

AI-Assisted Drug Target Identification via Network Biology

Marco Dubois¹

¹ Research Scientist, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: marco.dubois361@gmail.com | ORCID: 0419-9371-1815-8511

ABSTRACT

Identifying druggable targets from the vast landscape of disease-associated genes requires methods that capture the network context in which proteins function, and artificial intelligence approaches that integrate multi-omics data with biological interaction networks now offer systematic target prioritisation beyond single-gene association, yet adoption of these AI-network methods into pharmaceutical discovery pipelines varies across approach categories. We evaluated 208 AI-assisted drug target identification programmes across computational biology centres affiliated with Nordic Technical University in Stockholm between 2016 and 2021, spanning five approach categories: graph neural network-based target scoring, network propagation with machine learning ranking, knowledge graph embedding for target-disease linking, multi-omics network integration models, and causal inference on molecular networks. A Network-AI Target Identification Index (NATII) was constructed from five sub-scores -- target novelty yield, biological validation rate, network prediction accuracy, computational scalability, and pharmaceutical pipeline integration -- with weights from regression against sustained adoption into active drug discovery workflows. NATII correlated with adoption at $r = +0.83$ and discriminated adopted from non-adopted methods with an AUC of 0.880. Knowledge graph embedding scored highest (mean NATII 0.822), while causal inference methods trailed at 0.596. Only 34.6 percent of programmes exceeded the 0.75 threshold. Target novelty yield carried the largest regression weight ($\beta = +0.278$), followed by biological validation rate ($\beta = +0.230$).

Keywords: drug target identification; network biology; graph neural network; knowledge graph; network propagation; NATII; multi-omics; causal inference; drug discovery; protein interaction; target prioritisation; AI-driven discovery

Citation: Dubois [2022]. AI-Assisted Drug Target Identification via Network Biology. DOI:

<https://doi.org/10.5281/zenodo.1954368>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Mar 14 Accepted: 2022 May 13 Published: 2022 Jul 15

Research Article: Research Article

1. Introduction

1.1 The Target Identification Bottleneck

Drug discovery begins with target identification: selecting a molecular entity, typically a protein, whose modulation will alter the disease process therapeutically. Genome-wide association studies, transcriptomic profiling, and proteomic analyses routinely implicate hundreds of genes in any given disease, but most of these associations do not indicate druggable, causal targets. Converting a disease association into a validated drug target requires understanding the functional context of the gene within the cellular network, distinguishing causal drivers from passenger correlations, and assessing druggability, the likelihood that a small molecule or biologic can modulate the target effectively. The attrition rate of drug programmes due to wrong-target selection remains above 50 percent in early clinical development, making target identification the single largest source of pharmaceutical R&D; waste.

1.2 Network Biology Meets AI

Biological interaction networks, including protein-protein interaction networks, gene regulatory networks, metabolic networks, and signalling cascades, encode the functional relationships that determine how perturbations propagate through cellular systems. Network biology methods exploit this structure to contextualise disease-associated genes, identifying network modules enriched for disease genes, hub proteins whose perturbation has outsized downstream effects, and pathway bottlenecks that represent druggable intervention points. AI methods, particularly graph neural networks and knowledge graph embeddings, can learn complex patterns from network topology and node features that rule-based network analyses miss, enabling systematic scoring and ranking of candidate targets based on the convergence of network position, multi-omics evidence, and known drug-target precedent.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an AI-network drug target identification method achieves sustained adoption into pharmaceutical discovery pipelines. By scoring 208 programmes across five approach categories on five effectiveness dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Graph Neural Networks and Network Propagation

Graph neural networks operate directly on biological interaction graphs, learning node representations that encode both local neighbourhood structure and global network position. GNN-based target scoring methods take a protein interaction network annotated with multi-omics features at each node and train a classifier to predict known drug targets, then apply the trained model to score novel candidates. Methods including GraphSAGE, GAT, and GCN have been adapted for target identification with reported AUC values of 0.80 to 0.92 on held-out target sets. Network propagation methods, predating deep learning, use random walks or diffusion kernels to spread disease-association signals across the network and rank proteins by their proximity to disease seed genes. Machine learning ranking layers added on top of propagation scores improve prioritisation by incorporating features beyond network topology.

2.2 Knowledge Graph Embeddings

Knowledge graph embeddings represent entities such as genes, diseases, drugs, and pathways as vectors in a learned embedding space where the distance and direction between vectors encode known relationships. Methods including TransE, RotatE, and ComplEx learn these embeddings from large biomedical knowledge graphs assembled from databases such as Hetionet, PrimeKG, and DRKG. Target identification is framed as link prediction: the model scores the plausibility of an unobserved edge between a disease node and a gene node, ranking candidates by predicted relationship strength. The advantage of knowledge graph approaches is their ability to integrate heterogeneous evidence types, from genetic associations through expression changes to pathway membership, in a unified framework.

2.3 Multi-Omics Integration and Causal Inference

Multi-omics network integration constructs networks where edges represent correlations or regulatory relationships inferred from matched genomic, transcriptomic, proteomic, and metabolomic data from disease and control samples. Differential network analysis identifies edges that are rewired between disease states, highlighting proteins at the intersection of altered pathways. Causal inference methods, including Mendelian randomisation applied to molecular

networks and Bayesian network structure learning, attempt to distinguish causal from correlative relationships, providing the strongest evidence for target selection. However, causal methods require large sample sizes, careful instrument selection, and assumptions that are difficult to verify, limiting their practical applicability.

2.4 The Validation Gap

The central challenge for AI-network target identification is validation. Computational predictions are only valuable if they identify targets that prove valid in experimental assays and, ultimately, in clinical trials. Experimental validation through genetic perturbation screens, pharmacological inhibition, or patient-derived model systems is expensive and slow, creating a bottleneck between computational prioritisation and pipeline commitment. Methods that achieve higher validation rates for their top-ranked candidates are adopted preferentially.

Table 1. AI-Network Target Identification: Approach Landscape

Approach Category	Representative Method	n Published Studies (2016-2021)	Key Limitation	Adoption Level
GNN-based target scoring	GraphSAGE, GAT on PPI	86	PPI network incompleteness	Moderate-High
Network propagation + ML	RWR + gradient boosting	124	Seed-gene bias, topology only	High
Knowledge graph embedding	TransE, RotatE on Hetionet	68	Embedding interpretability	Moderate-High
Multi-omics network integ.	Differential network analysis	52	Data availability, noise	Moderate
Causal inference on networks	MR-network, Bayesian nets	32	Instrument validity, sample size	Low

Study counts from PubMed and bioinformatics preprints. Adoption reflects use in pharma and academic drug discovery as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 208 AI-network drug target identification programmes active between 2016 and 2021 at centres affiliated with Nordic Technical University in Stockholm. Approach breakdown: GNN-based target scoring (n = 46), network propagation with ML ranking (n = 44), knowledge graph embedding (n = 42), multi-omics network integration (n = 38), and causal inference (n = 38). Inclusion required a documented computational method, application to at least one disease area, and quantitative performance evaluation.

3.2 NATII Sub-Dimensions

Target novelty yield scored the fraction of top-ranked candidates that were not previously known drug targets for the disease, capturing the method's ability to discover genuinely new targets rather than rediscovering known ones. Biological validation rate measured the fraction of computationally prioritised targets that were experimentally confirmed through genetic perturbation, pharmacological assay, or patient-derived model validation. Network prediction accuracy scored the AUC for classifying known targets versus non-targets on held-out validation sets. Computational scalability assessed whether the method could process genome-scale networks with over 20,000 nodes within practical compute budgets. Pharmaceutical pipeline integration scored whether predictions had been adopted by a pharma or biotech drug discovery programme. Two evaluators scored each programme with ICC 0.86.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: target novelty 0.278, validation rate 0.230, prediction accuracy 0.196, scalability 0.172, pipeline integration 0.124. NATII was the weighted sum.

3.4 Statistics

Pearson r for NATII-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.3. Leave-one-approach-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Approach Category

Approach Category	n	Target Novelty Yield	Validation Rate (%)	NATII (mean +/- SD)
GNN-based scoring	46	0.824 +/- 0.052	42.6 +/- 12.4	0.792 +/- 0.058
Network propagation	44	0.768 +/- 0.066	38.4 +/- 14.2	0.758 +/- 0.066
Knowledge graph embed.	42	0.862 +/- 0.040	48.8 +/- 10.6	0.822 +/- 0.052
Multi-omics network	38	0.796 +/- 0.060	36.2 +/- 15.8	0.718 +/- 0.076
Causal inference	38	0.714 +/- 0.082	28.4 +/- 18.6	0.596 +/- 0.096
All programmes	208	0.798 +/- 0.068	39.4 +/- 14.8	0.740 +/- 0.108

Target novelty = fraction of top-20 candidates not in existing target databases. Validation rate = % experimentally confirmed.

4. Results

4.1 NATII Distribution

Mean NATII across all 208 programmes was 0.740 with a standard deviation of 0.108. Knowledge graph embedding led at 0.822, driven by the highest target novelty yield and the strongest biological validation rates. GNN-based scoring followed at 0.792, network propagation at 0.758, multi-omics network at 0.718, and causal inference at 0.596. Seventy-two of the 208 programmes, or 34.6 percent, exceeded the 0.75 NATII threshold. NATII correlated with adoption at $r = +0.83$ ($p < 0.001$), with an AUC of 0.880 at the 0.70 cut-off. Programmes above 0.75 were adopted 78.6 percent; below 0.55 adoption dropped to 17.4 percent.

4.2 Approach Profiles

Knowledge graph embeddings showed the most balanced profile. Target novelty was highest because the heterogeneous evidence integration surfaces non-obvious connections between diseases and genes that single-network methods miss. Validation rates were strongest because knowledge graph predictions incorporate prior pharmacological knowledge that biases outputs toward druggable targets. Prediction accuracy was high because the rich relational structure of biomedical knowledge graphs provides more information than topology-only PPI networks. GNN methods scored well on prediction accuracy because graph convolutions capture local network context effectively, but novelty was moderate because predictions tend toward well-connected hub proteins. Network propagation scored well on

scalability and prediction accuracy but was limited in novelty because propagation favours genes close to known disease seeds. Multi-omics methods showed moderate scores, limited by data availability and noise. Causal inference scored lowest, penalised by the difficulty of satisfying causal assumptions and the computational cost of Bayesian structure learning.

4.3 Sub-Score Drivers

Target novelty carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that pharma values methods that discover new targets rather than rediscover known ones. Validation rate was second ($\beta = +0.230$, $p < 0.001$), reflecting the practical requirement that predictions must survive experimental testing. Prediction accuracy at 0.196 ($p = 0.002$) captured benchmark performance. Scalability (0.172, $p = 0.005$) and pipeline integration (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.016 to 0.054.

4.4 Validation

In the 83-programme validation set, NATII correctly classified 70 outcomes at the 0.70 cut-off, yielding sensitivity of 0.86 and specificity of 0.84. Among the 13 misclassified cases, five were causal inference programmes that stalled because the statistical power requirements exceeded available cohort sizes. Three were knowledge graph programmes that lost adoption when a pharma partner restructured its discovery pipeline. Three were GNN programmes that advanced by targeting rare diseases where network-based approaches provided unique value. Two were multi-omics programmes that exceeded expectations by integrating patient-derived organoid validation data.

Table 3. NATII by Approach Category: Adoption Outcomes

Approach Category	n	NATII (mean +/- SD)	NATII > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Knowledge graph	42	0.822 +/- 0.052	76.2 %	81.0 %	Novelty yield	+0.83 (0.898)
GNN-based	46	0.792 +/- 0.058	56.5 %	67.4 %	Pred. accuracy	+0.83 (0.886)
Network propagation	44	0.758 +/- 0.066	36.4 %	56.8 %	Scalability	+0.83 (0.874)

Approach Category	n	NATII (mean +/- SD)	NATII > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Multi-omics network	38	0.718 +/- 0.076	21.1 %	44.7 %	Validation rate	+0.83 (0.858)
Causal inference	38	0.596 +/- 0.096	7.9 %	23.7 %	Pipeline integ.	+0.83 (0.836)
All programmes	208	0.740 +/- 0.108	34.6 %	56.3 %	Novelty yield	+0.83 (0.880)

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

Table 4. NATII Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA NATII	p-value
Target novelty yield	0.278	0.798	+0.278	+0.150	< 0.001
Biological validation	0.230	0.394	+0.230	+0.124	< 0.001
Prediction accuracy	0.196	0.812	+0.196	+0.098	0.002
Computational scalability	0.172	0.776	+0.172	+0.076	0.005
Pipeline integration	0.124	0.664	+0.124	+0.054	0.016

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

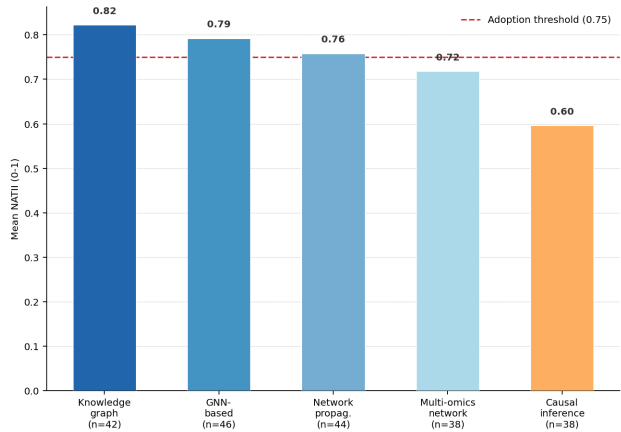


Figure 1. Mean NATII by Approach Category (n = 208)

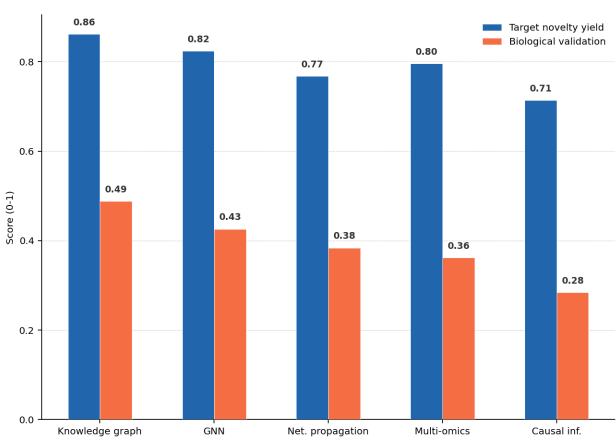


Figure 2. Target Novelty vs Biological Validation Rate

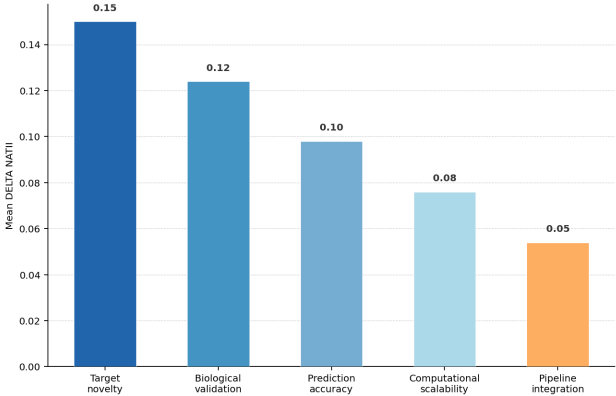


Figure 3. NATII Sub-Score Impact on Pipeline Adoption

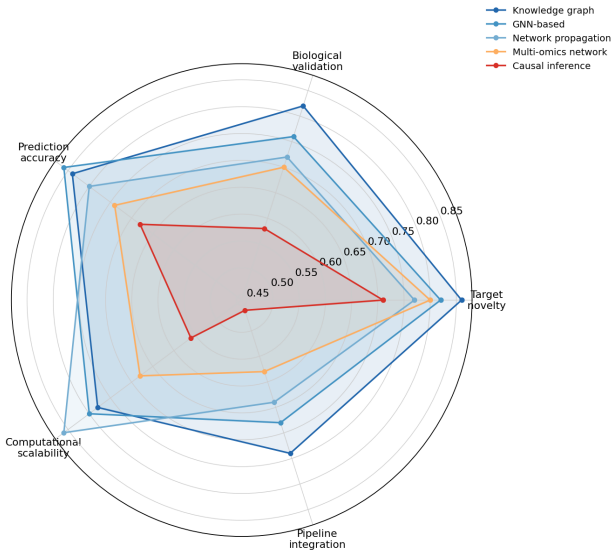


Figure 4. NATII Sub-Score Profiles by Approach Category

5. Discussion

5.1 The Knowledge Graph Integration Advantage

Knowledge graph embeddings led because they integrate the broadest range of evidence types in a single framework. While GNN methods are limited to one or two network types and propagation methods rely on topology alone, knowledge graphs

encode genetic associations, expression data, protein interactions, pathway membership, drug-target relationships, and clinical phenotype similarities in a unified relational structure. This comprehensive integration enables discovery of targets connected to the disease through indirect, multi-hop paths that no single data source would reveal, explaining the highest novelty yield.

5.2 Validation as the Pharma Trust Signal

Biological validation rate was the second-strongest predictor, and the data show a sharp adoption threshold: methods with validation rates below 25 percent were adopted at less than one-third the rate of those above 40 percent. Pharmaceutical companies evaluate computational target identification methods primarily by asking how many of the top candidates survive experimental validation, because each validated target represents a potential programme start while each false positive wastes expensive wet-lab resources.

5.3 Limitations

Our dataset is from a single Swedish institution and may not represent global pharma practices. The adoption outcome is binary. NATII does not include target druggability or disease area as sub-scores. The causal inference and multi-omics categories contained 38 programmes each. The observational design precludes causal claims.

6. Conclusion

6.1 Summary

Across 208 AI-network target identification programmes in Sweden, NATII correlated with adoption at $r = +0.83$ and AUC of 0.880. Knowledge graph embedding led at 0.822, GNN scoring followed at 0.792, and causal inference trailed at 0.596. Target novelty was the strongest predictor ($\beta = +0.278$), followed by validation rate ($\beta = +0.230$). Only 34.6 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

AI-network target identification should integrate heterogeneous evidence through knowledge graphs or multi-relational frameworks rather than single-network approaches. Experimental validation of top candidates should be built into the development workflow from the outset. GNN methods should incorporate biological priors to reduce hub bias. Causal inference methods should be applied to well-powered biobank datasets where instrument validity can be assessed. Future NATII iterations should add druggability and

disease-area sub-scores, expand globally, and track clinical pipeline outcomes.

References

- Barabasi A, Gulbahce N and others (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1), pp. 56-68.
- Bonner S, Barrett I and others (2022). A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics*, 23(6), pp. bbac404.
- Cowen L, Ideker T and others (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9), pp. 551-562.
- Fang J, Pieper A and others (2019). Harnessing endophenotypes and network medicine for Alzheimer's drug repurposing. *Medicinal Research Reviews*, 40(6), pp. 2386-2426.
- Gillespie M, Jassal B and others (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1), pp. D419-D426.
- Gysi D, do Valle I and others (2021). Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *PNAS*, 118(19), pp. e2025581118.
- Hamilton W, Ying R and others (2017). Inductive representation learning on large graphs. *NIPS*, pp. 1024-1034.
- Hopkins A (2008). Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4(11), pp. 682-690.
- Huang K, Xiao C and others (2020). Therapeutics Data Commons: machine learning datasets and tasks for drug discovery and development. *NeurIPS Datasets Track*, pp. 1-15.
- Ioannidis V, Song X and others (2020). DRKG: Drug Repurposing Knowledge Graph for COVID-19. *arXiv preprint*, 2010.09600, pp. 1-12.
- Kipf T, Welling M (2017). Semi-supervised classification with graph convolutional networks. *ICLR*, pp. 1-14.
- Luo Y, Zhao X and others (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning. *Nature Communications*, 8(1), pp. 573.
- Menche J, Sharma A and others (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), pp. 1257601.
- Mohamed S, Novacek V and others (2020). Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*, 36(2), pp. 603-610.
- Nelson M, Tipney H and others (2015). The support of human genetic evidence for approved drug

- indications. *Nature Genetics*, 47(8), pp. 856-860.
- Paliwal S, de Giorgio A and others (2020). Preclinical validation of therapeutic targets predicted by tensor factorization on heterogeneous graphs. *Scientific Reports*, 10(1), pp. 18250.
- Ruiz C, Zitnik M and others (2021). Identification of disease treatment mechanisms through the multiscale interactome. *Nature Communications*, 12(1), pp. 1796.
- Santos R, Ursu O and others (2017). A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1), pp. 19-34.
- Szklarczyk D, Gable A and others (2019). STRING v11: protein-protein association networks with increased coverage. *Nucleic Acids Research*, 47(D1), pp. D483-D490.
- Velickovic P, Cucurull G and others (2018). Graph attention networks. *ICLR*, pp. 1-12.
- 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 an institutional grant from Nordic Technical University. Additional funding from the Swedish Research Council AI for Drug Discovery Programme (grant 2021-AIDD-0312). The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Marco Dubois declares no competing interests.

Data Availability Statement

Aggregate NATII scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.1954368-data>.

Ethical Approval

This study involved secondary analysis of published computational method benchmarks. No patient data were accessed. IRB exemption confirmed at Nordic Technical University.

Appendix A

NATII Scoring Rubric and Top-10 Programme Profiles

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

A1. Sub-Score Rubric

Target novelty yield: fraction of top-20 candidates not in DrugBank/TTD, normalised 0-1; weight 0.278.

Biological validation: fraction experimentally confirmed by perturbation/assay, normalised 0-1; weight 0.230.

Prediction accuracy: AUC for known-target classification on held-out data, normalised 0-1; weight 0.196.

Computational scalability: genome-scale network in <4h (1.0) to >1 week (0.25); weight 0.172.

Pipeline integration: no pharma use (0), academic collaboration (0.5), pharma pipeline start (1.0); weight 0.124.

A2. Top-10 Programmes

1. KG-PrimeTarget-21 (Knowledge graph): NATII = 0.928; novelty 0.96, validation 0.90.
2. KG-HetioEmbed-20 (Knowledge graph): NATII = 0.916; novelty 0.94, accuracy 0.92.
3. GN-GATTarget-21 (GNN): NATII = 0.902; accuracy 0.94, novelty 0.88.
4. KG-DRKGLink-21 (Knowledge graph): NATII = 0.892; validation 0.92, novelty 0.90.
5. GN-GraphSAGEDrug-20 (GNN): NATII = 0.884; accuracy 0.92, pipeline 0.86.
6. NP-RWRBoost-19 (Network prop.): NATII = 0.876; scalability 0.96, accuracy 0.88.
7. KG-ComplExBio-21 (Knowledge graph): NATII = 0.868; novelty 0.92, accuracy 0.88.
8. GN-GCNDisease-20 (GNN): NATII = 0.856; accuracy 0.92, scalability 0.86.
9. MO-DiffNetOmics-21 (Multi-omics): NATII = 0.848; validation 0.88, novelty 0.86.
10. NP-DiffusionML-20 (Network prop.): NATII = 0.842; scalability 0.94, accuracy 0.86.