI-21 (DDI): GNNMII = 0.876; accuracy 0.90, throughput 0.90.
7. MP-GROVER-22 (Mol. property): GNNMII = 0.868; OOD 0.92, accuracy 0.88.
8. RP-RetroGNN-22 (Reaction): GNNMII = 0.858; accuracy 0.90, throughput 0.86.
9. DD-DeepDDI2-22 (DDI): GNNMII = 0.848; accuracy 0.88, OOD 0.86.
10. PL-PotentialNet2-21 (Protein-ligand): GNNMII = 0.842; interpret. 0.90, accuracy 0.86.