Methodological Insights into Cancer Driver Gene Prediction: Comparing Graph-Based and Traditional ML Approaches

Renan Soares de Andrades[1,2], **Mariana Recamonde-Mendoza**[1,2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]Bioinformatics Core, Hospital de Clínicas de Porto Alegre (HCPA)
Porto Alegre – RS – Brazil

rsandrades@inf.ufrgs.br, mrmendoza@inf.ufrgs.br

Abstract. *This study investigates the impact of key methodological choices on cancer driver gene (CDG) prediction using graph neural networks (GNNs) and traditional machine learning (ML) models. We evaluate three GNN architectures, four ML algorithms, three protein–protein interaction networks, multiple node feature configurations (single-omics, multi-omics, centrality measures), and three strategies to mitigate class imbalance. Graph Convolutional Networks consistently outperform other GNNs, while Gradient Boosted Trees remain competitive when structural features are included. Node centrality measures further enhance prediction across models. These results underscore the role of feature design and model selection in achieving accurate and robust CDG prediction.*

1. Introduction

Cancer is a multifactorial disease influenced by genetic and environmental factors, and remains a major global health challenge, with over 35 million new cancer cases predicted for 2050 – a 77% increase from estimates for 2022 [WHO 2024]. A key goal in cancer genomics is the identification of cancer driver genes (CDGs), whose alterations provide a selective advantage to tumor cells by promoting proliferation, survival, and metastasis [Rogers et al. 2020]. CDGs are often hidden among a large number of passenger mutations, which appear during tumor development but do not play a role in causing cancer. Despite advances in large-scale sequencing and computational analysis, identifying CDGs remains challenging due to their low mutation frequency, tissue- and patient-specific contexts, and the complexity of underlying molecular networks [Ostroverkhova et al. 2023].

Recent progress in machine learning (ML), particularly in deep learning, offer new tools for analyzing complex biological data. Among these, graph neural networks (GNNs) have gained prominence for modeling intricate relationships present in biological systems, such as protein-protein interaction (PPI) networks [Zhang et al. 2021]. By capturing the topological structure and contextual dependencies within these networks, GNNs can provide deeper insights into the functional relevance of genes and other molecular components. Integrating PPI networks with multi-omics data using GNNs has shown promise in enhancing the prediction of CDGs [Peng et al. 2023, Zhang et al. 2021], while also helping to refine PPI data by attenuating the impact of spurious or noisy connections. These methodological advances extend beyond CDG identification, improving broader applications in tumor classification, prognosis, and therapeutic target discovery.

Despite their promise, applying GNNs to CDG prediction involves significant methodological challenges. Key choices, such as node feature selection, strategies for handling class imbalance, and model architecture, can greatly affect predictive performance. Moreover, incorporating structural properties of nodes in PPI networks, such as degree, betweenness, and clustering coefficient, which are well-established metrics in systems biology, may improve model reliability and interpretability. Yet, whether these graph-based features can yield performance comparable to GNNs when used with traditional ML algorithms or contribute complementary information to GNN-based models, remains insufficiently explored in the current literature related to CDG prediction.

This study investigates the methodological implications of using GNNs and traditional ML algorithms for CDG prediction, focusing on how key design choices affect model performance. Rather than proposing a new state-of-the-art (SOTA) method, we aim to evaluate how factors such as omics data selection, feature engineering, class imbalance handling, and model architecture influence predictive outcomes. We compare the performance and behavior of GNNs and traditional ML models under controlled experimental conditions, testing three GNN architectures, four ML algorithms, six feature variants, and three imbalance-handling strategies across three different PPI networks. By systematically assessing these aspects, our study offers insights into the strengths and limitations of each modeling approach, supporting the development of more effective and reliable tools for cancer genomics and precision oncology.

2. Related works

Recent studies on ML methods for CDG prediction have explored a wide range of algorithms and data types, revealing both emerging trends and persistent methodological limitations, such as class imbalance, limited overlap between predictions from different methods, and the lack of consistently high-quality annotated datasets [Andrades and Recamonde-Mendoza 2022]. Among the most notable trends is the increasing use of deep learning techniques, particularly GNNs [Zhang et al. 2025]. This class of algorithms has been highlighted as particularly promising for CDG discovery, due to its success in other predictive tasks in bioinformatics [Zhang et al. 2021].

A prominent example is EMOGI, which uses Graph Convolutional Networks (GCNs) to integrate multi-omics data as node features for CDG prediction [Schulte-Sasse et al. 2021]. The model is evaluated across five PPI networks and compared to other ML- and network-based baseline models using only omics data, only network data, and hybrid (network and omics) methods from the literature. While their results demonstrate superior predictive performance for their approach, the study does not clearly isolate which methodological choices were most responsible for the observed improvements.

Other representatives approaches, such as CancerGATE [Jung et al. 2024] and MCDHGN [Wang et al. 2024], further demonstrate the potential of GNN-based architectures. CancerGATE uses graph attention autoencoders over PPI networks enriched with multi-omics data, reporting improved performance across several cancer types. An ablation study confirms that both the inclusion of multi-omics features and their embedding strategy contribute significantly to the results. MCDHGN, in turn, integrates heterogeneous biological networks with graph attention mechanisms and evaluates how changes

to the architecture or input vectors affect performance.

Song et al. [Song et al. 2023] combine multi-omics and network-derived features using a GNN composed of a Graph Attention (GAT) layer followed by two GraphSAGE layers over a PPI network. In addition to comparisons with SOTA methods, they evaluate performance against classical ML algorithms, such as Support Vector Machine (SVM) and Random Forest. An ablation study, removing model components or feature groups, shows that both the hybrid architecture and the use of multi-omics data are key to improving predictive performance.

Despite these advances, prior studies focus their comparative studies narrowly on components specific to their own frameworks. These works do not address broader methodological questions, such as the relative impact of omics feature choices across distinct algorithms, imbalance handling strategies, or the comparison between GNNs and traditional ML algorithms under standardized conditions. In contrast, our study addresses this gap by evaluating how key design decisions influence CDG prediction performance in both GNN-based and traditional ML models under unified experimental conditions.

3. Materials and Methods

In this section, we describe the PPI networks that define the graph structure, followed by the adopted omics data and the selection of positive and negative CDG examples as node labels. Next, we detail the data pre-processing and integration steps, and finally, the procedures for model training and evaluation. All datasets were obtained from publicly available sources.

3.1. Protein-protein interaction networks

For network data, the Network Data Exchange (NDEx) database was employed, a well-established resource in human disease research [Pratt et al. 2017]. Three pre-processed PPI networks were selected, all of which had been included in a systematic evaluation of molecular networks for their efficacy in identifying gene sets associated with human diseases [Huang et al. 2018]. To manage computational complexity and accommodate the extensive number of experiments conducted, the selected networks were limited to those containing fewer than 150k edges: the Human Protein Reference Database (HPRD), MultiNet, and IRefIndex (IREF). Table 1 provides detailed information about the original number of nodes and edges for each PPI network.

3.2. Omics data

To assess how the choice of omics data influences the performance of predictive models, we used multi-omics datasets from The Cancer Genome Atlas (TCGA), covering

Table 1. Number of nodes, edges and node labels distribution in the PPI networks

	Original		After pre-processing		Node Labels		
	Edges	Nodes	Edges	Nodes	Positive	Negative	Unlabeled
HPRD	36,867	9,453	36,844	9,438	793	4,873	3,772
MULTINET	109,595	14,445	108,568	13,987	868	8,486	4,633
IREF	133,548	14,667	133,095	14,627	874	8,838	4,915

over 8,000 samples across 16 cancer types. The data included single nucleotide variants (SNVs), copy number alterations (CNAs), DNA methylation, and gene expression.

Following established preprocessing methods, mutation rates, copy number changes, methylation levels in promoter regions, and gene expression differences were computed for each gene [Schulte-Sasse et al. 2021]. These measures were summarized into feature matrices, where each matrix had dimensions corresponding to the number of genes and cancer types. All features were normalized to ensure consistency across data types. Finally, the four individual matrices were combined into a single integrated dataset, capturing the molecular profile of over 22,000 genes across cancers. This dataset forms the basis for further analysis and the development of predictive models for CDGs.

3.3. Positive and negative examples of CDGs

In this study, genes were divided into three categories: driver genes (positively labeled), non-cancer genes or passenger genes (negatively labeled), and unlabeled genes, whose roles in cancer remain uncertain. The driver gene set was compiled from well-established sources, including the Network of Cancer Genes (NCG), the Tier 1 list from the Cancer Gene Census (CGC), and the OncoKB database, resulting in 907 unique genes after removing duplicates.

Unlabeled genes, those with possible but unconfirmed links to cancer, were drawn from various resources, such as Tier 2 CGC entries, cancer-related gene collections, biological pathways related to cancer, and disease-associated genes listed in the OMIM database. This category comprised 5,777 genes, representing a wide range of potential candidates. Defining non-cancer genes required stricter filtering. To exclude any genes related to cancer or other diseases, all genes with known disease associations in OMIM were removed from the dataset. The final classification revealed a substantial imbalance (Table 1), with driver genes significantly underrepresented, a challenge for building effective predictive models. We note that for the comparative experiments in this study, aimed at evaluating model performance, only genes with positive (driver) and negative (non-cancer) labels were used for evaluation purposes.

3.4. Data pre-processing and integration

During data pre-processing, measures were taken to ensure consistency across datasets. Gene identifiers were standardized using HGNC-approved symbols[1], verified with the HGNC Multi-symbol Checker. A mapping dictionary was created to maintain uniform naming across sources, which was essential for data integration. The analysis focused on genes present in both the PPI networks and the multi-omics dataset. Only overlapping genes were retained to ensure consistency in features and network information. Details on the resulting network structures are provided in Table 1.

Additionally, centrality measures, including degree, betweenness, closeness, and clustering coefficient, were calculated using the igraph package to capture each gene's structural relevance within the networks. Degree values were normalized to match the scale of the omics features. The final dataset integrated gene identifiers, omics profiles, and network-based features, serving as the foundation for the experimental and modeling phases of the study.

[1] https://www.genenames.org/

3.5. Class imbalance mitigation

To address the class imbalance inherent to the CDG prediction problem, we applied Random Undersampling, Balanced Cross-Entropy, and Focal Loss. Random undersampling equalizes the number of positive and negative labels. In our work, it means reducing the number of negative examples for training. Balanced cross-entropy adds a weighting factor to account for class imbalance, and was applied with $\alpha = 0.85$. Focal Loss [Lin et al. 2017] is a dynamic cross-entropy loss that employs a scaling factor $(1 - p_t)^\gamma$, where γ is a configurable hyperparameter that adjusts the rate at which easy examples are down-weighted. We explored three combinations determined after initial exploration: γ=0.5 and α=0.50; γ=1.0 and α=0.25; γ=2.0 and α=0.25.

3.6. Model training and evaluation

The study employed two modeling strategies: GNNs and traditional ML algorithms. Three GNN architectures, Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and GraphSAGE, were evaluated for a node classification task. Unless otherwise specified, the categorical cross-entropy was used as the default loss function. Models were implemented using the StellarGraph library, which offers a flexible framework for graph-based learning.

As baselines, four traditional ML algorithms were also applied: Support Vector Machines (SVM), Random Forests (RF), Gradient Boosting Trees (GBT), and Artificial Neural Networks (ANN). These models were developed using scikit-learn[2] and Keras, with default settings retained for most algorithms, except for ANN, which was adjusted to align more closely with the GNN configurations.

Model performance was assessed through an 80:20 data split, with 5-fold cross-validation on the training set to ensure reliable estimates. The test set was reserved exclusively for final performance reporting, while training and validation sets were monitored to prevent overfitting. A fixed random seed ensured consistent comparisons within a single method. The evaluation metrics included AUC-PR, AUC-ROC and accuracy. In this paper, we emphasize AUC-PR due to class imbalance, other metrics are presented in our project's GitHub repository[3].

4. Experiments and Results

The main objective of this work is to carry out a series of experiments aimed at evaluating and comparing various methodological decisions for predicting CDGs with GNNs and traditional ML algorithms. The subsequent sections present our experimental setup and results. Our analyses and discussion will be oriented by five main research questions:

- How does the choice of omics data impact GNNs' predictive performance?
- How do traditional ML algorithms compare to GNNs on the same omics data?
- Can GNNs benefit from node centrality measures as features?
- What is the best strategy to handle class imbalance in node prediction?
- Which graph neural network model performs best for CDG prediction?

Our discussion is focused mainly on the test set performance, with the most significant results presented in this section. Additional results can be found in our repository.

[2] https://scikit-learn.org/

[3] https://github.com/renandrades/Paper_GNNs_BSB_2025

4.1. Experiments setup

Hyperparameters configuration for GNNs and traditional ML algorithms can be found in the project's GitHub repository. Hyperparameter tuning was intentionally excluded from this study to concentrate our analysis on evaluating the impact of other methodological decisions in model development. For GNNs, most hyperparameters were set to the default values provided by the StellarGraph library. The ANN was configured with two hidden layers of 64 neurons each, using the cross-entropy loss function, a dropout rate of 0.01, and a learning rate of 0.001. Default hyperparameters from Scikit-learn (version 1.1.1) were used for RF, GBT, and SVM models. All experiments employed the cross-entropy loss function unless explicitly stated otherwise, such as in analyses addressing class imbalance mitigation strategies.

4.2. How does the choice of omics data used as node features impact the predictive performance of GNNs?

Similar to traditional ML algorithms, GNNs leverage node feature vectors to detect patterns associated with CDGs. By integrating these feature vectors with the local structural information of nodes within a graph, GNNs enhance their ability to predict node labels. Consequently, it is reasonable to hypothesize that different types of omics data used as node features may significantly influence model performance.

To investigate this, we conducted five experiments for each of the PPI networks (HPRD, MULTINET, and IREF). Initially, models were trained using a single omics data type (*i.e.,* mutations, CNA, DNA methylation, or gene expression) as node features. Then, a multi-omics approach was tested by concatenating all four omics types into a single feature vector per node. These experiments used the standard cross-entropy loss function and did not apply class imbalance mitigation strategies.

In addition to omics data, traditional node centrality measures, such as degree, betweenness, closeness, and clustering coefficient, were considered in our analyses. These metrics have been used as handcrafted features for disease gene prediction in the field of systems biology, capturing both global and local network characteristics before the advent of graph-based learning. To examine whether these features could still contribute information to GNNs and improve model performance, centrality measures were calculated for each PPI network using the Igraph[4] package and combined with the multi-omics features for a final round of training and evaluation.

The results, summarized in Table 2, present the AUC-PR scores, with the highest performance for each algorithm and PPI network highlighted in bold. Overall, the GNN models exhibited lower-than-expected classification performance, as indicated by modest AUC-PR values. Still, these results are comparable to performance values already reported in the literature [Peng et al. 2023, Jung et al. 2024], suggesting that challenges remain in accurately identifying true CDGs while minimizing false positives. Given the severe class imbalance in the dataset, these results are not unexpected and still offer valuable insights for model comparison.

The choice of node features significantly influenced GNN model performance, with substantial variation observed across different PPI networks and algorithms. GraphSage exhibited the highest sensitivity to changes in node features, while GAT and GCN

[4] https://igraph.org/

Table 2. AUC-PR performance comparison among GNNs for variations of types of omics data used as node features.

Features	PPI Network	GNNs			Traditional ML algorithms			
		GAT	GCN	GraphSAGE	SVM	RF	GBT	ANN
Mutations	HPRD	0.3565	0.4470	**0.4089**	0.3272	0.3664	0.3753	0.2252
CNA		0.2761	0.3992	0.3134	0.1950	0.1702	0.1846	0.1619
DNA Methylation		0.3844	0.3708	0.2402	0.1435	0.1519	0.1487	0.1324
Gene Expression		0.2806	0.2967	0.2397	0.1368	0.1559	0.1564	0.1563
Multi-Omics		0.3258	0.4984	0.3317	0.2595	0.3566	0.3565	0.2119
Multi-Omics and Centralities		**0.3992**	**0.5960**	0.3786	**0.4477**	**0.5231**	**0.5437**	**0.3376**
Mutations	MULTINET	0.2142	0.3523	**0.3842**	0.2819	0.3066	0.3212	0.1726
CNA		0.2865	0.3277	0.1996	0.1247	0.1136	0.1169	0.1220
DNA Methylation		0.2219	0.3185	0.1770	0.0951	0.1122	0.1089	0.0853
Gene Expression		0.2562	0.2506	0.1981	0.0997	0.1071	0.1151	0.0946
Multi-Omics		0.2182	0.3839	0.2671	0.2025	0.2976	0.3066	0.1614
Multi-Omics and Centralities		**0.3648**	**0.4019**	0.3174	**0.3126**	**0.4559**	**0.4505**	**0.2578**
Mutations	IREF	0.2606	**0.4344**	**0.3550**	**0.2598**	0.2929	0.3087	0.1848
CNA		0.3027	0.2844	0.2036	0.1204	0.1201	0.1268	0.1165
DNA Methylation		0.3088	0.2509	0.1482	0.0969	0.1023	0.1044	0.1081
Gene Expression		0.2211	0.2699	0.1850	0.0906	0.1016	0.1077	0.0981
Multi-Omics		0.2144	0.4315	0.2001	0.1958	0.2869	0.3029	0.1519
Multi-Omics and Centralities		**0.3598**	0.4274	0.2427	0.2368	**0.4052**	**0.4230**	**0.2117**

showed more stable performance. Among individual features, mutations yielded the best results in most cases, followed by multi-omics and DNA methylation, whereas CNA contributed the least to predictive performance according to our results.

The ranking of feature types varied depending on the combination of the PPI network and GNN model, reflecting the biological relevance of different omics data. Mutation features were the most informative, aligning with their established role as important evidence for cancer drivers. However, no single feature type consistently outperformed across all scenarios, highlighting the need for a more integrative approach. The multi-omics strategy, which combined all four omics data types, proved particularly beneficial, especially for the GCN model. This finding underscores the importance of leveraging diverse molecular data sources to improve classification accuracy in CDG prediction.

4.3. How do traditional ML algorithms compare to GNNs when applied to the same omics data?

We also aimed to assess the advantages of graph-based learning methods, which leverage the entire network structure, compared to traditional ML approaches that rely solely on hand-crafted features such as omics data and centrality measures. To investigate this, we conducted a series of experiments using SVM, RF, GBT, and ANN, which are widely adopted ML algorithms in bioinformatics.

For these conventional ML models, we framed the problem as a standard classification task in which each instance (*i.e.,* gene) was represented as a feature vector and assigned a corresponding label. When using only omics data as features, the structural information from PPI networks was excluded, meaning the predictive performance was determined solely by the informativeness of the molecular data. To ensure methodological consistency, we conducted experiments separately for each PPI network while maintaining fixed hyperparameters across all models. This approach allowed us to systematically

evaluate the impact of different omics feature sets on the classification performance of each algorithm.

The results for these experiments are presented in Table 2. Traditional ML algorithms demonstrated superior performance when utilizing mutation data; however, their overall predictive capability was lower compared to GNN-based models. Among the four ML algorithms evaluated, GBT consistently achieved the highest performance across all PPI networks, followed by RF. Additionally, the multi-omics feature set emerged as the second most informative data representation across all networks, further reinforcing the relevance of integrating multiple molecular data types for classification tasks.

4.4. Can GNNs benefit from the inclusion of node centrality measures as node features?

Node centrality measures, including degree, betweenness, closeness, and clustering coefficient, have traditionally been utilized in disease gene prediction models. To assess whether incorporating these measures along with multi-omics data could improve performance, we extracted centrality measures from each PPI network and integrated them as additional features. The results indicated that centrality measures significantly enhanced performance for nearly all models and networks (Table 2). While GNNs benefited from this integration, with GAT showing the most substantial improvement of up to 67.81%, traditional ML algorithms experienced even greater enhancement. For these traditional models, the combination of multi-omics data and centrality measures surpassed all other feature configurations.

A comparative analysis (Figure 1) revealed that the best-performing traditional ML algorithm, GBT, demonstrated competitive performance compared to graph-based models when utilizing both multi-omics and centrality features. GBT consistently outperformed GAT and GraphSAGE, and its performance was comparable to that of GCN, achieving the highest score among all evaluated approaches for the MULTINET network. These findings highlight that while GNNs inherently leverage network topology, tradi-

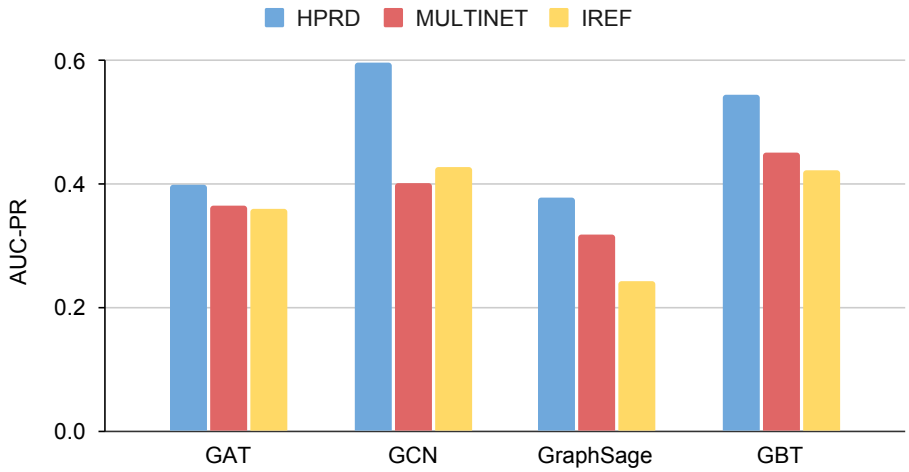

Figure 1. Comparison between GNNs and GBT using multi-omics and node centralities as features.

tional ML models can achieve similar predictive accuracy when provided with explicit structural information through centrality measures.

In summary, while predicting CDGs is generally more challenging for traditional ML algorithms compared to GNNs, these findings underscore the critical importance of considering the network structure. They demonstrate that when traditional models are provided with explicit structural information via centrality measures, they can achieve comparable predictive power. Moreover, these metrics can still help in the prediction process conducted by GNNs. This highlights the significance of centrality features as a complement to multi-omics data, suggesting they play a vital role in improving predictive accuracy in such scenarios.

4.5. What is the best strategy to handle class imbalance in the node prediction problem addressed?

Class imbalance is a recurring issue in cancer genomics and cancer driver prediction, with true driver mutations or genuine CDGs being rarer compared to passenger mutations or neutral genes. While other researchers also face this challenge, only a few discuss or compare strategies for addressing it. In our experiments, we applied five different strategies to mitigate class imbalance among positive and negative CDG examples for each GNN algorithm and PPI network, including Random Undersampling, Balanced Cross-entropy, and three variations of the Focal Loss function. Notably, Balanced Cross-Entropy resembles a Focal Loss function with $\gamma = 0$, and it utilizes the α parameter to balance the importance of positive and negative instances (we used $\alpha = 0.85$). The Cross-entropy cost function was employed in two scenarios: in the initial experiments (our "baseline") and when applying the Random Undersampling technique to the data.

Table 3 presents the results of our experiments, with the best results highlighted in bold. Multi-omics and centrality data were used as standard throughout all these experiments. In summary, we observed performance improvements in most cases when addressing class imbalance. Balanced Cross-Entropy delivered promising results, while the undersampling strategy was generally less effective, probably due to the decreased labeled dataset size. Although the results still fall below the desired level of predictive performance, they exhibit potential for further enhancement through hyperparameter optimization, particularly in the case of Balanced Cross-Entropy and Focal Loss. Given that Balanced Cross-Entropy is a special case of Focal Loss, we identified Focal Loss as the most promising approach for handling class imbalance in our domain.

4.6. Which graph neural network algorithm performs best for CDG prediction?

Finally, our last research question focuses on investigating the most suitable GNN algorithm for the task of predicting CDGs. Based on a comprehensive evaluation of the experimental results, GCN emerged as the most promising algorithm for CDG prediction. GCN consistently demonstrated superior performance across nearly all tested conditions, including different PPI networks and variations in node feature vector compositions (see additional data in our project's GitHub repository). Notably, the highest predictive performance in this study was achieved using GCN in conjunction with multi-omics data and centrality measures, as seen in Figure 2 for the HPRD network. However, similar conclusions were found in the other networks, highlighting GCN robustness in integrating

Table 3. AUC-PR peformance comparison for the test set between strategies to address the class imbalance issue.

Strategy	PPI Network	Algorithm		
		GAT	GCN	GraphSAGE
Cross-entropy	HPRD	0.3992	**0.5960**	0.3786
Undersampling		0.3275	0.5479	0.3332
Balanced cross-entropy (α=0.85)		**0.4215**	0.5918	0.4069
Focal loss (γ=0.5; α=0.50)		0.3873	0.5235	0.4230
Focal loss (γ=1; α=0.25)		0.3581	0.5756	0.3985
Focal loss (γ=2; α=0.25)		0.3792	0.5583	**0.4480**
Cross-entropy	MULTINET	0.3648	0.4019	0.3174
Undersampling		0.2817	0.4521	0.2188
Balanced cross-entropy (α=0.85)		0.3551	**0.4661**	**0.3929**
Focal loss (γ=0.5; α=0.50)		**0.4143**	0.4177	0.3264
Focal loss (γ=1; α=0.25)		0.3363	0.2978	0.3196
Focal loss (γ=2; α=0.25)		0.2753	0.4481	0.3822
Cross-entropy	IREF	0.3598	0.4274	0.2427
Undersampling		0.1714	0.4384	0.1887
Balanced cross-entropy (α=0.85)		0.3135	**0.4841**	**0.3013**
Focal loss (γ=0.5; α=0.50)		0.3546	0.4664	0.2296
Focal loss (γ=1; α=0.25)		0.2977	0.3776	0.2832
Focal loss (γ=2; α=0.25)		**0.3917**	0.4488	0.2928

diverse biological information sources. Therefore, GCN was pointed out as the best GNN algorithm for the addressed problem according to our experiments.

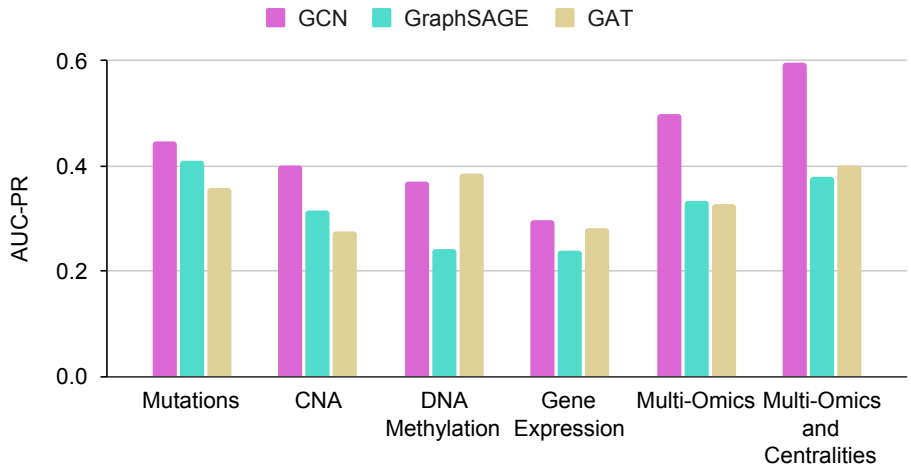

Figure 2. Test set performance for HPRD network.

5. Conclusion

Identifying CDGs remains a major challenge due to the complexity of tumor biology. Graph-based methods, particularly GNNs, have shown promise by leveraging the structure of biological networks, potentially outperforming traditional ML models. Despite

growing interest, the impact of specific design choices has not been systematically assessed. Likewise, comparative analyses between GNNs and traditional ML models under unified conditions remain limited. To address these gaps, we conducted a series of controlled experiments involving: (i) three GNN architectures; (ii) three protein–protein interaction (PPI) networks; (iii) multiple node feature configurations, including single-omics, multi-omics, and centrality-based features; and (iv) three class imbalance strategies. We also included four traditional ML algorithms as baselines to contextualize the performance of GNN-based models.

Based on our results, we confirmed that GNNs generally outperformed traditional ML algorithms by incorporating node connectivity into the learning and classification processes. Among the GNN architectures evaluated, GCN consistently achieved the most robust and reliable performance across all PPI networks and feature configurations. However, GBT is a strong candidate for a deeper investigation involving hyperparameters optimization. We found that both approaches are sensitive to variations in the node features. However, while traditional ML algorithms show variations in the AUC-PR values between 0.12 and 0.4 (considering the difference between the minimum and maximum scores), GNNs performance difference vary in a narrower range (0.12 and 0.21), suggesting higher robustness in their predictive capacity.

Finally, it was interesting to note that node centrality measures contributed to performance gains not only in traditional ML algorithms, but also in GNN-based models. Among traditional ML algorithms, 11 out of 12 combinations of algorithms and PPI networks achieved their best performance when centrality measures were concatenated with multi-omics. Regarding GNNs, 5 out of 9 combinations showed the same pattern. Given that GNNs, by nature, capture both local and global structural properties of the PPI networks, it is curious that explicitly including these well-known metrics still aggregate information for their learning process. This may open new perspectives for feature engineering in GNN-based approaches. Investigating whether the same conclusions are achieved for larger and more PPI networks would be an interesting topic for future works.

In general, our findings underscore the value of integrating biological network topology with omics data to enhance CDG identification. Rather than aiming to outperform existing approaches in the literature, this work focused on understanding how some methodological choices directly influence predictive performance. By providing a systematic and comparative analysis of traditional and graph-based learning methods, this study offers valuable insights to guide the development, selection, and comparison of predictive models in cancer genomics.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001; by grants from the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) [Projects No. 21/2551-0002052-0 and No. 22/2551-0000390-7 (CIARS)]; and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [Project No. 308075/2021-8].

References

Andrades, R. and Recamonde-Mendoza, M. (2022). Machine learning methods for prediction of cancer driver genes: a survey paper. *Briefings in Bioinformatics*, 23(3). bbac062.

Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., and Ideker, T. (2018). Systematic evaluation of molecular networks for discovery of disease genes. *Cell Systems*, 6(4):484–495.

Jung, S., Wang, S., and Lee, D. (2024). CancerGATE: Prediction of cancer-driver genes using graph attention autoencoders. *Computers in Biology and Medicine*, 176:108568.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Ostroverkhova, D., Przytycka, T. M., and Panchenko, A. R. (2023). Cancer driver mutations: predictions and reality. *Trends in Molecular Medicine*, 29(7):554–566.

Peng, W., Wu, R., Dai, W., and Yu, N. (2023). Identifying cancer driver genes based on multi-view heterogeneous graph convolutional network and self-attention mechanism. *BMC Bioinformatics*, 24(1):16.

Pratt, D., Chen, J., Pillich, R., Rynkov, V., Gary, A., Demchak, B., and Ideker, T. (2017). Ndex 2.0: a clearinghouse for research on cancer pathways. *Cancer Research*, 77(21):e58–e61.

Rogers, M. F., Gaunt, T. R., and Campbell, C. (2020). Prediction of driver variants in the cancer genome via machine learning methodologies. *Briefings in Bioinformatics*, 22(4). bbaa250.

Schulte-Sasse, R., Budach, S., Hnisz, D., and Marsico, A. (2021). Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence*, 3(6):513–526.

Song, H., Yin, C., Li, Z., Feng, K., Cao, Y., Gu, Y., and Sun, H. (2023). Identification of cancer driver genes by integrating multiomics data with graph neural networks. *Metabolites*, 13(3):339.

Wang, L., Zhou, J., Wang, X., Wang, Y., and Li, J. (2024). MCDHGN: heterogeneous network-based cancer driver gene prediction and interpretability analysis. *Bioinformatics*, 40(6):btae362.

WHO, W. H. O. (2024). Global cancer burden growing, amidst mounting need for services. `https://shorturl.at/8AUlY` [Accessed: May 2025].

Zhang, H., Lin, C., Chen, Y., Shen, X., Wang, R., Chen, Y., and Lyu, J. (2025). Enhancing molecular network-based cancer driver gene prediction using machine learning approaches: Current challenges and opportunities. *Journal of Cellular and Molecular Medicine*, 29(1):e70351.

Zhang, X.-M., Liang, L., Liu, L., and Tang, M.-J. (2021). Graph neural networks and their current applications in bioinformatics. *Frontiers in Genetics*, 12.