Adaptive Chebyshev Graph Neural Network for Cancer Gene Prediction with Multi-Omics Integration

Sa Li, Jonah Shader, Abhijeet Bhattacharya, Tianle Ma
Oakland University
{sa,jonahshader,bhattacharya,tianlema}@oakland.edu

***Abstract*— The rapid expansion of high-throughput molecular data introduces substantial computational challenges in identifying cancer driver genes. Since tumorigenesis is driven by both genetic and non-genetic factors, there is a critical need for predictive models that can integrate diverse data sources while remaining interpretable. We present ACGNN, a framework designed to predict cancer genes by integrating diverse data modalities, including PPI networks and pan-cancer multi-omics data, into a unified predictive model. It leverages graph convolutional networks (GCNs) to generate low-dimensional embeddings, which are further refined using adaptive Chebyshev graph neural networks for more precise driver gene identification. ACGNN dynamically adjusts the receptive field, allowing for more flexible and effective feature aggregation. Our method achieves a 25.86% average improvement in AUPRC over state-of-the-art methods, effectively identifying both well-established and novel cancer driver genes. We believe our method introduces a new approach with the ability to capture biologically relevant features, providing valuable insights for cancer research and precision medicine.**

***Index Terms*—cancer driver genes, graph neural network, node embeddings, Chebyshev networks.**

I. Introduction

CANCER is a genetic disease caused by the accumulation of mutations, but only a small subset of mutated genes—known as driver genes—actively contribute to tumor development [1]–[5]. Identifying these driver genes with high accuracy is essential for understanding cancer pathogenesis and developing targeted therapies [6]–[8]. Significant efforts, such as the Network of Cancer Genes (NCG) [9] and the COSMIC Cancer Gene Census (CGC) [10], have been made to annotate cancer genes based on mutation data.

Existing driver gene prediction methods primarily analyze patient groups within specific cancer types, leveraging gene expression and mutation data. These approaches are mainly categorized into mutation frequency-based methods [11], [12] and network-based methods [13]–[15]. Mutation frequency-based models rank genes by how significantly their mutation rates deviate from background expectations, while network-based methods incorporate pathway and gene interaction data [16], [17]. However, the reliability of network-based methods is often compromised by incomplete or noisy biological interactions [18]. Additionally, many existing approaches rely on omics data for gene representation learning while overlooking the structural information inherent in biomolecular networks [19], [20]. Methods that incorporate handcrafted network-based features often fail to capture the complex, non-linear structures of these networks, leading to suboptimal performance in driver gene identification [21], [22].

Graph Neural Networks (GNNs) [23] have shown promise in bioinformatics tasks, particularly in association prediction [24]–[26]. Cancer gene prediction requires models that effectively integrate diverse biological networks and omics data. Traditional Graph Neural Networks (GNNs), such as Graph Convolutional Networks (GCNs), rely on fixed-order convolutions, limiting their adaptability to varying graph structures. To overcome these limitations, we propose Adaptive Chebyshev Graph Neural Network (ACGNN), a novel approach that dynamically adjusts receptive fields using Chebyshev polyno-

mial approximations. Given the crucial role of gene feature interactions in refining node representations [27], ACGNN efficiently propagates information while capturing multi-scale topological dependencies. By leveraging Chebyshev polynomials, it flexibly adjusts the receptive field, enabling more effective feature aggregation and higher-order neighborhood learning. Our key contributions can be summarized as follows:

- We propose an ACGNN approach to identify cancer genes by integrating multiple PPI networks and multi-omics data, generating biologically meaningful gene representations using pretrained node embeddings from biomolecular networks.
- We employ adaptive Chebyshev graph convolution with residual connections to enhance feature propagation, addressing challenges posed by deep networks and noisy data.
- We leverage pretrained embeddings to improve model generalization, reducing dependence on large annotated datasets and ensuring more robust identification of cancer driver genes.
- ACGNN introduces an adaptive mechanism that dynamically tunes the polynomial order, improving expressiveness and robustness in multi-omics cancer gene prediction.

II. Related Work

Graph Neural Networks (GNNs) have become indispensable in biological network analysis. In [28], the authors propose EMGNN, a multilayer graph neural network that integrates multiple gene-gene interaction networks with pan-cancer multi-omics data to improve cancer gene prediction. Unlike single-network approaches, EMGNN captures tumorigenesis complexity across diverse networks while incorporating interpretability features to enhance transparency. Li and Nabavi [29] introduce a multimodal GNN for cancer molecular subtype classification, leveraging graph-based representations rather than traditional early or late fusion methods. Similarly, Li et al. [30] develop CGMega, an explainable graph neural network with attention mechanisms for cancer gene module dissection. Peng et al. [31] present MTGCN, a Multi-Task Graph Convolutional Network (GCN) that integrates PPI networks for improved driver gene identification. Further, Peng et al. [32] propose MNGCL, a contrastive learning framework integrating multi-omics datasets to enhance information interaction.

Schulte-Sasse et al. [33] develop EMOGI, a GCN-based explainable method integrating pan-cancer multi-omics and PPI networks, using layer-wise relevance propagation for interpretability. Zhang et al. [34] introduce HGDC, a Heterophilic Graph Diffusion Convolutional Network, utilizing Personalized PageRank (PPR) for improved message propagation in heterogeneous networks. Zhao et al. [36] propose MODIG, a GAT-based framework integrating multi-omics data with diverse gene networks, including co-expression patterns and KEGG pathways, to enhance driver gene identification.

Topology Adaptive Graph Convolutional Networks (TAGCNs) [37] extend traditional GCNs by dynamically adjusting convolutional operations based on graph structure, capturing both local and global topology. Singh et al. [38] further enhance this with GTAGCN, combining Generalized Aggregation Networks with TAGCNs for versatile applications. Despite these advancements, the potential of pretrained embeddings in multi-omics data integration remains underexplored.

Our methode optimizes the Chebyshev polynomial order, eliminates redundant computations, and employs efficient normalization techniques to enhance scalability without compromising predictive accuracy. While the Chebyshev Graph Convolutional Network (ChebNet) is a powerful spectral GNN variant, its high computational cost poses challenges. ACGNN addresses these limitations by improving computational efficiency while preserving interpretability and accuracy, making it well-suited for large-scale biological applications.

III. Materials and Methods

This section provides a detailed explanation of the data collection, graph construction, and embedding extraction processes, as well as the methodology that leverages Graph Neural Networks (GNNs) to compute high-dimensional embeddings.

A. Graph Convolution using Chebyshev Polynomials

Let $G = (V, E)$ be an undirected graph with adjacency matrix A and degree matrix D. The normalized Laplacian

matrix is defined as:

$$\mathcal{L} = I - D^{-1/2}AD^{-1/2} \tag{1}$$

Chebyshev graph convolutions approximate spectral filtering using Chebyshev polynomials:

$$X^{(l+1)} = \sum_{i=0}^{k} \theta_i T_i(\tilde{\mathcal{L}}) X^{(l)} \tag{2}$$

where:

- $X^{(l)}$ is the feature matrix at layer l,
- $T_i(\tilde{\mathcal{L}})$ are Chebyshev polynomials of the normalized Laplacian $\tilde{L}$,
- θ_i are trainable coefficients,
- k is the Chebyshev polynomial order.

B. Adaptive Receptive Fields

We introduce an adaptive mechanism where the polynomial order k is dynamically adjusted based on node importance:

$$k_v = \min(k_{max}, \alpha \cdot \log(\deg(v) + 1)) \tag{3}$$

where $\deg(v)$ is the degree of node v, and α is a tunable scaling factor.

C. Optimization and Training

To enhance computational efficiency, ACGNN incorporates several optimizations within its adaptive Chebyshev network. The Chebyshev order is reduced from $k = 3$ to $k = 2$, minimizing computational overhead while maintaining performance. Instead of LayerNorm, BatchNorm is employed to accelerate training. The architecture is further streamlined by reducing the number of ChebConv layers from three to two, lowering model complexity. Additionally, residual connections are introduced after the first layer, facilitating more efficient gradient flow and improving convergence speed. These modifications collectively enhance ACGNN's efficiency without compromising its predictive capabilities. Let $|E|$ be the number of edges and F the feature dimension, the computational complexity of ACGNN compared to standard ChebNet are shown in Table I.

Model	ChebConv Layers	Normalization	Complexity ($O(\cdot)$)
Standard ChebNet	3	LayerNorm	$3k\|E\|F^2$
Adaptive ChebNet	2	BatchNorm	$2k\|E\|F^2$

TABLE I – Comparison of Computational Cost

D. Data Collection and Preprocessing

In this study, we utilized the same PPI networks and multi-omics data as EMOGI and HGDC to evaluate our method's performance in identifying cancer driver genes. For completeness, we provide a brief overview of these datasets. The CPDB [39] is a specialized database that integrates protein-protein interaction (PPI) data with cancer-specific annotations [40]. It sources data from multiple experimental studies, such as The Cancer Genome Atlas (TCGA) [41]. STRING combines data from high-throughput experiments and curated interaction databases [42]. This resource includes experimental data from sources like the Human Protein Atlas [43], KEGG [44], Reactome pathways [45], [46], and Gene Ontology [47]. HIPPIE (Human Integrated Protein-Protein Interaction Environment) focuses on human-specific protein interactions [48]–[51]. It integrates high-quality experimental data and curated resources like the Molecular Interaction Database (MInt) [52] and BioGRID [53].

The PPI Network, obtained from the STRING database [54], limited to high-confidence interactions with scores above 0.85. The multi-omics data, sourced from the TCGA database, included gene mutations, DNA methylation, and gene expression profiles. Similar to EMOGI and MNGCL, we focused on cancer types with available data for gene mutations, gene expression, and DNA methylation in both tumor and normal tissues.

The three computed biological feature indices for genes are the gene mutation rate, the differential DNA methylation rate, and the differential gene expression rate. The gene mutation rate is calculated as the average of single nucleotide variations and copy number aberrations across all samples within a specific cancer type. The differential DNA methylation rate for each gene is calculated by averaging the methylation differences between cancer and normal samples. This averaging is performed across all cases for a given cancer type. For each gene i in cancer type c, a measure of dfferential DNA methylation (dm_{ci}) at its promoter can be defined as the difference in methylation signal between tumor (β_{ti}) and matched normal sample (β_{ni}), averaged across all available

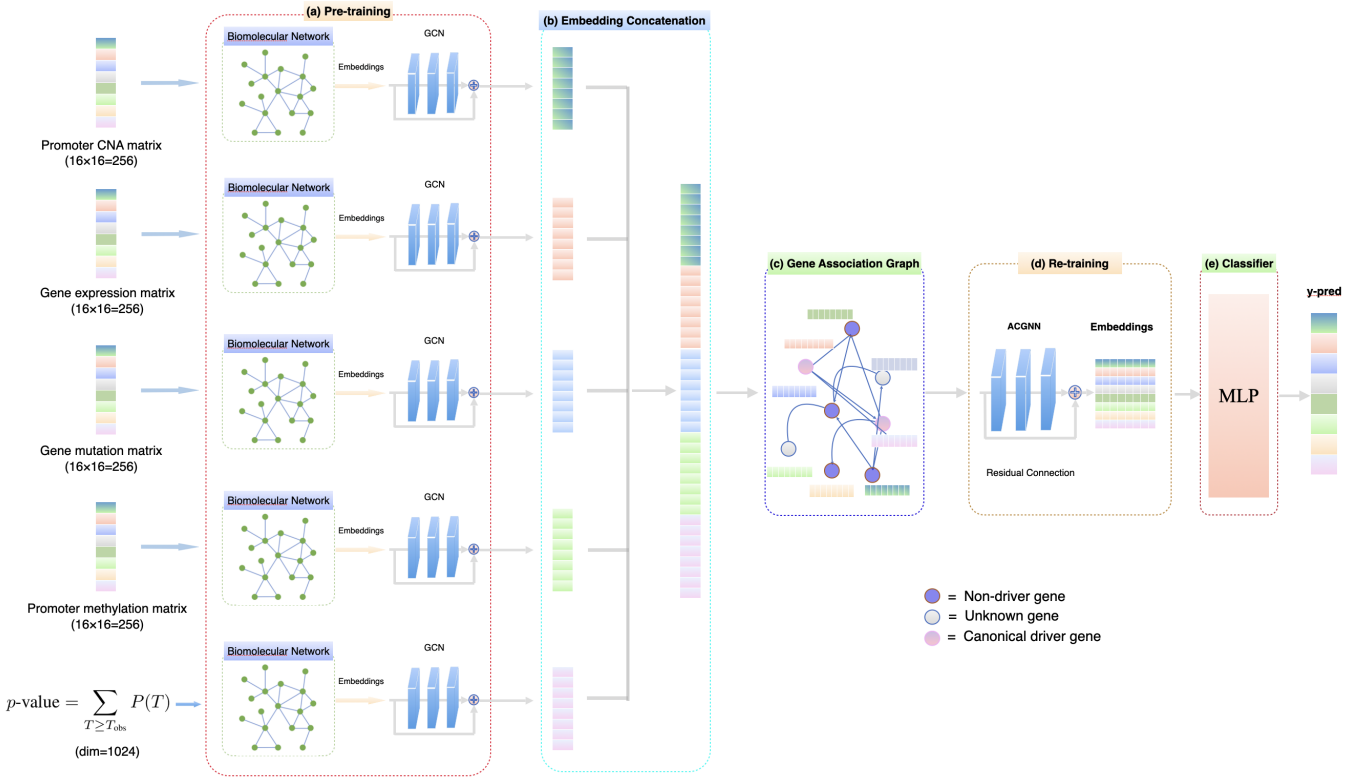


FIG. 1 – Overview of the ACGNN framework. (a) Multi-omics biomolecular networks are constructed for gene expression, mutation, and promoter DNA methylation data. Embeddings are extracted using GCN, and PPI network embeddings are integrated through Fisher's exact test. (b) Multi-omics embeddings are concatenated to form unified feature representations, combining topological, statistical, and multi-omics data. (c) A gene embedding graph is created, where nodes represent genes, and edges capture relationships between concatenated embeddings. (d) The ACGNN model refines node embeddings using an adaptive mechanism that dynamically tunes the polynomial order and residual connections, producing low-dimensional representations. (e) A binary classifier predicts driver and non-driver genes, outputting probabilities for each gene.

samples S_c for that c:

$$dm_{ci} = \frac{1}{|S_c|} \sum_{t \in S_c} (\beta_{ti} - \beta_{ni}), \tag{4}$$

The differential expression rate of each gene for a specific cancer type is calculated as the $\log_2$ fold change between expression in cancer and matched normal samples, averaged across all samples. The gene expression data, sourced from Wang et al. [55] was quantile-normalized and batch-corrected using the ComBat method [56].

E. Gene Enrichment Analysis

Gene enrichment analysis was conducted to identify significant associations among genes using Fisher's Exact Test [57]. The dataset, containing gene interactions, was analyzed to compute p-values for shared interaction partners between gene pairs. False discovery rates (FDR) were controlled using the Benjamini-Hochberg method [58], with adjusted p-values below 0.05 considered significant. Results included gene pairs, shared partners, p-values, and significance labels.

F. ACGNN

The Adaptive Chebyshev Graph Neural Network (ACGNN) framework integrates four types of multi-omics data: gene expression, gene mutation, promoter DNA methylation, and copy number alterations (CNA). For each data type, a corresponding biomolecular network $G_i = (V, E)$ is constructed, where V represents genes and E represents interactions. Node embeddings for each network are obtained using Chebyshev spectral filtering. The extracted embeddings capture essential molecular interactions within each data type (Figudetire 1(a)).

To incorporate additional topological information, PPI network embeddings are computed using Fisher's exact test on enrichment analysis based on gene sets, utilizing the p-value as a significance measure. These topological embeddings provide structural insights into gene interactions.

For each type of multi-omics data, 16 cancer-specific datasets are used to compute node embeddings of size 16, leading to a 256-dimensional feature vector per data type:

$$\mathbf{H}_i = \text{Concat}(\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{16}) \in \mathbb{R}^{256} \tag{5}$$

where $\mathbf{h}_j$ represents the embedding from the j-th dataset. The four vectors from different omics sources are concatenated to form a 1024-dimensional feature representation:

$$\mathbf{H}_{\text{multi-omics}} = \text{Concat}(\mathbf{H}_{\text{GE}}, \mathbf{H}_{\text{GM}}, \mathbf{H}_{\text{DM}}, \mathbf{H}_{\text{CNA}}) \in \mathbb{R}^{1024} \tag{6}$$

Finally, the integrated multi-omics representation is concatenated with the PPI network embeddings to incorporate additional structural information:

$$\mathbf{H}_{\text{final}} = \text{Concat}(\mathbf{H}_{\text{multi-omics}}, \mathbf{H}_{\text{topological}}) \in \mathbb{R}^{2048} \tag{7}$$

This final 2048-dimensional embedding serves as the comprehensive node representation, integrating molecular and structural information into the graph model (Figure 1(b)).

Using these embeddings, a gene embedding graph is constructed, where nodes represent genes and edges capture relationships between them (Figure 1(c)). The ChebNet model refines the node representations through residual connections, which preserve the original feature information while enabling deeper feature transformations (Figure 1(d)).

Finally, a binary classifier, implemented as a multi-layer perceptron (MLP) with a sigmoid activation function, predicts driver and non-driver genes (Figure 1(e)). The output is a probability score for each gene, representing its likelihood of being a cancer driver.

G. Model Training and Inference

We implement ACGNN using PyTorch and DGL. The network architecture consists of two Chebyshev convolution layers followed by a Multi-Layer Perceptron (MLP). First, bidirectional graphs are constructed for biomolecular networks, incorporating pretrained embeddings as node features. Node labels are assigned as follows: 1 for driver genes, 0 for non-driver genes, and -1 for unlabeled genes.

Algorithm 1 ACGNN Method

Input: Graph $G = (V, E)$, node features $\mathbf{H}^{(0)}$, rescaled Laplacian matrix $\tilde{L}$, Chebyshev order k, learnable weight matrices Θ_j, model parameters.

Output: Predicted output Y.

Initialize: Node features $\mathbf{H}^{(0)}$ and Laplacian matrix $\tilde{L}$.
$T_0(\tilde{L}) \leftarrow I$, $T_1(\tilde{L}) \leftarrow \tilde{L}$
for $j = 2$ to k **do** ▷ Compute Chebyshev polynomials
    $T_j(\tilde{L}) \leftarrow 2\tilde{L}T_{j-1}(\tilde{L}) - T_{j-2}(\tilde{L})$
end for
$\mathbf{H}^{(l+1)} \leftarrow \text{ReLU}\left(\sum_{j=0}^{k} \Theta_j T_j(\tilde{L})\mathbf{H}^{(l)}\right)$
$\mathbf{H}^{(l+1)} \leftarrow \text{BatchNorm}(\mathbf{H}^{(l+1)})$
$\mathbf{H}^{res} \leftarrow \mathbf{H}^{(l+1)}$ ▷ Store residual connection
$\mathbf{H}^{(l+2)} \leftarrow \text{ReLU}\left(\sum_{j=0}^{k} \Theta_j T_j(\tilde{L})\mathbf{H}^{(l+1)}\right)$
$\mathbf{H}^{(l+2)} \leftarrow \mathbf{H}^{(l+2)} + \mathbf{H}^{res}$ ▷ Add residual
$\mathbf{H}^{(l+2)} \leftarrow \text{Dropout}(\mathbf{H}^{(l+2)})$
$Y \leftarrow \text{MLP}(\mathbf{H}^{(l+2)})$
return Y

To define the training and testing sets, we apply the following masking strategy: the training mask includes all labeled nodes, while the testing mask allows all nodes to be evaluated. To ensure computational stability, self-loops are added to the graph, enabling nodes with no incoming edges to maintain connectivity. The model's feature dimensions are configured as follows: the input feature dimension is set to 128, the hidden feature dimension to 128, and the output dimension to 1 for binary classification.

We use focal loss function [59], given by Equation 8, to extend the standard Binary Cross-Entropy (BCE) loss to address class imbalance by reducing the loss contribution from well-classified examples and focusing more on hard-to-classify samples.

$$\begin{aligned} L_{\text{focal}} = &- \alpha\,(1-h)^{\gamma}\, y \log(h) \\ &- (1-\alpha) h^{\gamma} (1-y) \log(1-h). \end{aligned} \tag{8}$$

where y is the ground truth label, h is the predicted probability after applying the sigmoid activation function, α is a weighting factor that balances the positive and negative classes, and γ is a focusing parameter that adjusts the impact of easy versus hard examples. A higher value of γ reduces the relative loss contribution from well-classified samples, placing

greater emphasis on misclassified ones. The term $(1-h)^{\gamma}$ scales down the loss for correctly classified positive samples, whereas h^{γ} does the same for negative samples.

By incorporating these terms, focal loss mitigates the dominance of easily classified examples, ensuring that the model prioritizes learning from challenging cases. This is particularly useful in imbalanced datasets where the positive class, such as cancer driver genes, is significantly underrepresented compared to the negative class.

Node features and graph data are passed through the model to compute logits, which are then used to calculate the training loss based on the ground truth labels for nodes identified by the training mask. Predictions are generated for all nodes by applying a sigmoid activation to the computed logits, yielding the probabilities of each node belonging to the positive class. The pseudocode for executing our model is provided in Algorithm 1.

Method	Layers	Input	Hidden	Output	Params
ACGNN	5	2048	1024	1	7,346,177
HGDC	3	2048	1024	1	4,202,497
EMOGI	3	2048	1024	1	4,202,497
MTGCN	3	2048	1024	1	4,198,401
GCN	3	2048	1024	1	4,198,401
GAT	3	2048	1024	1	4,202,497
GraphSAGE	3	2048	1024	1	7,344,129
GIN	3	2048	1024	1	6,297,601
Chebnet	3	2048	1024	1	10,489,857

TABLE II – Details of model architectures across different methods.

IV. Experiments

A. Baseline Methods

To evaluate the performance of ACGNN in identifying cancer driver genes, we compared it against eight other methods. These included three state-of-the-art GCN-based approaches: EMOGI, MTGCN, and HGDC, which leverage multi-omics data as gene features and incorporate PPI networks to learn gene representations for cancer driver gene prediction. Additionally, we assessed five widely-used GNN methods: GCN [60], GIN [61], ChebNet [62], GraphSAGE [63], and GAT [64], which are traditional graph neural network models that aggregate features from neighboring nodes and themselves in different manners to learn new representations. All methods, including ACGNN, were tested on three distinct PPI networks, using the same network structure and gene feature matrix $X \in \mathbb{R}^{N\times 2048}$ as inputs. Here, N denotes the number of genes, and the pretrained embeddings are of dimension 2048, comprising four concatenated 256-dimensional biofeature vectors $(256+256+256+256)$ and a topological feature of dimension 1024. Details of the model architecture are provided in Table II.

B. Experiment Settings

Our algorithm is implemented in Python 3.9, utilizing PyTorch 2.0.1 and PyTorch Geometric 2.3.1, within an environment that also includes DGL. The Adam optimizer is employed, and we use FocalLoss ($\alpha = 0.25, \gamma = 2$) as the loss function. The learning rate is set to 0.001, with the number of training epochs fixed at 200 throughout all experiments, unless stated otherwise. All experiments were conducted on an Ubuntu server featuring an Intel CPU (2.4GHz, 128GB RAM) and an Nvidia RTX 4080 GPU. The time consumption and CPU/GPU usage for different methods are detailed in Table III.

C. Gene Embeddings Clustering

The initial task performed using the obtained embeddings involved clustering the nodes within the gene association graph. The process began with dataset preparation and gene mapping, followed by configuring the model with specific parameters, such as the number of layers and embedding dimensions. During training, multiple epochs were executed, where each batch was processed by computing logits, calculating the loss using a binary cross-entropy function with class imbalance weighting, and updating model parameters through backpropagation using the Adam optimizer. After each epoch, validation was performed to evaluate model performance, and the best model was selected based on the validation loss. Using the optimal model, the final node embeddings were generated, and cluster labels were assigned accordingly. These embeddings were further analyzed to extract cluster-specific insights, including the identification of significant nodes within each cluster. The final model state was saved to enable future use, providing a structured representation of the gene network through clustered node embeddings. Figure 2 presents the hierarchical clustering of genes based on pretrained node

Model	CPDB		STRING		HIPPIE	
	CPU/GPU Usage	Time/Epoch	CPU/GPU Usage	Time/Epoch	CPU/GPU Usage	Time/Epoch
ACGNN	3.17M/87.51M	0.0426s	3.78M/87.04M	0.0328s	3.14M/87.51M	0.0227s
HGDC	3.28M/55.75M	0.0573s	3.13M/55.29M	0.0227s	3.73M/55.75M	0.0358s
EMOGI	3.23M/55.75M	0.0573s	3.08M/55.29M	0.0348s	3.61M/55.75M	0.0227s
MTGCN	3.22M/55.72M	0.0497s	3.12M/55.25M	0.0301s	3.25M/55.72M	0.0490s
GCN	3.23M/55.72M	0.0470s	3.11M/55.25M	0.0329s	3.33M/55.72M	0.0429s
GAT	3.23M/55.75M	0.0880s	3.11M/55.29M	0.1417s	3.27M/55.75M	0.0626s
GrpahSAGE	3.17M/87.49M	0.0532s	3.25M/87.02M	0.0210s	3.18M/87.49M	0.0330s
GIN	3.17M/87.50M	0.0528s	3.26M/87.04M	0.0213s	3.20M/87.50M	0.0378s
ChebNet	3.22M/119.32M	0.0622s	3.13M/118.79M	0.0241s	3.16M/119.32M	0.0453s

TABLE III – Comparison of models in CPU/GPU usage and time cost per epoch on the different PPI networks: CPDB, STRING, and HIPPIE.

Method	CPDB		STRING		HIPPIE	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
ACGNN	0.9652	0.9783	0.9578	0.9738	0.9297	0.9597
HGDC	0.6776	0.7288	0.7133	0.7740	0.6525	0.7634
EMOGI	0.6735	0.7230	0.81846	0.8737	0.6672	0.7960
MTGCN	0.6862	0.7712	0.7130	0.7878	0.6762	0.7785
GCN	0.6915	0.7730	0.6688	0.7681	0.6708	0.7675
GAT	0.6670	0.7086	0.8166	0.8791	0.6478	0.7496
GraphSAGE	0.6664	0.7522	0.6166	0.7182	0.6571	0.7624
GIN	0.5836	0.6405	0.5173	0.5918	0.5844	0.6791
Chebnet	0.8017	0.8622	0.8777	0.9159	0.7409	0.8443

TABLE IV – Performance comparison of different methods across three biological networks: CPDB (pathway-based network), STRING (protein-protein interaction network), and HIPPIE (high-confidence protein interaction network).

	CPDB		STRING		HIPPIE	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
ACGNN	0.9652	0.9783	0.9578	0.9738	0.9297	0.9597
AC256x1	0.9315	0.9527	0.9054	0.9464	0.8551	0.9129
AC256x2	0.9382	0.9610	0.9413	0.9630	0.9112	0.9378
AC256x4	0.9495	0.9657	0.9489	0.9634	0.8934	0.9388
AC1024x1	0.8481	0.9045	0.8653	0.9182	0.9288	0.9504

TABLE V – The ablation experimental results of the different networks.

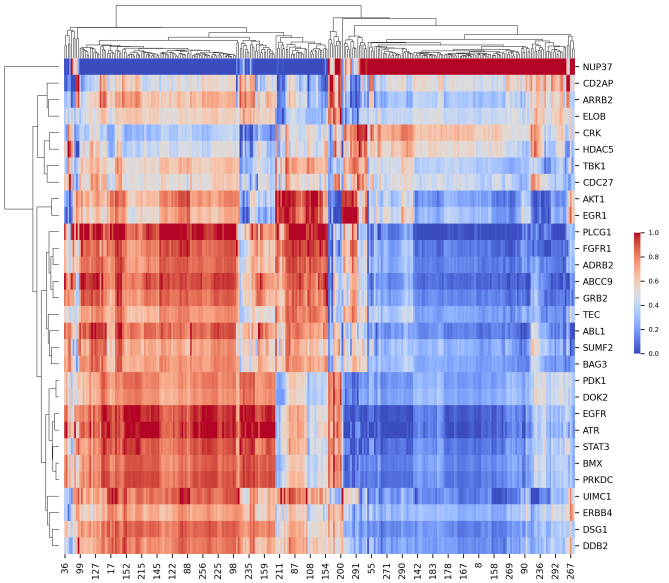
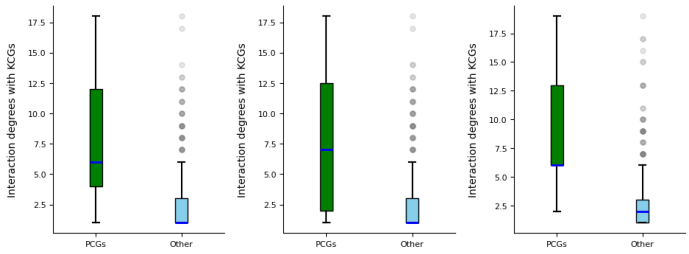


FIG. 2 – Demonstration of pretrained node embeddings with a dimension of 300 for clustering genes into k=30 clusters.

embeddings of dimension 300. The heatmap visualizes gene expression patterns, with red and blue indicating high and low expression levels, respectively. The hierarchical clustering dendrograms at the top and left of the heatmap reveal distinct gene clusters, highlighting structural similarities among genes. The presence of highly correlated gene groups suggests that the embedding model effectively captures biologically meaningful relationships. Notably, genes such as *NUP37*, *CD2AP*, and *AKT1* form distinct subclusters, implying potential functional similarity or co-regulation. In contrast, genes like *EGFR*, *ATR*, and *STAT3* are grouped in separate regions, potentially reflecting their involvement in distinct biological pathways. The structure of the heatmap demonstrates that the pretrained embeddings provide an informative representation of gene features, facilitating downstream analyses such as functional enrichment and pathway analysis.

D. *Performance on Driver Gene Prediction*

We evaluated our method and baseline approaches for predicting driver genes on pan-cancer multi-omics dataset, using the average AUROC and AUPRC from ten iterations of validation as performance metrics. Table II presents the architectural details of different models used in our study.



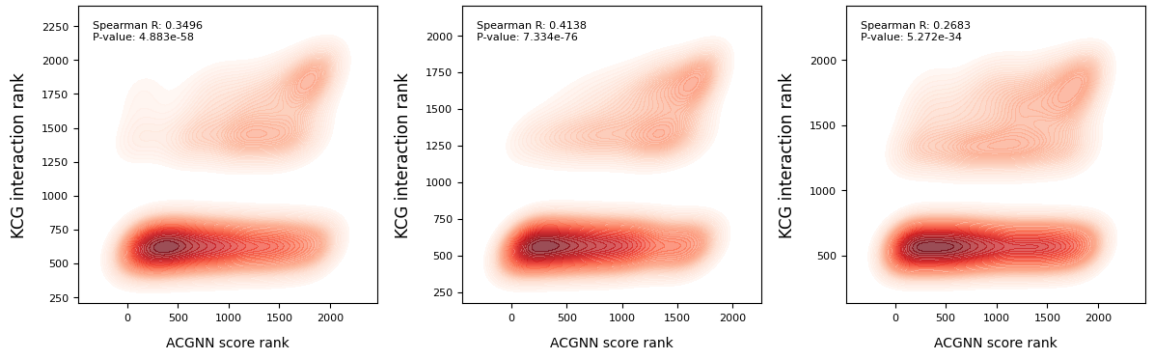
(a) CPDB (b) STRING (c) HIPPIE

FIG. 3 – The interaction distributions of Predicted Cancer Genes (PCGs) with Known Cancer Genes (KCGs) across three different protein-protein interaction (PPI) networks: CPDB, STRING, and HIPPIE.

ACGNN stands out with the highest parameter count of 7,346,177, attributed to its deeper architecture with five layers. In contrast, most baseline models, including HGDC, EMOGI, MTGCN, GCN, GAT, and GraphSAGE, employ a three-layer structure with a comparable number of parameters, ranging from approximately 4.19M to 4.20M. Notably, GraphSAGE has a higher parameter count of 7,344,129, indicating increased complexity despite having the same number of layers as the other baselines. GIN follows with 6,297,601 parameters, showing a moderate increase in complexity. ChebNet exhibits the highest parameter count of 10,489,857, highlighting its expressive power but also potential computational overhead. These architectural differences contribute to variations in model performance, as analyzed in the subsequent sections. Figure 6 presents the performance of various graph-based models on different biological networks. The x-axis represents the AUROC values, which indicate the classification ability of each model, while the y-axis represents the AUPRC values, highlighting precision-recall performance. From the plot, ACGNN achieves the highest AUROC and AUPRC, demonstrating strong predictive performance. HGDC, EMOGI, and MTGCN show competitive results, clustering around an AUROC of 0.75–0.85. The choice of biomolecular networks (CPDB, STRING, HIPPIE) influences model performance, as reflected in the varied shapes. GCN, GAT, and GraphSAGE exhibit moderate results, suggesting their dependency on network topology. Table IV presents the average performance of ACGNN and eight other methods across different biomolecular networks. ACGNN consistently achieves the highest performance, particularly excelling in CPDB and STRING. ChebNet also performs strongly, especially in STRING and HIPPIE. In contrast, GIN and GraphSAGE tend to underperform across multiple datasets, highlighting their limitations in certain network structures.

Figure 7 compares the AUPRC performance of various methods across different biological networks using independent test sets from OncoKB and ONGene. The x-axis represents the AUPRC for ONGene, while the y-axis represents the AUPRC for OncoKB. ACGNN continues to outperform other models, achieving the highest scores across both test sets. HGDC, EMOGI, and MTGCN display competitive results, clustering between 0.75 and 0.85 AUPRC. The distribution of models across CPDB, STRING, and HIPPIE networks indicates variations in predictive performance due to differences in biological network structures. The results demonstrate that methods leveraging advanced GNN architectures, particularly ACGNN, provide superior predictions in cancer driver gene identification compared to traditional models.

Figure 3 presents the interaction distributions of Predicted Cancer Genes (PCGs) with Known Cancer Genes (KCGs) across three different protein-protein interaction (PPI) networks: CPDB, STRING, and HIPPIE. The boxplots illustrate the interaction degrees between these gene categories, offering insights into their connectivity patterns. In the CPDB network (Figure 3 a), PCGs exhibit a wider distribution of interaction degrees, with a higher median and several outliers. This suggests that PCGs in CPDB tend to have stronger connectivity with KCGs. The STRING network (Figure 3 b) follows a similar pattern, though with slightly lower median values and fewer extreme outliers, indicating a moderate connectivity between PCGs and KCGs. The HIPPIE network (Figure 3 c) displays the highest median interaction degree among the three networks, suggesting that it might capture a denser set of biologically relevant interactions for PCGs. Across all three networks, PCGs demonstrate significantly higher interaction degrees with KCGs compared to other genes. The interquartile range (IQR) and median values for PCGs are consistently larger, while the presence of multiple outliers suggests that some PCGs have exceptionally high connectivity with KCGs. In contrast, the Other gene category has a much lower and more compact distribution, indicating weaker associations



(a) CPDB (b) STRING (c) HIPPIE

FIG. 4 – Kernel Density Estimation (KDE) plot illustrating the distribution of interaction degrees for predicted driver genes and other genes. The x-axis represents the inteaction degree, and the y-axis represents the density.

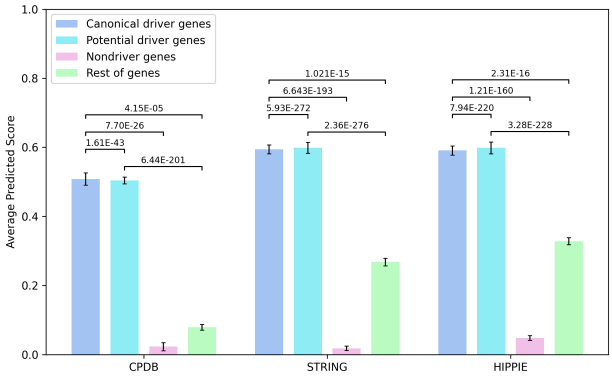


FIG. 5 – Average predicted scores for different networks and gene types. The bar plot displays the scaled predicted scores for canonical driver genes, potential driver genes, nondriver genes, and the rest of the genes across three biological networks: CPDB, STRING, and HIPPIE. Error bars represent the scaled standard errors of the mean (SEM). The statistical significance (p-values) for comparisons between gene categories within each network is annotated above the bars. Numbers above the brackets represent the p-values of the Spearman Rank Correlation Test, indicating the significance of differences between groups.

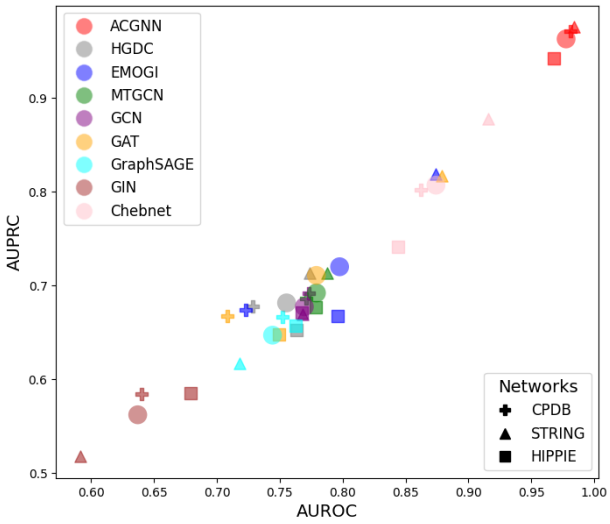


FIG. 6 – Performance comparison of graph-based models across different biological networks. The scatter plot shows the area under the receiver operating characteristic curve (AUROC) on the x-axis and the area under the precision-recall curve (AUPRC) on the y-axis for various models. The large circles represent the mean AUC values across biomolecular networks for each method.

with KCGs. These results reinforce the relevance of PCGs in cancer research, as their strong connectivity with known cancer genes suggests potential biological significance. The differences observed across the three networks highlight the impact of network-specific interaction curation, with HIPPIE capturing the densest set of interactions.

Kernel Density Estimation (KDE) (Figure 4) presents the average predicted scores for different gene types across the three biological networks. The bar plot categorizes genes into canonical driver genes, potential driver genes, nondriver genes, and the rest of the genes. The error bars represent the scaled standard errors of the mean (SEM), ensuring statistical robustness in the displayed values. Across all three networks, canonical driver genes consistently exhibit the highest predicted scores, followed by potential driver genes, while nondriver genes and the rest of the genes have significantly lower scores. This pattern suggests a strong alignment between the model's predictions and known biological classifications of driver genes. The statistical significance of the differences in predicted scores is annotated above the bars, with p-values from the Spearman Rank Correlation Test. The highly

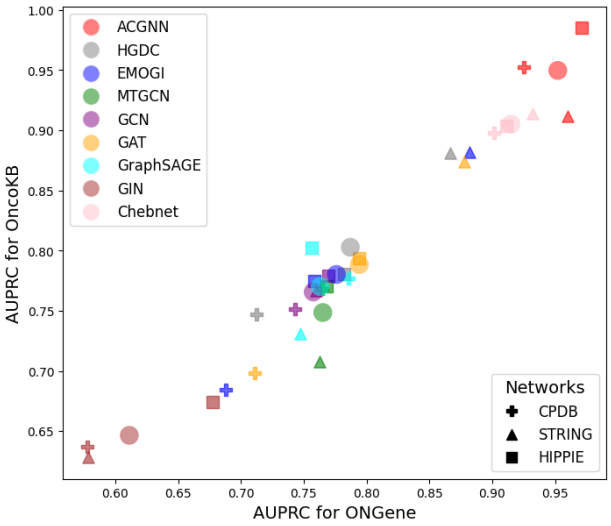


FIG. 7 – Performance comparisons of different methods on biomolecular networks. Results of all methods were evaluated on two independent test sets derived from the OncoKB and ONGene databases, respectively. The large circles represent the mean AUPRC values across biomolecular networks for each method.

significant p-values (e.g., $p < 10^{-40}$ in CPDB and similar magnitudes in STRING and HIPPIE) indicate strong differentiation between gene categories, reinforcing the predictive power of the model in distinguishing cancer-related genes. Additionally, STRING and HIPPIE networks exhibit a greater separation between gene categories than CPDB, suggesting that these networks may capture more biologically relevant interactions. The increased predicted scores for the "rest of the genes" category in HIPPIE compared to CPDB and STRING could indicate a broader connectivity pattern in HIPPIE's network structure. These results highlight the model's ability to predict meaningful biological classifications across different interaction networks, with STRING and HIPPIE demonstrating particularly strong discriminatory power.

E. Performance on OncoKB and ONGene

To investigate whether the models' performance is biased toward a specific dataset, we evaluated them on two additional cancer gene databases: OncoKB [65] and ONGene [66]. The cancer gene sets from OncoKB and ONGene are compiled from scientific literature or clinical studies and are therefore not explicitly informed by any of the data types used to train ACGNN . This demonstrates ACGNN 's capability to predict cancer genes broadly, independent of the methods or data used to define them, across the Gene Network, Pathway Network, and Protein-Protein Interaction Network. For this evaluation, the test set was masked exclusively for non-labeled genes.

The performance comparisons of different methods across biomolecular networks are illustrated in Figure 7. A clear positive correlation between AUPRC values for the ONGene and OncoKB datasets is observed, indicating that methods performing well on one dataset tend to perform well on the other. This highlights the robustness and generalizability of certain methods across diverse, independent cancer gene sets. Among the evaluated methods, ACGNN achieves the highest performance on both datasets, as reflected by its red markers clustered near the top-right corner of the plot. Furthermore, ACGNN maintains consistently high average AUPRC values, underscoring its reliability in network-based cancer gene prediction tasks. These results emphasize the critical importance of integrating multiple biomolecular networks and leveraging advanced graph-based models for accurately identifying cancer driver genes. The alignment between the results from the OncoKB and ONGene datasets further validates the relevance and applicability of these methods in advancing cancer research.

F. Feature Ablation Experiment

The feature ablation experiments provide insights into the influence of different multi-omics data and biomolecular networks on the performance of ACGNN, as shown in Table V. The table presents the ablation study results for different network variants trained on three biological networks: CPDB, STRING, and HIPPIE. The models tested include ACGNN, AC256x1, AC256x2, AC256x4, and AC1024x1. ACGNN, the full model incorporating all bio-features and topological information, achieves the highest AUROC and AUPRC across all networks, demonstrating the advantage of integrating comprehensive features. AC256x1, using only one type of biological feature, and AC256x2, utilizing two types, show a decline in performance, indicating that limited biological features reduce predictive power. AC256x4, incorporating four types of biological features, performs better than AC256x2, suggesting that more bio-features enhance model performance. AC1024x1, which relies solely on topological information, has the lowest AUROC and AUPRC scores across most networks, confirming the critical role of biological features.

These results highlight that incorporating multiple biological features significantly improves predictive accuracy, with the best performance achieved when combining all bio-features with topological information in ACGNN.

G. Evaluation of Cancer Type-Specific Driver Gene Prediction

We also investigate the effectiveness of ACGNN in detecting driver genes of a single cancer type. The results in Table VII presents the AUROC (Area Under the Receiver Operating Characteristic Curve) and AUPRC (Area Under the Precision-Recall Curve) scores for various cancer types using four different methods: HGDC, EMOGI, MTGCN, and ACGNN. These metrics assess the predictive performance of each model across three feature types: biological (Bio), topological (Topo), and combined (Comb). ACGNN consistently outperforms the other methods across all cancer types. The bolded values in the table indicate that ACGNN achieves the highest AUROC and AUPRC scores in nearly all cases, highlighting its effectiveness in predicting miRNA-disease associations compared to HGDC, EMOGI, and MTGCN. For instance, in bladder cancer (BLCA), ACGNN with combined features achieves an AUROC of 0.9644 and an AUPRC of 0.9766, significantly surpassing the next best method, MTGCN (Comb), which records an AUROC of 0.7525 and an AUPRC of 0.8174.

The integration of biological and topological features (Comb) consistently yields the best performance across all cancer types. In most cases, the AUROC and AUPRC values for the combined feature type are the highest, indicating that incorporating both biological and topological information enhances predictive accuracy. For example, in liver cancer (LIHC), ACGNN with combined features achieves an AUROC of 0.9704 and an AUPRC of 0.9809, outperforming the biological-only (AUROC = 0.9073, AUPRC = 0.9391) and topological-only (AUROC = 0.8513, AUPRC = 0.9022) configurations. Certain methods exhibit limitations with specific cancer types. HGDC consistently performs the worst, frequently yielding the lowest AUROC and AUPRC values. While EMOGI and MTGCN produce competitive results in some cases, they are generally outperformed by ACGNN. For instance, in cervical cancer (CESC), HGDC with topological features achieves an AUROC of 0.6635 and an AUPRC of 0.7085, considerably lower than ACGNN with combined features, which attains an AUROC of 0.9760 and an AUPRC of 0.9871. The superior performance of ACGNN across multiple cancer types suggests strong generalizability.

H. Prediction of Novel Cancer Driver Genes

In this section, we investigate the performance of ACGNN in identifying novel cancer genes across three biomolecular networks: CPDB, STRING, and HIPPIE. The CPDB network consists of 5,693 nodes and 31,761 edges, while STRING contains 10,430 nodes and 58,559 edges. HIPPIE, the largest of the three, comprises 12,981 nodes and 84,969 edges. These networks provide diverse interaction patterns, enabling a comprehensive evaluation of ACGNN's predictive capabilities. To evaluate the performance of ACGNN across different biomolecular networks, we computed the predicted scores for various gene categories, including canonical driver genes (CDGs), nondriver genes, potential cancer genes curated in the NCG 6.0 database, and other genes. The analysis was conducted across three biomolecular networks: CPDB, STRING, and HIPPIE. The results provide insights into how ACGNN distinguishes between different gene types within these networks, highlighting its capability to identify potential cancer-associated genes with high confidence.

Figure 5 illustrates the average predicted scores of different gene categories (canonical driver genes, potential driver genes, nondriver genes, and the rest of genes) across three biomolecular networks: CPDB, STRING, and HIPPIE. The results indicate that canonical and potential driver genes consistently achieve significantly higher predicted scores compared to nondriver genes and the rest of the genes, suggesting strong model confidence in identifying cancer-associated genes. The statistical significance between different gene categories is denoted by p-values, all of which are extremely small, confirming the robustness of the model's predictions. Notably, the predicted scores of canonical and potential driver genes exhibit minimal variance, further reinforcing the model's reliability across diverse biomolecular networks. The STRING and HIPPIE networks show slightly higher predicted scores for driver genes compared to CPDB, indicating that these networks might provide more informative gene interactions for the model.

To identify novel cancer driver genes, we trained the model using the CPDB network on the full training dataset with 500 epochs. Predictions were filtered by applying a threshold of 0.99, ensuring that only potential cancer driver genes were included while excluding previously labeled genes. This process resulted in 352 predicted cancer driver genes, of which 55 are already known cancer drivers. Additionally, 204 of the novel predictions have at least one supporting piece of evidence suggesting their potential as cancer drivers, based on sources such as NCG's candidate cancer genes [70], OncoKB's manually curated cancer genes, ONGene's literature-curated cancer genes, and IntOGen's catalog of cancer-related mutations [71]. Among the top 60 predicted novel driver genes, 71.67% (43 out of 60) have at least one supporting source indicating their potential role in cancer (Table VI).

V. Discussion and Conclusions

Many genes exhibit context-dependent functions, being recurrently mutated in some cancers while undergoing epigenetic alterations in others [72]–[74]. To gain a comprehensive understanding of the cancer landscape, we emphasize the importance of integrating diverse molecular data types into a unified model for accurately predicting cancer genes with distinct characteristics. ACGNN provides a scalable and adaptive framework for cancer gene prediction by leveraging Chebyshev polynomials, attention-based receptive fields, and pan-cancer multi-omics data integration. We evaluated the performance of ACGNN across multiple cancer gene sets and PPI networks, comparing it with other established methods. While no single method consistently outperformed the others across all scenarios, ACGNN trained on pan-cancer data demonstrated, on average, superior AUPRC values.

While ACGNN demonstrated stable performance across various cancer gene datasets, the performance of cancer driver gene prediction was highly dependent on the dataset used. This variability likely reflects the differing biases inherent in the collection of these cancer gene sets. For instance, the PPIs from STRING includes functional associations (e.g., co-expression or shared pathways) that do not necessarily indicate direct physical interactions between proteins.

The ACGNN framework is highly versatile, capable of integrating various types of omics data and networks beyond those utilized here. This flexibility allows its application beyond the field of cancer genomics, enabling the study of other complex diseases where multi-omics data are available, and functional gene connections play a critical role in classifying disease-related genes. Our future work will explore hybrid models combining ChebNet with Transformers.

Acknowledgements

This work is supported by the National Science Foundation under Grant No. 2245805.

References

[1] N. D. Dees *et al.*, "MuSiC: Identifying Mutational Significance in Cancer Genomes," *Genome Research*, vol. 22, pp. 1589–1598, 2012. doi: 10.1101/gr.134635.111.

[2] B. Vogelstein *et al.*, "Cancer Genome Landscapes," *Science*, vol. 339, pp. 1546–1558, 2013. doi: 10.1126/science.1235122.

[3] M. D. M. Leiserson, F. Vandin, H.-T. Wu, *et al.*, "Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations Across Pathways and Protein Complexes," *Nature Genetics*, vol. 47, pp. 106–114, 2015. doi: 10.1038/ng.3168.

[4] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[5] A. Bashashati *et al.*, "DriverNet: Uncovering the impact of somatic driver mutations on transcriptional networks in cancer," *Genome Biology*, vol. 13, no. 12, p. R124, 2012. doi: 10.1186/gb-2012-13-12-r124.

[6] M. S. Lawrence *et al.*, "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes," *Nature*, vol. 499, pp. 214–218, 2013. doi: 10.1038/nature12213.

[7] L. B. Alexandrov *et al.*, "Signatures of Mutational Processes in Human Cancer," *Nature*, vol. 500, pp. 415–421, 2013. doi: 10.1038/nature12477.

[8] J. P. Hou and J. Ma, "DawnRank: Discovering Personalized Driver Genes in Cancer," *Genome Medicine*, vol. 6, no. 7, p. 56, 2014. doi: 10.1186/s13073-014-0056-8.

[9] D. Repana *et al.*, "The Network of Cancer Genes (NCG): A Comprehensive Catalogue of Known and Candidate Cancer Genes from Cancer Sequencing Screens," *Genome Biology*, vol. 20, pp. 1–12, 2019. doi: 10.1186/s13059-019-1760-7.

[10] Z. Sondka *et al.*, "The COSMIC Cancer Gene Census: Describing Genetic Dysfunction Across All Human Cancers," *Nature Reviews Cancer*, vol. 18, pp. 696–705, 2018. doi: 10.1038/s41568-018-0060-1.

[11] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, "OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes," *Bioinformatics*, vol. 29, no. 18, pp. 2238–2244, Sep. 2013.

[12] R. Gillman, M. A. Field, U. Schmitz, R. Karamatic, and L. Hebbard, "Identifying cancer driver genes in individual tumours," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 5028–5038, 2023.

Gene	Confirmed Sources	Gene	Confirmed Sources	Gene	Confirmed Sources
ACTB	OncoKB, NCG, IntOGen	VDAC1		SRP68	IntOGen
CXCL2	OnGene	BMP1	IntOGen	CHMP6	
TRAF2	OncoKB, NCG, IntOGen	CDK2		DDB1	IntOGen
RBBP4	IntOGen	CRKL	OncoKB, OnGene	FHL2	OnGene
UBC	IntOGen	KAT2B		HCK	IntOGen
LIN37		ATR	OncoKB, NCG, IntOGen	ZWINT	NCG
FREM2	NCG, IntOGen	CHRD	IntOGen	ACTN2	IntOGen
LEF1	OncoKB, OnGene, NCG, IntOGen	SIAE		DAXX	OncoKB, OnGene, NCG, IntOGen
PEX5		UBXN7	NCG	TEAD4	IntOGen
ORC5		TPM3	OncoKB, NCG	RBM39	NCG, IntOGen
NFKB1	IntOGen	ABHD5		HDAC1	OncoKB, OnGene, NCG
HDAC2	OncoKB, NCG	BIRC2	OnGene	GRB2	NCG, IntOGen
FOXK1	NCG	GAB1	OncoKB	EGFR	OncoKB, OnGene, NCG, IntOGen
PURA		TPM1		KAT5	
HDAC3	IntOGen	RBL1	IntOGen	ATF5	IntOGen
PRKDC	OncoKB, NCG, IntOGen	CDK9		MYOM1	IntOGen
NUP37		TBP	NCG, IntOGen	ERCC6	NCG, IntOGen
CDC20	NCG	TBCA		GNA13	OncoKB, OnGene, NCG, IntOGen
CASP7		PLK1	OnGene, NCG, IntOGen	SF3B3	NCG
TRAF6	OnGene, NCG, IntOGen	HDHD3	IntOGen	UBE2B	

TABLE VI – The top 60 predicted driver genes along with their confirmed sources, with unconfirmed genes left blank. Among these, 71.67% of the novel genes have at least one supporting piece of evidence indicating their potential as cancer drivers. In total, 43 out of 60 genes are confirmed.

[13] J. Song, W. Peng, F. Wang, and J. Wang, "Identifying driver genes involving gene dysregulated expression, tissue-specific expression and gene-gene network," *BMC Med. Genomics*, vol. 12, no. S7, p. 168, Dec. 2019.

[14] W. Peng, S. Yi, W. Dai, and J. Wang, "Identifying and ranking potential cancer drivers using representation learning on attributed network," *Methods*, vol. 192, pp. 13–24, Aug. 2021.

[15] W. Peng, Q. Tang, W. Dai, and T. Chen, "Improving cancer driver gene identification using multi-task learning on graph convolutional network," *Briefings Bioinf.*, vol. 23, no. 1, Jan. 2022, Art. no. bbab432.

[16] J. Song, W. Peng, and F. Wang, "An entropy-based method for identifying mutual exclusive driver genes in cancer," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 3, pp. 758–768, May 2020.

[17] S. W. Zhang, J. Y. Xu, and T. Zhang, "DGMP: Identifying cancer driver genes by jointing DGCN and MLP from multi-omics genomic data," *Genomics Proteomics Bioinformatics*, vol. 20, no. 5, pp. 928–938, Oct. 2022, doi: 10.1016/j.gpb.2022.11.004. Epub Dec. 1, 2022. PMID: 36464123; PMCID: PMC10025764.

[18] F. Cheng, J. Zhao, and Z. Zhao, "Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes," *Briefings Bioinf.*, vol. 17, no. 4, pp. 642–656, Jul. 2016.

[19] O. Collier, V. Stoven, and J.-P. Vert, "LOTUS: a single- and multitask machine learning algorithm for the prediction of cancer driver genes," *PLoS Computational Biology*, vol. 15, 2019, Article e1007381. doi: 10.1371/journal.pcbi.1007381.

[20] H.-C. Yi, Z.-H. You, D.-S. Huang, *et al.*, "Graph representation learning in bioinformatics: trends, methods and applications," *Briefings in Bioinformatics*, vol. 23, 2021, Article bbab340. doi: 10.1093/bib/bbab340.

[21] T. P. Mourikis, L. Benedetti, E. Foxall, *et al.*, "Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma," *Nature Communications*, vol. 10, p. 3101, 2019. doi: 10.1038/s41467-019-11069-0.

[22] J. Nulsen, H. Misetic, C. Yau, *et al.*, "Pan-cancer detection of driver genes at the single-patient resolution," *Genome Medicine*, vol. 13, p. 12, 2021. doi: 10.1186/s13073-021-00820-7.

[23] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61-80, Jan. 2009. doi: 10.1109/TNN.2008.2005605.

[24] X. Liu and M. Yang, "Research on conversational machine reading comprehension based on dynamic graph neural network," *Journal of Integrated Technology*, vol. 11, no. 2, pp. 67–78, 2022.

[25] W. Peng, T. Chen, and W. Dai, "Predicting drug response based on multi-omics fusion and graph convolution," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1384–1393, Mar. 2022.

[26] W. Peng, H. Liu, W. Dai, N. Yu, and J. Wang, "Predicting cancer drug response using parallel heterogeneous graph convolutional networks with neighborhood interactions," *Bioinformatics*, vol. 38, no. 19, pp. 4546–4553, Sep. 2022.

[27] A.-I. Albu, "An approach for predicting protein-protein interactions using supervised autoencoders," *Procedia Computer Science*, vol. 207, pp. 2023–2032, 2022, doi: 10.1016/j.procs.2022.09.261.

[28] M. Chatzianastasis, M. Vazirgiannis, and Z. Zhang, "Explainable Mul-

tilayer Graph Neural Network for Cancer Gene Prediction," *Bioinformatics*, Oct. 2023. doi: 10.1093/bioinformatics/btad643. Available: https://doi.org/10.1093/bioinformatics/btad643.

[29] B. Li and S. A. Nabavi, "A Multimodal Graph Neural Network Framework for Cancer Molecular Subtype Classification," *BMC Bioinformatics*, vol. 25, p. 27, 2024. Available: https://doi.org/10.1186/s12859-023-05622-4.

[30] H. Li, Z. Han, Y. Sun, F. Wang, P. Hu, C. Ren, X. Xu, H. Chen, Y. Yang, and X. Bo, "CGMega: Explainable Graph Neural Network Framework with Attention Mechanisms for Cancer Gene Module Dissection," 2023. doi: 10.21203/rs.3.rs-3180743/v1. Available: https://doi.org/10.21203/rs.3.rs-3180743/v1.

[31] W. Peng, Q. Tang, W. Dai, and T. Chen, "Improving Cancer Driver Gene Identification Using Multi-task Learning on Graph Convolutional Network," *Briefings in Bioinformatics*, vol. 23, no. 1, Article bbab432, Jan. 2022. Available: https://doi.org/10.1093/bib/bbab432.

[32] W. Peng, Z. Zhou, W. Dai, N. Yu, and J. Wang, "Multi-Network Graph Contrastive Learning for Cancer Driver Gene Identification," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 4, pp. 3430-3440, Jul.-Aug. 2024. doi: 10.1109/TNSE.2024.3373652.

[33] R. Schulte-Sasse, S. Budach, D. Hnisz, and et al., "Integration of Multiomics Data with Graph Convolutional Networks to Identify New Cancer Genes and Their Associated Molecular Mechanisms," *Nature Machine Intelligence*, vol. 3, pp. 513–526, 2021. Available: https://doi.org/10.1038/s42256-021-00325-y.

[34] T. Zhang, S.-W. Zhang, M.-Y. Xie, and Y. Li, "A Novel Heterophilic Graph Diffusion Convolutional Network for Identifying Cancer Driver Genes," *Briefings in Bioinformatics*, vol. 24, no. 3, May 2023, bbad137. Available: https://doi.org/10.1093/bib/bbad137.

[35] I. G. Iván and V. Grolmusz, "When the Web Meets the Cell: Using Personalized PageRank for Analyzing Protein Interaction Networks," *Bioinformatics*, vol. 27, no. 3, pp. 405–407, 2011. doi: 10.1093/bioinformatics/btq671.

[36] W. Zhao, X. Gu, S. Chen, J. Wu, and Z. Zhou, "MODIG: Integrating Multi-Omics and Multi-Dimensional Gene Network for Cancer Driver Gene Identification Based on Graph Attention Network Model," *Bioinformatics*, vol. 38, no. 21, pp. 4901-4907, Oct. 2022. doi: 10.1093/bioinformatics/btac622.

[37] J. Du, S. Zhang, G. Wu, J. M. F. Moura, and S. Kar, "Topology adaptive graph convolutional networks," *arXiv preprint*, vol. 1710, no. 10370, 2018. [Online]. Available: https://arxiv.org/abs/1710.10370.

[38] S. Singh, A. Sharma, and V. K. Chauhan, "GTAGCN: Generalized Topology Adaptive Graph Convolutional Networks," *arXiv*, Mar. 2024. Available: https://arxiv.org/abs/2403.15077.

[39] Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., & Herwig, R. "ConsensusPathDB: toward a more complete picture of cell biology." *Nucleic Acids Research*, vol. 39, Database issue, pp. D712-D717, 2011. https://doi.org/10.1093/nar/gkq1156.

[40] Li, Z., et al. "CPDB: A comprehensive cancer protein database." *Bioinformatics*, vol. 33, no. 7, pp. 1073-1080, 2017.

[41] Weinstein, J. N., et al. "The Cancer Genome Atlas Pan-Cancer analysis project." *Nature Genetics*, vol. 45, no. 10, pp. 1113-1120, 2013.

[42] Szklarczyk, D., et al. "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets." *Nucleic Acids Research*, vol. 47, no. D1, pp. D607-D613, 2019.

[43] Uhlén, M., et al. "The Human Protein Atlas: an integrated exploration of the human proteome." *Cell*, vol. 162, no. 2, pp. 383-393, 2015.

[44] Kanehisa, M., et al. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research*, vol. 28, no. 1, pp. 27-30, 2000.

[45] Jassal, B., et al. "Reactome: a database of reactions, pathways and biological processes." *Nucleic Acids Research*, vol. 48, no. D1, pp. D497-D503, 2020.

[46] G. Wu, X. Feng, and L. Stein, "A human functional protein interaction network and its application to cancer data analysis," *Genome Biology*, vol. 11, no. 5, p. R53, May 2010. doi: 10.1186/gb-2010-11-5-r53. PMID: 20482850; PMCID: PMC2898064.

[47] The Gene Ontology Consortium. "Gene Ontology: tool for the unification of biology." *Nature Genetics*, vol. 25, no. 1, pp. 25-29, 2000.

[48] Lopes, F. M., et al. "HIPPIE: a human integrated protein-protein interaction database." *Nucleic Acids Research*, vol. 45, no. D1, pp. D207-D212, 2017.

[49] Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Research*, 45(D1), D408–D414. https://doi.org/10.1093/nar/gkw985

[50] Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., and Ideker, T. "Systematic evaluation of molecular networks for discovery of disease genes." *Cell Systems*, vol. 6, no. 4, pp. 484-495.e5, 2018. https://doi.org/10.1016/j.cels.2018.03.001.

[51] Khurana, E., Fu, Y., Chen, J., and Gerstein, M. (2013). Interpretation of genomic variants using a unified biological network approach. *PLoS Computational Biology*, 9(3), e1002886. https://doi.org/10.1371/journal.pcbi.1002886

[52] Rual, J. F., et al. "MIntAct: an open source database and analysis environment for protein-protein interaction data." *Nucleic Acids Research*, vol. 33, suppl. 1, pp. D225-D229, 2005.

[53] Chatr-Aryamontri, A., et al. "The BioGRID interaction database: 2015 update." *Nucleic Acids Research*, vol. 43, no. D1, pp. D470-D478, 2015.

[54] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic Acids Res*, vol. 51, no. D1, pp. D638-D646, Jan. 2023. doi: 10.1093/nar/gkac1000.

[55] Wang, Q., Armenia, J., Zhang, C., et al. "Unifying cancer and normal RNA sequencing data from different sources." *Scientific Data*, vol. 5, 180061, 2018. https://doi.org/10.1038/sdata.2018.61.

[56] Johnson, W. E., Li, C., & Rabinovic, A. "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007. https://doi.org/10.1093/biostatistics/kxj037.

[57] R. A. Fisher, "Statistical methods for research workers," in *Breakthroughs in Statistics*, S. Kotz and N. L. Johnson, Eds., Springer Series in Statistics. New York, NY: Springer, 1992, pp. 66-70. doi: 10.1007/978-1-4612-4380-9.

[58] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the*

Royal Statistical Society, Series B: Statistical Methodology, vol. 57, no. 1, pp. 289–300, 1995.

[59] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *arXiv preprint*, 2018. Available: arXiv:1708.02002.

[60] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2017).

[61] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How Powerful are Graph Neural Networks?," *ICLR*, 2019.

[62] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," *Advances in Neural Information Processing Systems*, vol. 29, pp. 3844–3852, 2016.

[63] Hamilton, William L., Rex Ying, and Jure Leskovec. "Inductive representation learning on large graphs." Advances in neural information processing systems 30 (2017).

[64] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018. Available: https://openreview.net/forum?id=rJXMpikCZ. [Accepted as poster].

[65] D. Chakravarty *et al.*, "OncoKB: A Precision Oncology Knowledge Base," *JCO Precision Oncology*, vol. 2017, p. PO.17.00011, Jul. 2017. doi: 10.1200/PO.17.00011. PMID: 28890946; PMCID: PMC5586540.

[66] Y. Liu, J. Sun, and M. Zhao, "ONGene: A literature-based database for human oncogenes," *Journal of Genetics and Genomics*, vol. 44, no. 2, pp. 119–121, Feb. 2017. doi: 10.1016/j.jgg.2016.12.004. PMID: 28162959.

[67] Ogata H., Goto S., Sato K., Fujibuchi W., Bono H., Kanehisa M. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999. https://doi.org/10.1093/nar/27.1.29.

[68] Zheng Zhang, Lei Cui, Jianping Wu. "Exploring an edge convolution and normalization based approach for link prediction in complex networks." *Journal of Network and Computer Applications*, vol. 189, 103113, 2021. https://doi.org/10.1016/j.jnca.2021.103113.

[69] Zhang Z., Xu H., Zhu G. "Incorporating high-frequency information into edge convolution for link prediction in complex networks." *Scientific Reports*, vol. 14, no. 1, p. 5437, 2024. https://doi.org/10.1038/s41598-024-56144-9.

[70] Repana, D. *et al.* "The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens." *Genome Biology*, vol. 20, p. 1, 2019. https://doi.org/10.1186/s13059-019-1701-7.

[71] Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., Gonzalez-Perez, A., & Lopez-Bigas, N. "A compendium of mutational cancer driver genes." *Nature Reviews Cancer*, vol. XX, pp. XXX–XXX, 2020. https://doi.org/10.xxxx.

[72] Bell, C. C. & Gilan, O. "Principles and mechanisms of non-genetic resistance in cancer." *British Journal of Cancer*, vol. 122, pp. 465–472, 2019. https://doi.org/10.1038.

[73] Repana, D., et al. "The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens." *Genome Biology*, vol. 20, p. 1, 2019. https://doi.org/10.1186/s13059-018-1612-0.

[74] Baylin, S. B. & Jones, P. A. "Epigenetic determinants of cancer." *Cold Spring Harbor Perspectives in Biology*, vol. 8, no. 1, p. a019505, 2016. https://doi.org/10.1101/cshperspect.a019505.

Cancer	Method	AUROC			AUPRC			Cancer	Method	AUROC			AUPRC		
		Bio	**Topo**	**Comb**	**Bio**	**Topo**	**Comb**			**Bio**	**Topo**	**Comb**	**Bio**	**Topo**	**Comb**
BLCA	HGDC	0.7371	0.7130	0.7240	0.7957	0.7382	0.7837	LIHC	HGDC	0.6939	0.6463	0.6724	0.7632	0.7037	0.7465
	EMOGI	0.6335	0.7404	0.7698	0.6812	0.7903	0.8011		EMOGI	0.7215	0.7263	0.7136	0.7595	0.7866	0.7929
	MTGCN	0.7491	0.6800	0.7525	0.8180	0.7677	0.8174		MTGCN	0.7235	0.7022	0.8819	0.8110	0.7824	0.8523
	ACGNN	**0.8585**	**0.8507**	**0.9644**	**0.9079**	**0.9014**	**0.9766**		ACGNN	**0.9073**	**0.8513**	**0.9704**	**0.9391**	**0.9022**	**0.9809**
BRCA	HGDC	0.6406	0.5941	0.6980	0.7023	0.6205	0.7249	LUAD	HGDC	0.7260	0.6777	0.7149	0.7692	0.7297	0.7620
	EMOGI	0.7539	0.7181	0.7127	0.8252	0.7816	0.7592		EMOGI	0.7351	0.7230	0.7512	0.7825	0.7685	0.8196
	MTGCN	0.7095	0.7145	0.6853	0.7876	0.7906	0.7705		MTGCN	0.6916	0.6918	0.7127	0.7828	0.7724	0.7967
	ACGNN	**0.9017**	**0.8105**	**0.9287**	**0.9359**	**0.8714**	**0.9533**		ACGNN	**0.9384**	**0.8068**	**0.9026**	**0.9590**	**0.8704**	**0.9337**
CESC	HGDC	0.6031	0.6635	0.7071	0.6765	0.7085	0.7663	LUSC	HGDC	0.7030	0.6497	0.6589	0.7372	0.7119	0.6921
	EMOGI	0.7592	0.7393	0.7196	0.8226	0.7823	0.7792		EMOGI	0.7761	0.6713	0.6724	0.8320	0.7278	0.7203
	MTGCN	0.7272	0.6969	0.7182	0.8060	0.7801	0.7984		MTGCN	0.7337	0.6976	0.7010	0.8048	0.7824	0.7826
	ACGNN	**0.9636**	**0.7343**	**0.9760**	**0.9749**	**0.8214**	**0.9871**		ACGNN	**0.9569**	**0.7995**	**0.9282**	**0.9716**	**0.8666**	**0.9510**
COAD	HGDC	0.7364	0.6693	0.6849	0.7989	0.7205	0.7249	PRAD	HGDC	0.7499	0.6984	0.7512	0.7840	0.7452	0.8134
	EMOGI	0.6961	0.7492	0.7229	0.7260	0.7820	0.7751		EMOGI	0.7227	0.6942	0.7495	0.7760	0.7404	0.7903
	MTGCN	0.7458	0.7073	0.7400	0.8167	0.7886	0.8088		MTGCN	0.7441	0.6984	0.7023	0.8140	0.7786	0.7858
	ACGNN	**0.8672**	**0.8467**	**0.9376**	**0.9034**	**0.8990**	**0.9565**		ACGNN	**0.9108**	**0.8073**	**0.9466**	**0.9423**	**0.8675**	**0.9654**
ESCA	HGDC	0.6980	0.6656	0.7482	0.7359	0.6874	0.8006	READ	HGDC	0.7362	0.6040	0.6933	0.7753	0.6441	0.7448
	EMOGI	0.7600	0.7627	0.7745	0.7939	0.8219	0.8208		EMOGI	0.7984	0.7276	0.7755	0.8451	0.7775	0.8199
	MTGCN	0.7303	0.7000	0.7266	0.7966	0.7807	0.7935		MTGCN	0.7215	0.6814	0.6934	0.8013	0.7740	0.7828
	ACGNN	**0.9490**	**0.8005**	**0.9757**	**0.9672**	**0.8640**	**0.9845**		ACGNN	**0.9439**	**0.8173**	**0.9773**	**0.9607**	**0.8692**	**0.9846**
HNSC	HGDC	0.7067	0.7301	0.7404	0.7307	0.8105	0.7872	STAD	HGDC	0.6896	0.6036	0.8673	0.7381	0.6569	0.7058
	EMOGI	0.7131	0.7662	0.7073	07699	0.8272	0.7384		EMOGI	0.7557	0.7144	0.7545	0.8453	0.7577	0.7881
	MTGCN	0.7197	0.7045	0.7322	0.7937	0.7803	0.7981		MTGCN	0.7305	0.7182	0.7110	0.8080	0.7940	0.7908
	ACGNN	**0.9468**	**0.8300**	**0.9209**	**0.9643**	**0.8866**	**0.9502**		ACGNN	**0.9439**	**0.8042**	**0.9560**	**0.9628**	**0.8719**	**0.9721**
KIRC	HGDC	0.6861	0.6279	0.7195	0.7458	0.67973	0.7802	THCA	HGDC	0.7152	0.6249	0.6734	0.7731	0.6916	0.7038
	EMOGI	0.7285	0.6366	0.7480	0.7768	0.6995	0.7662		EMOGI	0.7226	0.7165	0.7519	0.7690	0.7739	0.8019
	MTGCN	0.7260	0.7167	0.7159	0.8048	0.7924	0.7906		MTGCN	0.6751	0.7038	0.7410	0.7667	7865	0.8083
	ACGNN	**0.8862**	**0.8643**	**0.9617**	**0.9212**	**0.9073**	**0.9739**		ACGNN	**0.8550**	**0.8312**	**0.9088**	**0.9011**	**0.8832**	**0.9416**
KIRP	HGDC	0.6949	0.6895	0.6944	0.7416	0.7289	0.7477	UCEC	HGDC	0.6596	0.6302	0.7559	0.7181	0.6734	0.8011
	EMOGI	0.8170	0.6623	0.7000	0.8631	0.7056	0.7387		EMOGI	0.8164	0.7696	0.7917	0.8718	0.8144	0.8434
	MTGCN	0.7023	0.7002	0.7227	0.7830	0.7829	0.7948		MTGCN	0.7371	0.7139	0.6608	0.8202	0.7924	0.7483
	ACGNN	**0.8667**	**0.7986**	**0.9168**	**0.9110**	**0.8684**	**0.9403**		ACGNN	**0.9567**	**0.8029**	**0.9852**	**0.9718**	**0.8661**	**0.9902**

TABLE VII – Performance comparison on cancer type-specific driver gene prediction on CPDB network. **Bio**: Biological features-based model. **Topo**: Topological network features-based model. **Comb**: Combined biological and topological features.