Graph Neural Network-Based Prediction of Drug-Target Interactions Using Heterogeneous Biomedical Data Integration

Radoslaw Drywa
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
radoslaw.drywa240@ucf.edu

Abstract

Drug-target interaction (DTI) prediction must reconcile sparse, noisy, and heterogeneous evidence ranging from chemical structure and protein sequence to pathway membership, expression context, and literature-derived associations. Many existing computational models treat DTI prediction as homogeneous link prediction or concatenate features before learning, which obscures relation-specific semantics and can propagate degree and assay biases. This manuscript develops a heterogeneous graph neural network framework for DTI prediction that represents drugs, proteins, diseases, pathways, and related biomedical entities as typed nodes and encodes known interactions and contextual relations as typed edges. Type-specific encoders map molecular fingerprints, sequence representations, and omics or knowledge-base attributes into a shared latent space; relation-aware attention aggregates evidence across heterogeneous neighborhoods; and a calibrated scoring function estimates interaction probabilities with explicit handling of positive-unlabeled bias, edge confidence, and distribution shift. The framework specifies leakage-resistant evaluation, cold-start diagnostics, uncertainty calibration, and computational strategies for sparse large-scale graphs. Analytical assessment indicates that the main gains arise not from adding data indiscriminately but from preserving relation semantics, controlling noisy or indirect edges, and separating biological mechanism from statistical association. The resulting design offers a technically grounded route to more transferable and interpretable DTI prediction and clarifies the conditions under which heterogeneous biomedical integration is likely to improve prioritization rather than merely increase model complexity.

Keywords

drug-target interaction; heterogeneous graph neural networks; biomedical data integration; positive-unlabeled learning; graph attention; uncertainty calibration

1. Introduction

Drug-target interaction prediction is an evidence-integration problem rather than a simple binary classification task. A candidate drug-protein pair may be supported by direct binding assays, inferred from chemical similarity, suggested by pathway membership, or associated through disease phenotypes, yet these forms of evidence carry different reliability, granularity, and causal proximity to the interaction itself. Sequence-based deep models have shown that learned representations of compounds and proteins can capture binding-relevant patterns beyond hand-engineered descriptors [1]. Network-based integration has further demonstrated that heterogeneous biomedical relations can support drug repositioning and interaction inference when multiple entity types are considered jointly [2]. Graph convolutional models have also been used to model polypharmacy and side-effect relations over biomedical graphs

[3], establishing that message passing over relational structure can exploit dependencies that feature concatenation misses.

The central methodological difficulty is that a DTI graph is not merely large and sparse; it is semantically typed. A drug-drug similarity edge, a protein-protein interaction edge, a pathway-membership edge, and a known drug-target binding edge are not interchangeable channels of information. Generic graph convolutional networks [4] and attention-based graph neural networks [5] provide useful building blocks, but they must be adapted to typed relations and edge-specific reliability. Relational graph convolutional models [6] and knowledge-aware attention mechanisms [7] show how relation-specific transformations can preserve edge semantics. Heterogeneous graph attention models [8] further indicate that metapath-aware or type-aware aggregation can improve representation learning when multiple node and edge types coexist. These developments motivate a DTI framework in which relation type, node type, edge confidence, and biological context are represented explicitly rather than flattened into a single adjacency matrix.

A second difficulty concerns supervision. Confirmed interactions are positive labels, but unobserved pairs are not verified negatives. Positive-unlabeled learning addresses this setting by treating unlabeled instances as a mixture of hidden positives and negatives [10]. In biomedical databases, missing interactions may reflect untested pairs, assay failures, weak binding below a threshold, or curation lag rather than biological absence. At the same time, non-graph classifiers such as gradient-boosted trees [9] remain important baselines because they are strong on tabular descriptors and resistant to some forms of overfitting. A credible graph-based method must therefore demonstrate value beyond strong feature-based baselines and must avoid confusing graph coverage with biological evidence.

The contribution of this manuscript is a heterogeneous graph neural network framework for DTI prediction that integrates biomedical data through typed encoders, relation-aware message passing, edge-confidence weighting, positive-unlabeled training, and calibrated uncertainty estimation. The framework is not presented as an empirical benchmark result; rather, it is developed as a technically specified and analytically assessed methodology. The design emphasizes three requirements: preservation of relation semantics during information fusion, explicit treatment of unreliable or indirect evidence, and evaluation protocols that separate genuine generalization from leakage and degree-driven shortcuts. These requirements lead to concrete architectural choices, objective functions, calibration procedures, and failure-mode analyses that can guide implementation and subsequent empirical study.

2. Problem Formulation and Biomedical Data Heterogeneity

2.1 Drug-target interaction as typed relation prediction

Let a biomedical knowledge graph be represented as a heterogeneous graph with typed entities and typed relations. The prediction target is whether a drug node d and a protein node t are linked by a validated interaction relation. The model may exploit direct drug-target edges, drug-drug similarity, protein-protein interactions, drug-disease associations, disease-gene associations, pathway membership, side-effect relations, expression context, and literature-derived relations. The graph is formally written as follows.

$$G = (V, E, T(V), T(E), X)$$

Here V is the set of entity nodes, E is the set of typed edges, T(V) is the set of node types, T(E) is the set of relation types, and X contains node features. Each edge e = (u, v, r) connects a source node u to a target node v through relation type r. An edge may also carry a confidence weight and auxiliary attributes, such as assay type, binding affinity, tissue context, or curation provenance. The DTI prediction problem is to estimate a scoring function for candidate drug-target pairs that combines graph structure, node attributes, and relation-specific evidence.

This formulation differs from homogeneous link prediction because the meaning of a path depends on the sequence of relation types. A path from drug to protein through a disease-gene relation is not equivalent to a path through direct binding evidence. The model must therefore learn relation-conditioned transformations and should not aggregate all neighbors with a single shared weight matrix.

2.2 Heterogeneous biomedical entities and relations

The node types typically include drugs, proteins, diseases, pathways, side effects, and expression contexts. Drug nodes may be represented by molecular fingerprints, learned molecular graph embeddings, physicochemical descriptors, or target-profile features. Protein nodes may be represented by amino-acid sequence embeddings, domain annotations, structural descriptors, or expression profiles. Disease and pathway nodes may use ontology embeddings, gene-set memberships, or text-derived representations. The graph may also include auxiliary nodes for assays, publications, cell lines, or anatomical contexts when such data are available and sufficiently curated.

The edge types are equally important. Known drug-target interactions should be separated by evidence strength, for example direct binding, functional assay, or inferred pathway relation. Drug-drug similarity can be derived from chemical structure or shared targets, but similarity edges may duplicate target information and induce leakage if the evaluation split is not controlled. Protein-protein interaction edges can propagate functional context, but high-degree hub proteins can dominate message passing and obscure specific binding evidence. Disease-gene and drug-disease edges can support repositioning hypotheses, but they are indirect with respect to molecular binding. These distinctions motivate relation-specific transformations and edge-confidence weights.

2.3 Label reliability and positive-unlabeled structure

A confirmed DTI is a positive observation, but the absence of a recorded interaction is not a verified negative. The unlabeled set contains both untested pairs and true non-interactions, and the proportion between them varies by protein family, assay technology, and database curation policy. Treating all unobserved pairs as negatives creates a training signal that rewards database coverage rather than biological discrimination. Positive-unlabeled learning provides a principled starting point by modeling the unlabeled distribution as a mixture of positives and negatives [10]. In a graph setting, this issue is compounded because unlabeled pairs may be connected through short paths that share drug or target neighbors with labeled positives, increasing the risk of optimistic evaluation if splitting is performed only at the edge level.

The framework therefore treats labels as evidence-weighted observations rather than as a complete binary partition. Edge weights can encode assay confidence, binding strength, or curation quality, while the loss can down-weight uncertain positives and use positive-

unlabeled corrections for the unlabeled set. This does not eliminate label noise, but it aligns training with the structure of biomedical evidence.

3. Related Work and Design Rationale

Sequence-based DTI models established that learned compound and protein representations can improve affinity or interaction prediction relative to descriptor-based pipelines [1]. Their limitation is that they primarily model pairwise molecular information and do not naturally represent multi-hop biomedical context. Network-integration approaches address this limitation by combining drug, target, disease, and pathway relations [2], but early integration methods often relied on hand-designed similarity kernels or matrix factorization rather than end-to-end relation-aware learning. Graph neural networks provide a natural bridge between these traditions because they can propagate information over explicit relational structure while learning task-specific representations.

Graph convolutional networks [4] and graph attention networks [5] are the basic architectural precedents for neighborhood aggregation and adaptive weighting. Relational graph convolutional networks [6] extend this idea to multiple edge types by assigning relation-specific transformations, which is directly relevant to DTI graphs containing binding, similarity, pathway, and disease relations. Knowledge-aware attention models [7] demonstrate that attention over heterogeneous relations can improve recommendation and link prediction when entities are connected through varied semantic paths. Heterogeneous graph attention networks [8] further show that type-aware aggregation and metapath reasoning can outperform homogeneous aggregation in graphs with multiple node and relation types. These works support the design choice to separate relation types rather than collapse them into a single weighted adjacency matrix.

Strong non-graph baselines remain essential. Gradient-boosted decision trees [9] are effective on tabular molecular descriptors and can reveal whether a graph model is adding information beyond conventional features. Positive-unlabeled learning [10] is also central because DTI datasets rarely contain reliable negatives. The proposed framework combines these insights: it uses graph structure for contextual propagation, typed encoders for heterogeneous entity features, relation-aware attention for semantic discrimination, and positive-unlabeled weighting for realistic supervision.

4. Methodological Framework

4.1 Typed graph construction

The graph construction stage separates entity types, relation types, and evidence sources. Drug and protein nodes are mandatory; disease, pathway, side-effect, and context nodes are optional and should be included only when their provenance and coverage are sufficient. Each edge receives a relation label and a confidence weight. Direct drug-target binding edges receive higher confidence than inferred or text-mined edges. Similarity edges are retained only when they are not simple transformations of the prediction target, because otherwise the model may learn a shortcut that inflates validation performance. Temporal metadata, when available, should be preserved so that evaluation can simulate prospective prediction rather than random edge masking.

The typed graph is not assumed to be complete. Missing relations are represented as absent edges rather than as negative labels. The model must therefore learn from partially observed neighborhoods and should include a mechanism for handling low-degree nodes.

Node features are normalized within type, and edge weights are clipped to a bounded range to prevent a small number of high-confidence or high-degree edges from dominating aggregation.

4.2 Type-specific encoders

Each node type has a different feature space. Drugs may use molecular fingerprints or learned molecular graph representations; proteins may use sequence embeddings or domain annotations; diseases and pathways may use ontology or gene-set embeddings. A shared encoder across all node types would impose an inappropriate inductive bias because the statistical structure of chemical fingerprints differs fundamentally from that of protein sequences or disease ontologies. The framework therefore uses type-specific encoders followed by a projection into a common latent space. The initial representation is defined as follows.

$$h(v,0) = f\big(\tau(v)\big)\Big(x(v); \theta\big(\tau(v)\big)\Big)$$

In this expression, h(v,0) is the initial hidden state of node v, τ(v) is the node type, x(v) is the raw feature vector, f(τ(v)) is the encoder for that type, and θ(τ(v)) are its parameters. The encoder may be a multilayer perceptron for tabular features, a convolutional or attention-based sequence encoder for proteins, or a graph encoder for molecular graphs. The projection into a shared space allows message passing across types while preserving type-specific feature extraction.

4.3 Relation-aware attention and aggregation

At each message-passing layer, a node receives information from neighbors connected by different relation types. The framework computes a relation-specific attention score for each neighbor and normalizes scores within the same relation neighborhood. The normalization is given by the following expression.

$$\alpha(v,u,r) \;=\; exp(e(v,u,r)) \;/\; \Sigma\ over\ w\ in\ N(r,v)\ exp(e(v,w,r))$$

Here α(v,u,r) is the attention weight from neighbor u to node v under relation r, N(r,v) is the set of neighbors of v connected by relation r, and e(v,u,r) is a relation-specific compatibility score. The score is computed by applying a LeakyReLU activation to a concatenation of relation-specific projections of the two node states, followed by a learnable attention vector for that relation. This design allows the model to distinguish, for example, a direct binding neighbor from a pathway neighbor even when both connect to the same protein.

The normalized attention weights are used to aggregate messages across relations. The update rule is written as follows.

$$h(v,l+1) \;=\; \sigma(\Sigma\ over\ r\ \Sigma\ over\ u\ in\ N(r,v)\ \alpha(v,u,r)\ W(r)\ h(u,l))$$

In this equation, h(v,l+1) is the updated representation of node v at layer l+1, σ is a nonlinear activation, W(r) is a relation-specific transformation matrix, and h(u,l) is the representation of neighbor u at layer l. The sum over r allows the model to integrate evidence from multiple relation types, while W(r) prevents relation-specific signals from being mixed prematurely. Multiple attention heads can be used, with their outputs concatenated or averaged. Residual connections and layer normalization stabilize training in deep stacks.

Edge confidence can be incorporated by multiplying each message by a bounded weight before aggregation. This is important because a high-confidence binding edge and a low-confidence text-mined edge should not contribute equally. The confidence weight is not a

substitute for label correction; it is a soft prior that reduces the influence of noisy relations while preserving their potential utility.

4.4 Fusion and DTI scoring

After L message-passing layers, each drug and protein receives a contextual representation that encodes its local graph neighborhood. The final DTI score combines these representations through a bilinear or multilayer scoring function. A bilinear form with a bias term is written as follows.

$$s(d,t) = \sigma\big(z(d)^T M z(t) + b(d,t)\big)$$

Here s(d,t) is the predicted interaction probability for drug d and protein t, z(d) and z(t) are the final contextual embeddings, M is a learned interaction matrix, b(d,t) is an optional pair-specific bias derived from features such as assay type or tissue context, and σ is the logistic sigmoid. The bilinear term captures compatibility between drug and target representations, while the bias term allows contextual features to adjust the baseline propensity for interaction. A multilayer perceptron can replace the bilinear form when nonlinear pair interactions are expected, but the bilinear version provides a useful inductive bias toward low-rank drug-target compatibility.

The scoring function should be trained with a calibration-aware objective and evaluated on pairs that are not trivially recoverable from shared neighbors. If the model is used for ranking, the raw score can be used; if it is used for decision support, the calibrated probability is more appropriate.

4.5 Objective and positive-unlabeled weighting

The training objective must account for positive-unlabeled structure and class imbalance. A weighted binary cross-entropy loss with regularization is defined as follows.

$$L = -\Sigma\, w(d,t)[y \log s(d,t) + (1-y)\log(1-s(d,t))] + \lambda\|\Theta\|^2$$

In this expression, the sum runs over labeled and sampled pairs, y is the observed label, w(d,t) is a pair-specific weight, λ is the regularization strength, and Θ denotes the model parameters. The weight can reflect assay confidence, class balancing, or positive-unlabeled correction. For unlabeled pairs, the effective negative weight should be reduced or estimated rather than set to one, because some unlabeled pairs are hidden positives. The regularization term controls model complexity and is especially important when heterogeneous features create many weak associations.

The objective can be extended with a ranking loss that encourages positive pairs to score above sampled unlabeled pairs, or with an auxiliary relation-prediction loss that trains the graph encoder to reconstruct held-out edges. Auxiliary losses can improve representation learning, but they must not use the same edges that are reserved for DTI evaluation. Otherwise, the model may learn the test relation through an auxiliary objective and produce inflated performance.

4.6 Scalable execution and hardware-aware computation

Heterogeneous biomedical graphs can contain millions of nodes and tens of millions of edges, and relation-specific message passing introduces irregular memory access and sparse matrix operations. Full-batch training is often infeasible. Mini-batch neighbor sampling, relation-aware subgraph construction, and cached embeddings for high-degree nodes are necessary for practical execution. The computational pattern is dominated by sparse gathers,

scatter-add operations, and small matrix multiplications whose efficiency depends on the graph structure and the number of relation types.

Hardware-level acceleration of graph convolution is a relevant complement to algorithmic scalability because irregular sparse operations can underutilize dense accelerators [11]. A DTI implementation should therefore separate the graph sampling layer from the dense neural network layer, batch edges by relation type, and avoid materializing one dense adjacency matrix for the entire heterogeneous graph. Mixed-precision arithmetic can reduce memory pressure, but attention scores and aggregation accumulators may require higher precision to avoid numerical instability when neighborhoods are large.

Training is typically performed with adaptive stochastic optimization [12]. Early stopping should be based on a validation split that respects biological separation, not on random edge masking alone. Gradient clipping and weight decay help stabilize relation-specific attention, particularly when some relation types are rare. The number of message-passing layers should be kept modest unless the graph is carefully sampled, because overly deep propagation can mix distant and weakly related evidence.

4.7 Calibration and uncertainty

A ranking score is not automatically a calibrated probability. A model may rank interactions well while systematically overestimating or underestimating absolute risk, which is problematic when scores are used to prioritize assays or repositioning candidates. A post hoc calibration layer can map raw scores to calibrated probabilities using a monotone transformation. One simple form is given below.

$$p(cal) = \sigma(a \cdot logit(s(d,t)) + b)$$

Here p(cal) is the calibrated probability, logit(·) is the inverse sigmoid, and a and b are calibration parameters estimated on a held-out calibration set. More flexible calibrators, such as isotonic regression or temperature scaling, can be used when sufficient validation data are available. Calibration should be evaluated on splits that differ by drug family, target family, or temporal curation period, because calibration error often increases under distribution shift.

Uncertainty can be represented through predictive entropy, ensemble variance, or Monte Carlo dropout [13]. Calibration metrics and proper scoring rules are necessary because modern neural networks can be overconfident even when their ranking performance is strong [14]. In DTI applications, uncertainty estimates are useful not only for abstention but also for active learning, where the next assay should be chosen to reduce ambiguity among plausible targets.

5. Evaluation Logic and Error Analysis

5.1 Data sources and benchmark construction

A rigorous evaluation of heterogeneous DTI models requires datasets with explicit provenance, positive evidence, and controlled splits. Public binding databases such as BindingDB [16] and DrugBank [17] provide interaction and drug annotation data, while kinase selectivity panels offer well-characterized positive and negative measurements for specific target families [15]. Protein interaction and functional association resources such as STRING [18] can supply protein-protein and pathway context, and perturbation resources such as the Connectivity Map [19] can provide expression-based drug context. Integrated knowledge graphs such as Hetionet [20] illustrate how heterogeneous biomedical relations

can be assembled for systematic drug repurposing, although their coverage and curation policies must be considered when they are used as model inputs.

The benchmark should distinguish three prediction settings: warm-start prediction, in which both drugs and targets have observed interactions; semi-cold-start prediction, in which one entity has limited interactions; and cold-start prediction, in which an entity is unseen or has very few edges. Random edge splitting is insufficient for cold-start assessment because it allows the same drug or target to appear in both training and test sets. Drug-wise, target-wise, and temporal splits provide complementary views of generalization. Temporal splitting is particularly informative because it simulates prospective prediction and reduces the chance that literature-derived edges leak future knowledge into training.

5.2 Metrics and calibration assessment

Because DTI datasets are highly imbalanced, accuracy is not an appropriate primary metric. Area under the receiver operating characteristic curve measures ranking across thresholds, while area under the precision-recall curve is more informative when positives are rare. Precision at fixed recall, recall at fixed precision, and enrichment among the top-ranked candidates are directly relevant to experimental prioritization. For affinity prediction, regression metrics such as mean squared error or Pearson correlation can be reported alongside classification metrics, but the two tasks should not be conflated.

Calibration should be evaluated with reliability diagrams and expected calibration error, and proper scoring rules such as the Brier score should accompany ranking metrics. Performance should be reported by target family, drug class, degree stratum, and edge-confidence stratum. A model that performs well only on high-degree proteins or well-studied drugs may be exploiting database coverage rather than learning transferable binding determinants.

5.3 Failure modes and leakage risks

Several failure modes are specific to heterogeneous DTI prediction. Similarity leakage occurs when drug-drug or target-target similarity edges are computed from the same interaction labels used for evaluation. Neighborhood leakage occurs when a test pair is connected by a short path through a training edge that directly reveals the test relation. Degree bias occurs when high-degree hubs dominate message passing and cause the model to overpredict interactions for well-studied entities. Relation sparsity occurs when rare relation types are overwhelmed by abundant but weakly informative edges.

These risks require explicit diagnostics. The model should be retrained with similarity edges removed or recomputed without test labels. Neighbor sampling should be audited to ensure that test nodes are not included in sampled subgraphs during training. Performance should be stratified by degree, and ablation studies should remove each relation type to determine whether it contributes unique information or merely duplicates feature-based signals. A relation that improves validation performance but degrades cold-start performance may be acting as a shortcut rather than a generalizable biological signal.

6. Analytical Assessment and Boundary Conditions

6.1 Information gain versus noise amplification

Heterogeneous integration is beneficial only when added relations carry information about the target relation that is not already present in node features. A pathway edge can help

if pathway membership constrains the set of plausible targets, but it can also introduce noise if pathways are broad or tissue-independent. A disease association can support repositioning, but it is not direct binding evidence. The proposed framework addresses this distinction through relation-specific transformations and edge-confidence weights, but it cannot create information that is absent from the graph. The practical boundary condition is that added data sources should improve coverage of relevant biology while remaining sufficiently curated to avoid dominated noise channels.

This point has an architectural consequence. If a new relation type is added, it should receive its own transformation and attention parameters, and its contribution should be evaluated by ablation. Adding many weak relations to a shared message-passing channel can increase variance without increasing signal. The model's capacity should therefore be allocated according to relation informativeness, not simply according to relation availability.

6.2 Cold-start behavior and degree bias

Cold-start prediction is the most demanding setting because graph propagation depends on neighborhood coverage. For a drug with no known targets, message passing through drug-drug similarity may still provide useful context, but only if those similarities are based on chemical structure rather than shared target labels. For a protein with few interaction partners, sequence and domain features must carry more of the predictive burden. Type-specific encoders help in this setting because they preserve feature information even when graph structure is sparse. Relation-aware attention also helps because it can down-weight uninformative neighborhoods rather than forcing propagation through them.

Degree bias remains a structural risk. High-degree nodes appear in many sampled subgraphs and can dominate training. Weighted normalization and degree-aware sampling can reduce this effect, but they must not be tuned on the test set. Stratified reporting is essential: a model with strong overall performance but poor performance on low-degree targets may be less useful for novel target discovery than a slightly weaker model with more uniform performance.

6.3 Interpretability and biological plausibility

Relation-aware attention provides a partial interpretability mechanism because it identifies which relation types and neighbors contribute to a prediction. However, attention weights are not causal explanations. A high attention weight on a pathway neighbor indicates that the model used that relation, not that the pathway causes the interaction. Interpretability should therefore be evaluated through ablation, relation-perturbation analysis, and comparison with known biological mechanisms. Pathway enrichment among top-ranked targets, enrichment of predicted targets in known drug mechanisms, and consistency across related drugs can provide external plausibility checks. These checks are not substitutes for experimental validation, but they help distinguish biologically coherent predictions from graph-structural artifacts.

6.4 Distribution shift and temporal drift

Biomedical knowledge accumulates over time, and interaction databases change as assays are performed and curation improves. A model trained on historical data may encounter new drugs, new target families, and revised relation confidence in deployment. Temporal validation and calibration on later data are therefore necessary. The framework's edge-confidence weighting and calibration layer can partially mitigate drift, but they do not replace

prospective monitoring. If the model is used for candidate prioritization, predictions should be accompanied by uncertainty estimates and by evidence summaries that identify which relations supported the score.

7. Implications for Translational Use

The framework is most naturally positioned as a prioritization tool rather than an autonomous decision system. Its output can rank candidate drug-target pairs for experimental follow-up, suggest repositioning hypotheses, or identify targets that are supported by multiple relation types. The calibrated probability and uncertainty estimate are important in this context because experimental resources are limited and false positives have real costs. A top-ranked candidate supported by direct binding evidence, pathway context, and expression data is more interpretable than a candidate supported only by a high-degree graph neighborhood.

The design also has implications for data curation. Because the model distinguishes relation types and edge confidence, it creates an incentive to record assay provenance, binding strength, and curation status. Biomedical databases that provide such metadata enable more reliable integration than databases that record only binary interactions. This is not merely an engineering preference; it affects the statistical validity of the model because the meaning of a positive label depends on how the interaction was measured.

8. Conclusion

This manuscript developed a heterogeneous graph neural network framework for drug-target interaction prediction using integrated biomedical data. The framework represents drugs, proteins, diseases, pathways, and contextual entities as typed nodes; encodes relation semantics through type-specific transformations; aggregates evidence with relation-aware attention; weights edges by confidence; and trains with positive-unlabeled awareness and calibrated uncertainty. The analytical assessment shows that the value of heterogeneous integration depends on preserving relation semantics, controlling leakage and degree bias, and evaluating calibration and cold-start performance alongside ranking metrics. The main scientific implication is that graph-based DTI prediction should be treated as evidence synthesis under uncertainty rather than as a homogeneous link-prediction problem. This framing provides a technically grounded basis for implementation, benchmark design, and translational prioritization.

References

1. Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: Deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821–i829.

2. Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., & Zeng, J. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nature Communications, 8, 573.

3. Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457–i466.

4. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In Proceedings of the International Conference on Learning Representations.

5. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In Proceedings of the International Conference on Learning Representations.

6. Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. In The Semantic Web (ESWC 2018) (pp. 593–607). Springer.

7. Wang, X., He, X., Cao, Y., Liu, M., & Chua, T.-S. (2019). KGAT: Knowledge graph attention network for recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 950–958).

8. Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous graph neural network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 793–803).

9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

10. Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 213–220).

11. Xie, X., Peng, H., Hasan, A., Huang, S., Zhao, J., Fang, H., ... & Ding, C. (2023, October). Accel-gcn: High-performance gpu accelerator design for graph convolution networks. In 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD) (pp. 01-09). IEEE.

12. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations.

13. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059).

14. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330).

15. Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., & Zarrinkar, P. P. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nature Biotechnology, 29(11), 1046–1051.

16. Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Research, 44(D1), D1045–D1053.

17. Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Research, 46(D1), D1074–D1082.

18. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., ... & von Mering, C. (2021). The STRING database in 2021: Customizable protein–protein

networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research, 49(D1), D605–D612.

19. Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., ... & Golub, T. R. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell, 171(6), 1437–1452.e17.

20. Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., ... & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife, 6, e26726.