Briefings in Bioinformatics, 2025, **26(4)**, bbaf412
https://doi.org/10.1093/bib/bbaf412
Problem Solving Protocol

DriverOmicsNet: an integrated graph convolutional network for multi-omics exploration of cancer driver genes

Yang-Hong Dai [1,2], Chia-Jun Chang [3], Po-Chien Shen [1], Wun-Long Jheng[4,5], Ding-Jie Lee [6,7], Yu-Guang Chen [8,9,*]

[1]Department of Radiation Oncology, Tri-Service General Hospital, National Defense Medical University, No. 325, Sec. 2, Chenggong Rd., Neihu District, Taipei City 114202, Taiwan, Republic of China
[2]Department of Oncology, University of Oxford, Oxford, OX3 7DQ, United Kingdom
[3]Department of Biomedical Engineering, National Cheng Kung University, No. 1, University Rd., Tainan City 701, Taiwan, Republic of China
[4]Cancer Center, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Sec. 3, Zhongyang Rd., Hualien City 970473, Taiwan, Republic of China
[5]Shizuoka Center for Molecular Simulation and Digital Health, Hamamatsu City, Shizuoka 430-0928, Japan
[6]Division of Nephrology, Department of Internal Medicine, Tri-Service General Hospital Keelung Branch, National Defense Medical University, No. 325, Sec. 2, Chenggong Rd., Neihu District, Taipei City 114202, Taiwan, Republic of China
[7]Department of Biological Science and Technology, Institute of Bioinformatics and System Biology, National Yang Ming Chiao Tung University, No. 75, Boai St., East Dist., Hsinchu City 300197, Taiwan, Republic of China
[8]Division of Hematology/Oncology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, No. 325, Sec. 2, Chenggong Rd., Neihu District, Taipei City 114202, Taiwan, Republic of China
[9]The Center for Cell Therapy and Regenerative Medicine, Tri-Service General Hospital, Taipei, Taiwan, Republic of China

*Corresponding author. Division of Hematology/Oncology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, No. 325, Sec. 2, Chenggong Rd., Neihu District, Taipei City 114202, Taiwan, Republic of China. E-mail: song123456tw@gmail.com

Abstract

Cancer is a complex and heterogeneous group of diseases driven by genetic mutations and molecular changes. Identifying and characterizing cancer driver gene is crucial for understanding cancer biology and guiding precision oncology. Integrating multi-omics data can reveal the intricate molecular interactions underlying cancer progression and treatment responses. We developed a graph convolutional network (GCN) framework, DriverOmicsNet, that integrates multi-omics data using STRING protein–protein interaction networks and correlation-based weighted gene correlation network analysis (WGCNA). We applied this framework to 15 cancer types, analyzing 5555 tumor samples to predict cancer-related features such as homologous recombination deficiency, cancer stemness, immune clusters, tumor stage, and survival outcomes. DriverOmicsNet demonstrated superior predictive accuracy and model performance metrics across all target labels when compared with GCN models based on STRING network alone. Gene expression emerged as the most significant feature, reflecting the dynamic and functional state of cancer cells. The combined use of STRING PPI and WGCNA networks enhanced the identification of key driver genes and their interactions. Our study highlights the effectiveness of using GCNs to integrate multi-omics data for precision oncology. The integration of STRING PPI and WGCNA networks provides a comprehensive framework that improves predictive power and facilitates the understanding of cancer biology, paving the way for more tailored treatments.

Keywords: cancer driver genes; multi-omics; graph convolutional networks; STRING PPI; WGCNA; precision oncology

Introduction

Cancer is a complex and heterogeneous group of diseases and continues to pose a significant challenge to medical science. The intricate web of molecular alterations within cancer cells, driven by a myriad of genetic mutations and molecular changes, has made it crucial to decipher the underlying mechanisms orchestrating the disease's initiation, progression, and response to treatment [1]. Among these molecular alterations, the identification and characterization of cancer driver genes (CDgs) have emerged as pivotal milestones in our quest to understand and combat cancer [2].

Driver genes, unlike their passenger counterparts, play a crucial role in cancer development by endowing tumor cells with a selective growth advantage [3]. Mutations in these genes initiate a cascade of molecular events, promoting uncontrolled cell growth, replicative immortality, invasion, metastasis, deregulation of energy metabolism, and evasion of immune suppression [4]. Consequently, discerning the molecular signatures and functional implications of CDgs has become imperative in guiding precision oncology, where treatments are tailored to the unique genetic makeup of each patient's tumor.

In recent years, the advent of next-generation sequencing and high-throughput multi-omics technologies has brought us closer

Received: April 12, 2025. **Revised:** July 14, 2025. **Accepted:** July 16, 2025

to comprehending the intricate landscape of cancer biology [5]. These technologies have enabled researchers to simultaneously explore DNA mutations, gene expression, copy number variations (CNVs), epigenetic modifications, protein levels, and metabolic profiles, collectively referred to as "multi-omics" data. These multi-omics data sources offer unprecedented insights into the molecular underpinnings of cancer, making it possible to uncover the complex interplay between genetic alterations, gene expression patterns, and clinical outcomes. In addition to mutations that characterize "dysfunctional" events in genes, other multi-omics data contribute to a holistic understanding by shedding light on "dysregulation" events in oncogenesis [6]. However, the challenge lies in integrating these diverse layers of information to elucidate the causal relationships between molecular signatures and cancer phenotypes.

Numerous approaches have emerged over the years to tackle the integration of multi-omics data. Many of these efforts have centered on unsupervised data integration, often lacking the context of sample labels [7–9]. However, with the rise of personalized medicine, the availability of meticulously curated datasets with comprehensive sample annotations, characterizing phenotypes or traits, has expanded significantly. Consequently, there is a growing interest in supervised multi-omics integration techniques, capable of discerning disease-related biomarkers and making predictions for new samples.

In recent years, deep learning (DL) has showcased its formidable prowess in bioinformatics [10]. For example, Wang *et al*. applied DL algorithms, specifically leveraging graph neural networks (GNNs) to circumvent potential biases inherent in traditional methods like feature concatenation and ensemble techniques [11]. In their pioneering work (MOGONET), they harnessed the capabilities of GNNs to uncover correlations among diverse omics data types. They adopted a novel approach by training distinct GNNs, each tailored to a specific omics data type, based on weighted similarity networks. They significantly improved patient classification, revealing pivotal biomarkers for Alzheimer's disease and breast cancer. However, their methodology fell short in the integration of all omics data into a unified graph and only offered a datatype-specific patient embedding [12]. To better understand how multi-omics data can elucidate relationships among CDgs, it is necessary to integrate the multi-omics features simultaneously for a local graph network. Additionally, MOGONET does not train on curated network structure such as protein–protein interaction (PPI) network, which plays essential roles in structuring and mediating biological processes [13].

To address the aforementioned limitations, we developed a framework of graph convolutional network (GCN) named DriverOmicsNet that combined multi-omics graph embedding derived from STRING PPI network and correlation-based weighted gene correlation network analysis (WGCNA) for CDgs across multiple cancer types. DriverOmicsNet demonstrates remarkable predictive accuracy and excels in various model performance metrics, including tasks related to homologous DNA repair deficiency, cancer stemness, immune cluster, tumor stage, and survival prediction. Moreover, we have delved into the interpretability of our GCN models, revealing their ability to discern essential driver markers. By leveraging GNNs and a wealth of multi-omics data, our study embarks on a transformative journey that redefines the precision oncology paradigm. Our goal is to bridge the gap between the detailed molecular phenotypes of cancer cells and clinical outcomes, paving the way for tailored treatments that improve patient care.

Methods

Tumor samples for model construction

Multi-omics data were obtained from the UCSC Xena platform (https://xena.ucsc.edu) using the R package UCSCXenaTools [14]. This approach enabled the efficient retrieval and preprocessing of high-quality cancer genomics data. To ensure statistical robustness and meaningful tumor–normal comparisons, we excluded cancer types with fewer than 100 tumor samples or lacking anatomically matched normal tissues in the GTEx database. In particular, cancer types such as TCGA-LGG (lower-grade glioma) and SARC (sarcoma) were excluded due to incomplete clinical annotation, such as missing or non-applicable American Joint Committee on Cancer (AJCC) staging, which is essential for consistent comparison across cancer subtypes. Additionally, although some cancer types initially met the sample size criterion, further filtering for complete multi-omics and clinical label availability reduced their effective sample size below the inclusion threshold. Finally, tissues such as the brain, which comprises highly heterogeneous subregions in GTEx, were excluded to avoid introducing variability that could confound biological interpretation [15]. After applying these criteria, a final cohort of 15 cancer types, comprising 5555 tumor samples, was retained for downstream analyses.

Multi-omics data type and preprocessing

mRNA gene expression

RSEM (RNA-Seq by expectation–maximization) expected counts from the UCSC Toil Recompute Compendium were retrieved and processed following the protocol by Chen et al [16, 17]. Counts were back-transformed and normalized using the voom method from the limma package for linear modeling [18]. An additional 2656 cancer-free tissue samples from GTEx served as controls. The voom-normalized expression data were used for WGCNA and downstream multi-omics integration via GNN models. To evaluate the reliability of candidate CDgs in distinguishing cancer subtypes from matched normal tissues, we calculated gene significance (GS) using WGCNA. GS was defined as the $-\log_{10}$(*P*-value) derived from linear regression of gene expression against the trait of interest, providing a quantitative measure of each gene's association with tumor-related phenotypes. No explicit GS threshold was applied; all candidate CDgs were retained to preserve the full spectrum of potentially relevant signals.

CNV and mutation

The retrieved CNV data were estimated by GISTIC2 and thresholded, with −2, −1, 0, 1, 2 indicating two copy deletions, one copy deletion, no change, amplification, and high amplification [19]. Somatic mutation data was determined from the MC3 project, including nonsilent mutation such as single nucleotide polymorphisms and insertion-deletions [20]. Binary mutation call was used as the omics data in our study.

Methylation

We obtained β values from the Illumina HumanMethylation450 array, which served as our representation of DNA methylation at CpG sites [21]. CpG probes $p \in \mathbb{R}$ corresponding to each candidate CDg were grouped by genomic context—CpG island, non-island, promoter, non-promoter, and enhancer—based on the probeMap file (hg38) [22]. Probe vectors of varying lengths (m) were zero-padded to a fixed length (M) per cancer type. To obtain a compact, biologically meaningful representation, we applied an Autoencoder (AE) with three encoder layers and LeakyReLU activations. An optional binary masking mechanism excluded padded values

during training, preserving only valid input data. The output of this AE model is a compact, low-dimensional representation of DNA methylation data for each CDg:

$$f_{AE} : P_{padded} \in \mathbb{R}^M \rightarrow P_{enc} \in \mathbb{R}^n,$$

where n is set to 1 in our study. The low-dimensional output was verified and visualized by t-SNE (t-distributed stochastic neighbor embedding), based on the mean methylation values across all methylated probes in different regulatory regions.

Determination of cancer drivers

DriverDBv3 (http://driverdb.bioinfomics.org), an integrative multi-omics database, was used to identify candidate CDgs across cancer types [6]. It leverages established bioinformatics algorithms to detect CDgs and further provides insights into CNV and methylation drivers, offering a comprehensive view of cancer-associated dysregulatory events.

Labels for binary classification tasks

To evaluate DriverOmicsNet's ability to predict clinically relevant cancer features, we performed binary classification on homologous recombination deficiency (HRD), cancer stemness, immune clusters, tumor stage, and overall survival [23, 24]. HRD, transcriptomic stemness scores, and immune signature (IM) scores were retrieved from the TCGA Pan-Cancer hub via UCSC Xena [25]. HRD and stemness were binarized using median values (Supplementary Figs. 1–2). The IM scores, comprising 68 IMs, were clustered using hierarchical clustering (*hclust* in R; Supplementary Fig. 3). Tumor stages were grouped as early (stage I–II) or late (stage III–IV), and survival outcomes were dichotomized based on median survival.

Overview of DriverOmicsNet

DriverOmicsNet is a unified framework for multi-omics-based binary classification. It comprises two main components (Fig. 1): (i) network construction using STRING-based PPI and WGCNA-derived correlation networks, and (ii) fusion of latent embeddings from dual GCN models. Final predictions are made by concatenating the output vectors from both networks, enabling robust integration of complementary biological information for classification tasks.

GCN for DriverOmicsNet

The goal of DriverOmicsNet is to take advantage of GCN and explore the key graphlets that is crucial for label classification. First, STRING PPI graph $g_{p(v,e)}$ and correlation-based WGCNA graph $g_{w(v,e)}$ were constructed for each cancer type, where v and e represented node set and edge set, respectively. STRING is a well-established database for experimentally verified interactions among query proteins, therefore providing a rich foundation for our analysis [26]. For $g_{p(v,e)}$, undirected edges connecting each CDg were defined based on a combined score > 0.4. For constructing $g_{w(v,e)}$, we first obtained a similarity co-expression matrix among CDgs. In addition to traditional Pearson's correlation, we used Distance correlation (*dcor* package in python) to explore its ability of capturing biologically meaningful networks [27]. Hub genes for each network were defined as the genes with the highest node degrees. Next, adjacency matrices were obtained by using the soft-thresholding power, which was selected to approximately fit a scale-free topology. Subsequently, topological overlap matrices $TOM_{w,p}$ and $TOM_{w,d}$ were generated. To enhance the comparability with $g_{p(v,e)}$, thresholding was applied for each cancer type to approximate the edge set number of $g^{n_1}_{p(v,e)}$:$e \in Z^{n_1}$ with $g^{n_2}_{w(v,e)}$:$e \in Z^{n_2}$, so that $n_1 \sim n_2$.

In order to integrate multi-omics data in the processes of message passing and aggregation for GCN, the graphs $g^{n_1}_{p(v,e)}$ and $g^{n_1}_{w(v,e)}$ were converted to data structures acceptable for training a GCN model, which is a node feature matrix $X \in \mathbb{R}^{nxk}$ where n is the number of CDgs and k is the feature size (size = 8 for gene expression, CNV, mutation, CpG island, non-CpG island, promoter, nonpromoter, and enhancer), and a data structure called edge index:

$$E = [(u_1, v_1), (u_2, v_2), \ldots (u_m, v_m)] \in Z^{2xm},$$

where (u_m, v_m) represents a pair of node indices connected by an edge and m denotes the total number of edges in an undirected graph. For $g_{w(v,e)}$, edge attribute ϵ for each edge is given as an additional parameter, which is the value from $TOM_{w,p}$ and $TOM_{w,d}$.

Specifically, for $g^{n_1}_{p(v,e)}$ and $g^{n_2}_{w(v,e)}$, the GCN models can be denoted as $GCN^c_l(X, E)_p$ and $GCN^c_l(X, E, \epsilon)_w$ where $l \in Z^5$ is the target label of interest and $c \in Z^{15}$ is the cancer type. Generally, each model has the same structure, with four GCN layers and three fully connected layers (multilayer perceptrons, MLPs). Each GCN layer is activated by ReLU, which is followed by TopKPooling to reduce the number of nodes while retaining the most informative ones. The pooled nodes are then globally pooled to obtain a graph-level representation, resulting in four layers:

$$L = (L_1, L_2, L_3, L_4) \in \mathbb{R}^{kxh},$$

where h is is the hidden dimension during training. The four L layers are concatenated to create a comprehensive graph-level representation:

$$L^* : concat\,(L_1, L_2, L_3, L_4) \in \mathbb{R}^{kxh},$$

which is followed by the three MLPs. The output of the last linear layer is passed through a softmax function to obtain class probabilities for classification. Configuration of the GCN structure is shown in Supplementary Fig. 4.

DriverOmicsNet for binary classification

We randomly split the total samples of each cancer type into training and testing sets with a 8:2 ratio. We fine-tuned the hyperparameters (batch size, hidden dimension, learning rate, and weight decay rate) for $GCN^c_l(X, E)_p$ and $GCN^c_l(X, E, \epsilon)_w$ with HyperOPT and Ray [28]. Maximal number of epochs of 500 with early stopping was used with the patience of 10. Initially, $GCN^c_l(X, E)_p$ and $GCN^c_l(X, E, \epsilon)_w$ were trained in an end-to-end fashion. To establish a final model that is able to capture local network structure from STRING PPI and WGCNA networks, L^* from each model L^*_p and L^*_w are concatenated, forming

$$L^*_f : concat\left(L^*_p, L^*_w\right),$$

which is followed by two MLPs to make final prediction (Supplementary Fig. 5)

Model performance metrics

To evaluate training and testing accuracies, we performed fivefold cross-validation for both DriverOmicsNet and baseline graphlet-specific models. Accuracy values across folds were then compared using paired two-tailed t-tests to assess whether the observed improvements were statistically significant ($P < .05$). Additionally,

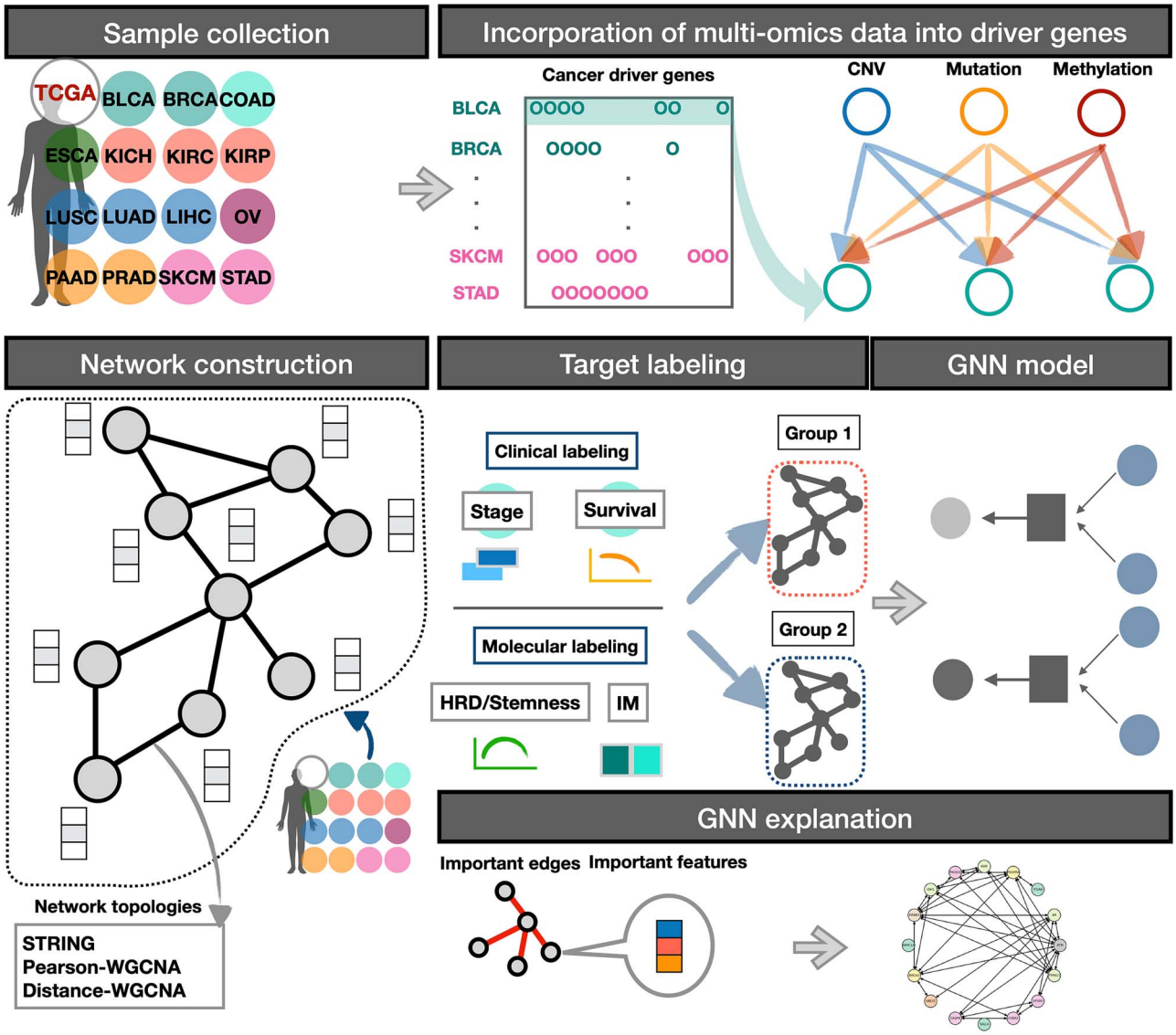

Figure 1. Overview of DriverOmicsNet.

to assess the overall performance and robustness of DriverOmicsNet, we applied five-fold cross-validation and independent testing. Model evaluation was based on multiple metrics, including accuracy, F1-score, Matthews correlation coefficient (MCC), and area under the receiver operating characteristic curve.

We compared our model with existing machine learning (ML) algorithms including an attention-based GCN (AGCN) [29], K-nearest neighbor, support vector machine, logistic regression, random forest, least absolute shrinkage and selection operator (LASSO), elastic net, and MLP. For all baseline models, the multi-omics features were concatenated into a unified input matrix. The AGCN model integrates a GCN architecture with three complementary attention mechanisms, squeeze-and-excitation blocks, column-wise self-attention, and row-wise self-attention, to model both omics-level feature contributions and topological influences from neighboring nodes. The STRING PPI was used as the underlying graph structure to guide message passing between CDgs.

Identification of key CDgs for binary classification

We applied GNNExplainer to our GCN model over 100 training epochs, generating node and edge importance scores via learned feature and edge masks. This process identified the most influential components driving model predictions. A similar procedure was conducted for the distance network. Cumulative importance scores across the dataset were used to select the top 100 nodes and edges, with a threshold set at the 100th highest score. These key elements were then used to construct and visualize subgraphs, offering insights into their structural and functional relevance.

External validation

To externally validate the predictive performance of DriverOmicsNet, we applied the model structure to independent multi-omics datasets from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Specifically, we utilized data from pancreatic ductal adenocarcinoma (PAAD), retrieved via LinkedOmics (http://www.linkedomics.org), comprising matched gene expression (log_2-transformed RSEM values, upper quartile normalized), CNV (GISTIC2), somatic mutation, and gene-level DNA methylation (β values). For binary label construction, immune subtypes were defined by median dichotomization of the ESTIMATE immune scores obtained from the CPTAC metadata. HRD status was inferred from hallmark DNA repair pathway scores, also binarized at the median. Tumor stage and survival data were directly extracted from CPTAC annotations.

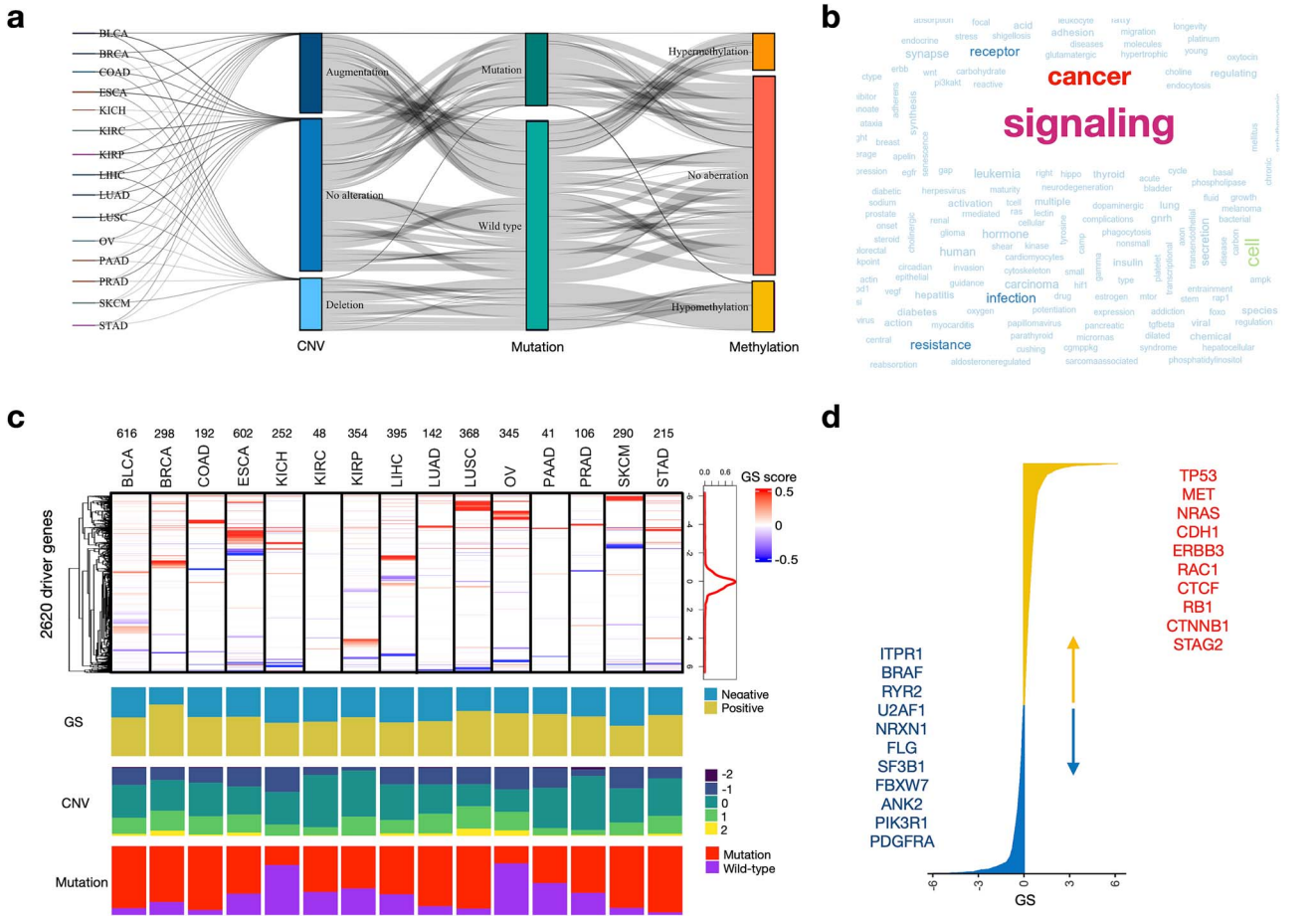

Figure 2. Characterization of the CDgs used in the study for each cancer type. (a) Sankey diagram showing the associations among CNVs, mutations and abnormal methylations for CDgs. Nodes in the graph with more connections suggest a greater influence. (b) Enriched gene ontology terms for all CDgs across all cancer types. (c) Heatmap shows GS score for all CDgs across all cancer types used in DriverOmicsNet. The density curve summarizes GS values from all cancer types (N = 2620 genes). Stacked bar plots indicate proportion of positive and negative GS, mutational status and CNV for each cancer type. (d) Cumulative GS scores for all CDgs across all cancer types. WGCNA = weighted gene co-expression network analysis. GS = gene significance. CNV = copy number variation.

Notably, DNA methylation features were directly input into the model without AE transformation, in contrast to other omics modalities, which followed the standard preprocessing pipeline of DriverOmicsNet. The model was subsequently retrained using the restructured CPTAC data, and classification performance was evaluated using the same metrics as those applied in the initial training phase.

Results

Multi-omics features of driver genes in cancer

Across 15 cancer types, DriverDBv3 identified 2620 candidate CDgs, integral to TCGA cohorts and cancer progression. Multi-omics correlations among these CDgs are summarized in Fig. 2a, based on the database's driver summary (http://driverdb.bioinfomics.org). Over-representation analysis (ORA) highlighted cancer signaling as the most enriched pathway (Fig. 2b). GS analysis revealed strong correlations between CDg expression and tumor traits (Fig. 2c), with notable subsets exhibiting negative GS values, including oncogenes such as *PDGFRA*, *PIK3R1*, and *BRAF* (Fig. 2d) [30–32]. In contrast, *TP53* consistently showed positive GS across all cohorts, reaffirming its central role as a universal driver gene [33]. Analysis of CNV and mutation data further identified *TP53* and *PIK3CA* as the most frequently altered CDgs across cancers (Supplementary Fig. 6) [34]. Collectively, these findings underscore the diversity of multi-omics profiles in TCGA, allowing GNN models to generalize better to different scenarios and make more accurate predictions or classifications.

Encoding of DNA methylation data

Supplementary Fig. 7 shows the distribution of β values across genomic regions for each cancer type, revealing region-specific methylation patterns. An AE model with LeakyReLU activation functions was used to reduce dimensionality while preserving key biological signals (Fig. 3a and b). This model uncovered distinct DNA methylation clusters across multiple cancers (Fig. 3c; Supplementary Fig. 8), demonstrating its capacity to capture meaningful patterns. To evaluate predictive potential, we trained a MLP using only the one-dimensional AE output. The model achieved high accuracy, with AUC values of 0.92 (95% CI: 0.86–0.98) for promoter regions in breast invasive carcinoma (BRCA), 0.86 (95% CI: 0.81–0.91) for CpG islands in lung adenocarcinoma (LUAD), 0.92 (95% CI: 0.86–0.99) for non-CpG islands in stomach adenocarcinoma (STAD), and 0.92 (95% CI: 0.87–0.99) for enhancers in skin cutaneous melanoma (SKCM). These results highlight the framework's ability to simplify and accurately stratify DNA methylation landscapes across cancers (Fig. 3c).

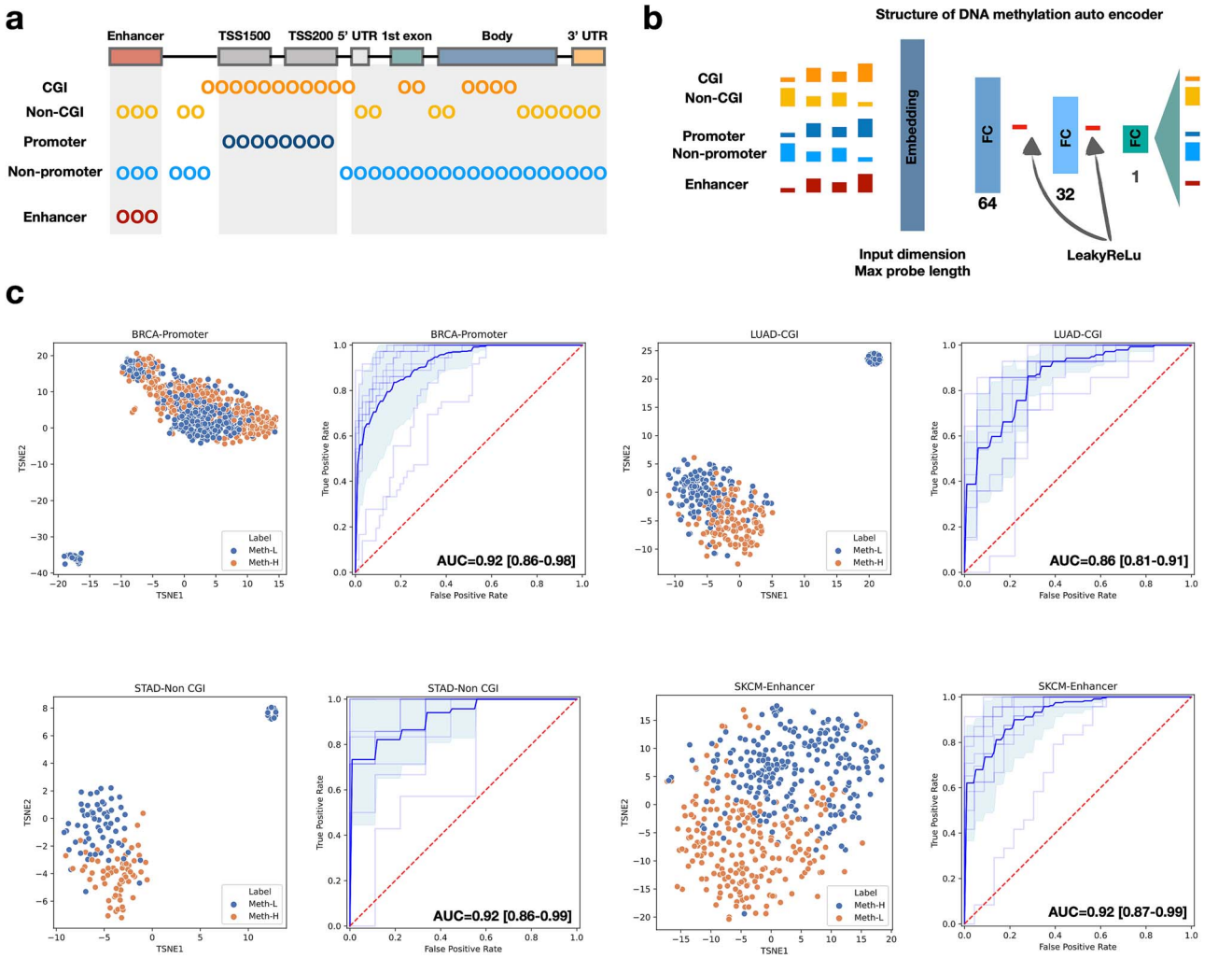

Figure 3. Encoding of methylation for multi-omics integration. (a) Illustration of different genomic regulatory regions targeted by the HumanMethylation450 array, including CGI, non-CGI regions, promoter and nonpromoter regions, and enhancers. These categories are further mapped to genomic contexts such as TSS1500, TSS200, 5′ UTR, 1st exon, gene body, and 3′ UTR, highlighting the spatial distribution of methylation probes. (b) Schematic architecture of the DNA methylation AE used to reduce high-dimensional probe-level β-values into compact, biologically relevant embeddings. The AE comprises FC layers with LeakyReLU activation, and the input structure accounts for multiple regulatory categories with variable probe lengths. (c) t-SNE plots and area under curve plots showing distribution of decoded outputs from AE and their capabilities of predicting mean methylation statuses in BRCA, LUAD, STAD and SKCM across different regulatory regions. CGI = CpG islands. AE = autoencoder. FC = fully connected. BRCA = breast invasive carcinoma. LUAD = lung adenocarcinoma. STAD = stomach adenocarcinoma. SKCM = skin cutaneous melanoma.

Predicting outcomes based on separate network structure

The number of undirected edges derived from STRING PPI varies across different cancer types, closely mirroring the respective input gene numbers (Supplementary Fig. 9). Supplementary Fig. 10 illustrates the thresholds of correlation coefficients used to ensure the comparability of PPI and CDg co-expression modules. In the case of $g^{n_1}_{p(v,e)}$, the hub genes remain consistent across all cancer cohorts, with *TP53* exclusively identified as the hub gene (Supplementary Table 1). However, for $g^{n_2}_{w(v,e)}$, there is a striking divergence in hub genes across various cancer types. Most cancer types exhibit shared hub genes between the Pearson- and Distance-WGCNA networks, except for a few exceptions, including esophageal carcinoma, liver hepatocellular carcinoma (LIHC), ovarian serous cystadenocarcinoma (OV), and STAD.

A total of 225 GCN models were evaluated based on distinct network structures. WGCNA-based models consistently outperformed those using STRING PPI networks across all five clinical labels (Fig. 4a). Among WGCNA variants, models built on distance correlation showed slight yet consistent improvements over Pearson-based models. Based on these findings, we selected the top-performing Distance-WGCNA and STRING models for integration. An additional 75 fusion models were trained using latent vector concatenation. This combined approach significantly boosted training and testing accuracy across cancer types (all *P*-value <.05), with testing improvements ranging from 0.1% (HRD) to 15.2% (survival) compared to individual models (Fig. 4b; Supplementary Fig. 11). The comparison with other ML algorithms revealed that the combined models consistently achieved high testing accuracy across all five classification tasks (Supplementary Fig. 12).

The top 30 combined models

To evaluate the predictive potential of our combined models, we focused on the top 30 models, representing the top 40% in performance. These were predominantly trained to predict IM, followed by stemness and tumor stage (Supplementary Fig. 13). Notably, none of the top models targeted survival, indicating

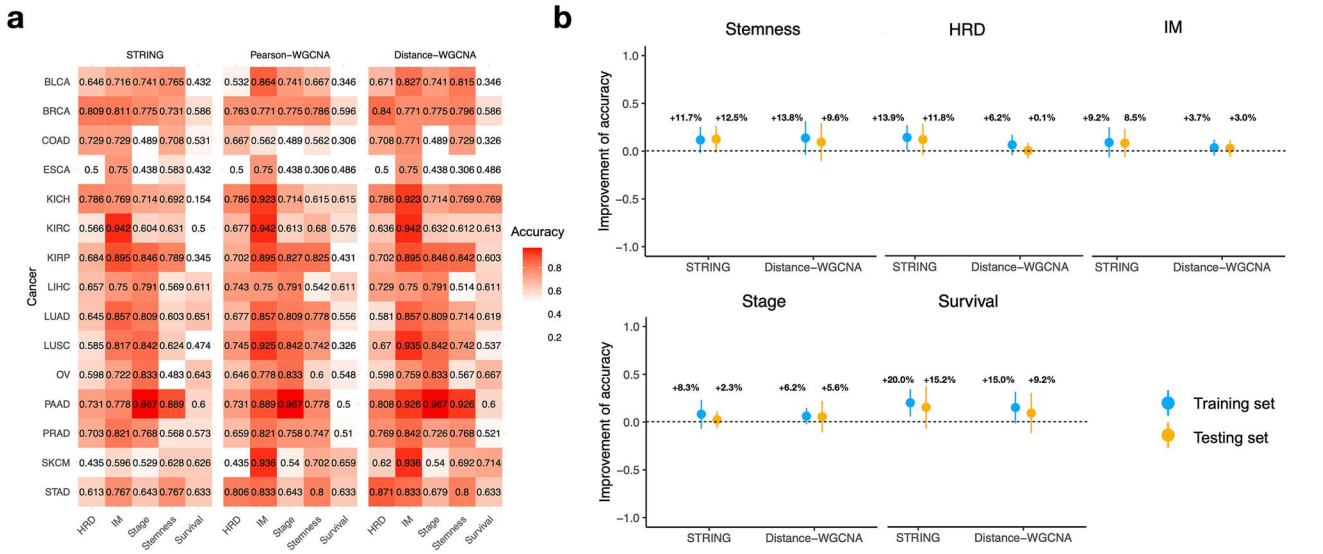

Figure 4. Model accuracies of GCN. (a) Model testing accuracies based on network structures of STRING, Pearson correlation-based WGCNA, and Distance correlation-based WGCNA across five target features. Color bar indicates the testing accuracy. (b) Improvement of accuracies observed in DriverOmicsNet. Improvements of accuracy are displayed as mean ± standard deviation for training and testing sets. GCN models built from STRING and Distance correlation-based WGCNA networks are set as the baseline. GCN = graph convolutional network. WGCNA = weighted gene co-expression network analysis.

limited predictive power for this outcome. Among the top 10 models, seven predicted IM clusters, achieving testing accuracies ranging from 89.47% for kidney renal papillary cell carcinoma (KIRP) to 95.15% for kidney renal clear cell carcinoma (KIRC). We then selected the highest-performing models for each label type: STAD for HRD, KIRP for both stemness and stage, and SKCM for IM prediction (Fig. 5). The STAD-HRD model achieved the strongest performance (Precision = 0.9307 ± 0.0108; F1 = 0.9102 ± 0.0938; MCC = 0.8368 ± 0.1540; AUC = 0.9604 ± 0.0642), followed by the SKCM-IM model (Precision = 0.9091 ± 0.0656; F1 = 0.8749 ± 0.1078; MCC = 0.7745 ± 0.1837; AUC = 0.9195 ± 0.0960). In contrast, the KIRP-stage model had the weakest performance, potentially due to class imbalance (label ratio = 2.33). Two additional stage prediction models with more balanced label ratios also underperformed (data not shown), further suggesting that stage classification may be more challenging in this setting. Although class imbalance may have compromised the performance of certain models, the combined model framework consistently outperformed traditional ML algorithms and AGCN (Supplementary Table 3).

Omics importance and graphlet identification for molecular features

To assess the contribution of each omics feature to graph prediction, we selected 17 models with balanced labels (label ratio < 3) to reduce bias and ensure robustness. Cumulative node feature importance and subgraphs with overlapping CDgs from both STRING and Distance WGCNA models are shown in Supplementary Fig. 14. Gene expression consistently emerged as the most influential feature, followed by CNV, while mutation status had the lowest importance. Methylation data displayed high variability across cancer types and graph structures, which was also reflected in subgraph connectivity. These results highlight gene expression and CNV as key drivers of graph-based predictions and emphasize the importance of accounting for cancer-type-specific variation when integrating methylation data.

Validation with the CPTAC PAAD dataset

External validation using the CPTAC PAAD dataset demonstrated that DriverOmicsNet maintained strong predictive performance across multiple clinical and molecular labels. As shown in Fig. 6a, the model achieved robust accuracy, precision, F1-score, MCC, and AUC in classifying tumor stage, survival, IM, and DNA repair status. Notably, prediction of IM and DNA repair status yielded the highest classification metrics among the four tasks. Cumulative node feature importance analysis from the Distance WGCNA models predicting IM and DNA repair status revealed that gene expression and gene-level DNA methylation were the most influential features (Fig. 6b), while CNV and somatic mutation features contributed less substantially. These trends were consistent with the feature importance patterns observed in the TCGA-PAAD cohort (Supplementary Fig. 14). These findings support the model's ability to generalize independent multi-omics data and highlight the complementary value of integrating diverse omics modalities.

Discussion

Our study highlights the effectiveness of DriverOmicsNet, a GCN-based framework, in integrating multi-omics data to predict key cancer-related features, including HRD, cancer stemness, and immune clusters. By combining STRING PPI and correlation-based WGCNA networks, DriverOmicsNet demonstrates strong potential for precision oncology. Notably, we introduce a novel autoencoder-based method for processing DNA methylation data, which captures biologically meaningful patterns across genomic regions while reducing dimensionality. This approach enhances both the interpretability and predictive power of methylation features, offering a more refined and biologically relevant representation for multi-omics integration.

Notably, the performance of DriverOmicsNet in predicting survival outcomes was relatively limited compared to other classification tasks. This may reflect the inherent complexity of survival as a phenotype, which is influenced by a wide array of genetic, clinical, and treatment-related factors that may not be fully captured by multi-omics data alone. Additionally, dichotomizing survival for classification purposes likely reduced the granularity of prognostic signals. Future extensions of DriverOmicsNet may benefit from incorporating survival-specific modeling approaches,

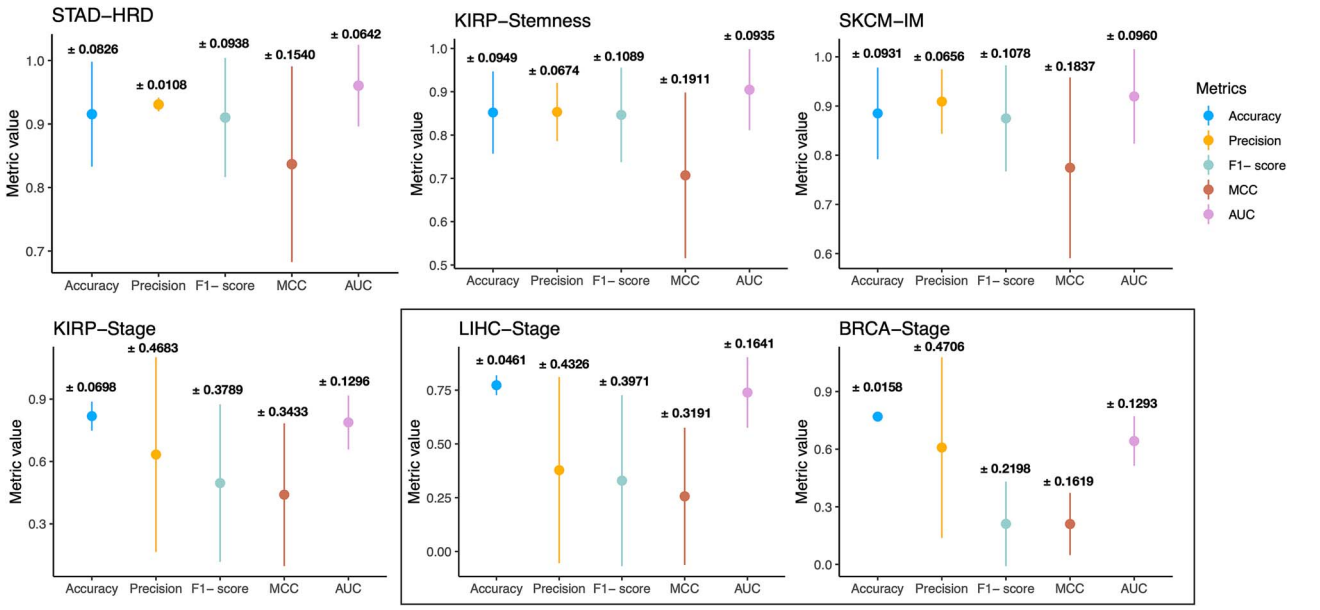

Figure 5. Model performance for selected models with the highest testing accuracy for HRD, stemness, IM, and stage predictions. Additional two models are demonstrated for stage. HRD = homologous recombination deficiency. IM = immune signature. STAD = stomach adenocarcinoma. KIRP = kidney renal papillary cell carcinoma. SKCM = skin cutaneous melanoma. LIHC = liver hepatocellular carcinoma. BRCA = breast invasive carcinoma. MCC = Matthews correlation coefficient. AUC = area under curve.

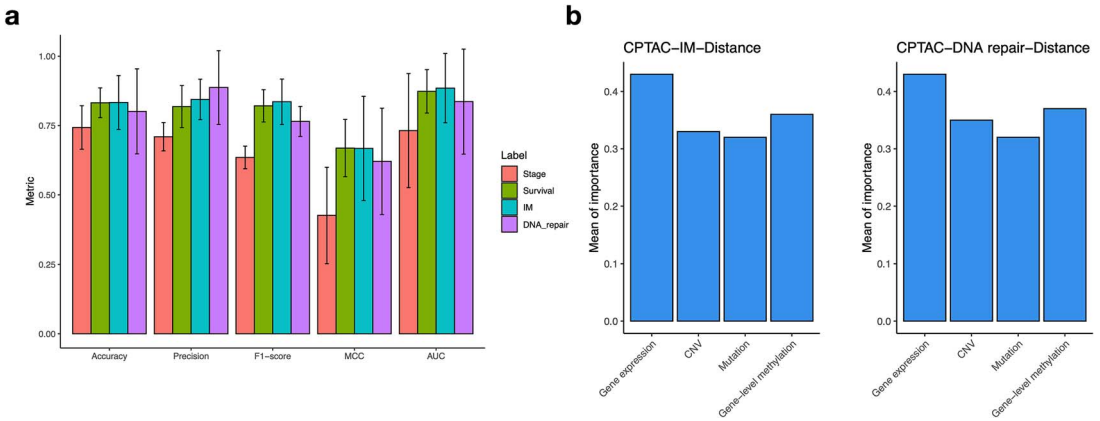

Figure 6. External validation of DriverOmicsNet on the CPTAC PAAD dataset. (a) Classification performance of DriverOmicsNet across four clinical and molecular endpoints—tumor stage, survival, IM, and DNA repair status. (b) Cumulative node feature importance derived from Distance correlation-based WGCNA models for predicting IM (left) and DNA repair status (right). IM = immune signature. MCC = Matthews correlation coefficient. AUC = area under curve.

such as Cox-based or time-to-event DL frameworks, to better capture the continuous nature of survival data.

While the STRING PPI network is comprehensive and regularly updated, it has key limitations [35]. Its coverage of experimentally validated interactions is relatively limited for certain diseases, such as Alzheimer's, compared to other databases like hPRINT [36]. Additionally, inconsistencies in how interactions are classified—what STRING considers experimentally verified may not align with classifications in other resources—can lead to confusion [26]. Despite these issues, STRING remains a valuable tool, offering broad coverage across over 5090 organisms and 24.6 million proteins, making it a widely used resource for systems biology research [35].

STRING integrates experimental data, computational predictions, and text mining to provide a comprehensive overview of PPIs. However, constructing PPI networks using WGCNA-based co-expression offers complementary advantages. WGCNA builds networks that reflect condition-specific gene expression patterns, enabling the identification of interactions relevant to particular tissues, developmental stages, or disease states [37]. Unlike STRING's static interactions, WGCNA captures dynamic expression changes, revealing context-dependent interactions that may arise only under specific stimuli or conditions [38]. Moreover, by directly incorporating gene expression data, WGCNA facilitates a clearer understanding of regulatory mechanisms and the relationship between network topology and gene activity.

Additionally, WGCNA offers quantitative metrics of gene connectivity within modules, enabling the identification of hub genes that are central to network function. Its focus on co-expression minimizes false positives often introduced by STRING's integration of heterogeneous data sources. While STRING's broad coverage is valuable, it may include spurious interactions that lack relevance in specific biological contexts.

A key advantage of WGCNA is its flexibility in selecting co-expression metrics. Using distance-based co-expression instead of Pearson correlation offers several benefits. Distance metrics

can capture nonlinear gene relationships often missed by Pearson correlation [27], which is crucial in complex biological systems. They are also more robust to outliers and better accommodate variations in gene expression across conditions, leading to more accurate identification of modules and interactions. This makes distance-based WGCNA especially suited for analyzing heterogeneous datasets, such as those in cancer research.

The use of GCNs in our study provides distinct advantages by enabling the integration of multi-omics data within a unified graph framework. GCNs effectively model complex gene interactions by combining the static, experimentally validated connections from STRING with the dynamic, condition-specific relationships from WGCNA [39]. Unlike the MOGONET framework by Wang *et al.*, which employs separate GNNs for each omics layer [11], our approach merges all omics into a single cohesive model, offering a comprehensive view of molecular interactions. While AGCN incorporates three complementary attention mechanisms (squeeze-and-excitation, column-wise, and row-wise self-attention), its dependence on a fixed topological structure—such as PPI networks alone—fundamentally limits its adaptability [29]. In contrast, our framework uniquely integrates both PPI and co-expression network topologies, enabling a more comprehensive representation of biological interactions. This dual-network approach not only enhances predictive performance but also improves the identification of key driver genes, offering deeper mechanistic insights into cancer biology and accelerating discoveries in precision oncology.

Our findings highlight the central role of CDgs and emphasize gene expression as the most influential molecular feature across cancer types and clinical outcomes. Gene expression is a powerful indicator of cellular function, reflecting active biological pathways and enabling precise characterization of cancer phenotypes. Its dynamic nature captures condition-specific changes, offering insights into tumor adaptation and treatment response. When integrated through WGCNA, gene expression correlates with other genomic alterations, such as CNVs and mutations, enhancing the interpretability of multi-omics data and reducing false-positive driver gene identification, thereby improving model accuracy and biological relevance [40]. The task-specific feature importance identified by DriverOmicsNet, derived from both STRING and distance correlation-based WGCNA network structures, reveals biologically meaningful associations between gene expression and phenotype-relevant processes across cancer types. For the IM clustering task, strong predictive performance in cancers such as STAD and SKCM (Supplementary Fig. 12) corresponds with the identification of hub genes such as *ANK2* and *ACTB*, respectively. *ANK2* has been reported to correlate with response to immune checkpoint inhibitors in LUAD [41], while *ACTB* is associated with immune cell infiltration across multiple cancer types [42]. The involvement of *ANK2* in stratifying immune states in STAD highlights a potential biomarker that warrants further investigation. In the context of HRD, key genes such as *PTEN* in PRAD and *BIRC5* in BRCA are consistent with their established roles in maintaining genomic stability and DNA damage response pathways [43, 44]. Collectively, the concordance between the model-identified hub genes and established biological mechanisms highlights the interpretability and translational potential of the integrative multi-omics framework provided by DriverOmicsNet. Further research is still needed to decipher the precise biological functions of these genes across various cancer types, enhancing their translational value and clinical applicability.

Despite encouraging results, several limitations must be noted. First, the accuracy and generalizability of our multi-omics integration depend heavily on data quality and completeness; gaps or biases may affect model performance. Second, although our framework integrates static and dynamic networks, it does not capture temporal variations in gene expression and protein interactions, which are key to understanding cancer evolution and treatment response. Additionally, the computational complexity of GCNs may hinder broader clinical adoption. Lastly, while our models performed well across multiple cancer types, their applicability to rare cancers and small datasets requires further validation. Future work should prioritize expanding dataset diversity, enhancing data consistency, and optimizing computational efficiency.

Conclusion

In conclusion, DriverOmicsNet's integration of STRING and WGCNA networks harnesses the complementary strengths of static and dynamic interaction data, enhancing the biological relevance and robustness of PPI networks for precision oncology. This dual-network approach enables a deeper understanding of cancer biology and supports the development of more accurate predictive models and therapeutic strategies. The incorporation of distance-based co-expression in WGCNA further improves network fidelity, making it a valuable method for constructing biologically meaningful gene interaction networks.

Key Points

- We propose DriverOmicsNet, a GNN-based model that integrates multi-omics data and gene co-expression networks to identify CDgs across 15 cancer types.
- DriverOmicsNet incorporates both STRING PPI and WGCNA-derived co-expression networks to capture gene–gene interactions and co-regulatory structures, enhancing biological interpretability.
- The model employs a GCN architecture to integrate diverse omics features, including gene expression, mutation, CNV, and DNA methylation. An autoencoder is used to compress methylation data for effective integration.
- External validation using the CPTAC PAAD dataset demonstrates the robustness of DriverOmicsNet's predictions.
- Comparative analyses demonstrate that DriverOmicsNet achieves competitive performance relative to baseline models, with improved interpretability and biological relevance of identified CDgs.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflicts of interest: None declared.

Funding

This research was further supported by several grants from the National Science and Technology Council Funding & Awards (114-2314-B-016-004-MY2), VGH, TSGH, AS Joint Research Program

(113DCA0200005), the Tri-Service General Hospital (TSGH-E-113225/TSGH-E-114227) and the National Defense Medical Centre (MND-MAB-D-114084) awarded to YGC.

Data and software availability

The data used in this study were obtained from the UCSC Xena browser (https://xena.ucsc.edu) and LinkedOmics (http://www.linkedomics.org). The source codes are available at https://github.com/YangHongDai/DriverOmicsNet.

References

1. Chakravarthi BV, Nepal S, Varambally S. Genomic and Epigenomic alterations in cancer. *Am J Pathol* 2016;**186**:1724–35. https://doi.org/10.1016/j.ajpath.2016.02.023.
2. Ostroverkhova D, Przytycka TM, Panchenko AR. Cancer driver mutations: Predictions and reality. *Trends Mol Med* 2023;**29**:554–66. https://doi.org/10.1016/j.molmed.2023.03.007.
3. Martinez-Jimenez F, Muinos F, Sentis I. *et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* 2020;**20**:555–72. https://doi.org/10.1038/s41568-020-0290-x.
4. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell* 2011;**144**:646–74. https://doi.org/10.1016/j.cell.2011.02.013.
5. Heo YJ, Hwa C, Lee GH. *et al.* Integrative multi-omics approaches in cancer research: From biological networks to clinical subtypes. *Mol Cells* 2021;**44**:433–43. https://doi.org/10.14348/molcells.2021.0042.
6. Liu SH, Shen PC, Chen CY. *et al.* DriverDBv3: A multi-omics database for cancer driver gene research. *Nucleic Acids Res* 2020;**48**:D863–70. https://doi.org/10.1093/nar/gkz964.
7. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 2009;**25**:2906–12. https://doi.org/10.1093/bioinformatics/btp543.
8. Kim D, Joung JG, Sohn KA. *et al.* Knowledge boosting: A graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. *J Am Med Inform Assoc* 2015;**22**:109–20. https://doi.org/10.1136/amiajnl-2013-002481.
9. Tini G, Marchetti L, Priami C. *et al.* Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief Bioinform* 2019;**20**:1269–79. https://doi.org/10.1093/bib/bbx167.
10. Oh JH, Choi W, Ko E. *et al.* PathCNN: Interpretable convolutional neural networks for survival prediction and pathway analysis applied to glioblastoma. *Bioinformatics* 2021;**37**:i443–50. https://doi.org/10.1093/bioinformatics/btab285.
11. Wang T, Shao W, Huang Z. *et al.* MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun* 2021;**12**:3445. https://doi.org/10.1038/s41467-021-23774-w.
12. Kesimoglu ZN, Bozdag S. SUPREME: Multiomics data integration using graph convolutional networks. *NAR Genom Bioinform* 2023;**5**:lqad063. https://doi.org/10.1093/nargab/lqad063.
13. Xia J, Benner MJ, Hancock RE. NetworkAnalyst–integrative approaches for protein-protein interaction network analysis and visual exploration. *Nucleic Acids Res* 2014;**42**:W167–74. https://doi.org/10.1093/nar/gku443.
14. Shixiang Wang XL. The UCSCXenaTools R package: A toolkit for accessing genomics data from UCSC Xena platform, from cancer multi-omics to single-cell RNA-seq. *Journal of Open Source Software* 2019;**4**:1627. https://doi.org/10.21105/joss.01627.
15. de Klein N, Tsai EA, Vochteloo M. *et al.* Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat Genet* 2023;**55**:377–88. https://doi.org/10.1038/s41588-023-01300-6.
16. Vivian J, Rao AA, Nothaft FA. *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nat Biotechnol* 2017;**35**:314–6. https://doi.org/10.1038/nbt.3772.
17. Chen HM, MacDonald JA. Network analysis of TCGA and GTEx gene expression datasets for identification of trait-associated biomarkers in human cancer. *STAR Protoc* 2022;**3**:101168. https://doi.org/10.1016/j.xpro.2022.101168.
18. Law CW, Chen Y, Shi W. *et al.* Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:R29. https://doi.org/10.1186/gb-2014-15-2-r29.
19. Mermel CH, Schumacher SE, Hill B. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;**12**:R41. https://doi.org/10.1186/gb-2011-12-4-r41.
20. Ellrott K, Bailey MH, Saksena G. *et al.* Scalable Open Science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst* 2018;**6**:271–281.e7. https://doi.org/10.1016/j.cels.2018.03.002.
21. Zhou W, Triche TJ Jr, Laird PW. *et al.* SeSAMe: Reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res* 2018;**46**:e123. https://doi.org/10.1093/nar/gky691.
22. Song MA, Tiirikainen M, Kwee S. *et al.* Elucidating the landscape of aberrant DNA methylation in hepatocellular carcinoma. *PLoS One* 2013;**8**:e55761. https://doi.org/10.1371/journal.pone.0055761.
23. Malta TM, Sokolov A, Gentles AJ. *et al.* Machine learning identifies Stemness features associated with oncogenic dedifferentiation. *Cell* 2018;**173**:e315.
24. Nguyen L, WMM J, Van Hoeck A. *et al.* Pan-cancer landscape of homologous recombination deficiency. *Nat Commun* 2020;**11**:5584. https://doi.org/10.1038/s41467-020-19406-4.
25. Wolf DM, Lenburg ME, Yau C. *et al.* Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. *PLoS One* 2014;**9**:e88309. https://doi.org/10.1371/journal.pone.0088309.
26. Bajpai AK, Davuluri S, Tiwary K. *et al.* Systematic comparison of the protein-protein interaction databases from a user's perspective. *J Biomed Inform* 2020;**103**:103380. https://doi.org/10.1016/j.jbi.2020.103380.
27. Hou J, Ye X, Feng W. *et al.* Distance correlation application to gene co-expression network analysis. *BMC Bioinformatics* 2022;**23**:81. https://doi.org/10.1186/s12859-022-04609-x.
28. CE JBBK, Yamins D, Cox DD. Hyperopt: A python library for model selection and hyperparameter optimization. *Computational Science & Discovery* 2015;**8**:014008. https://doi.org/10.1088/1749-4699/8/1/014008.
29. Guo H, Lv X, Li Y. *et al.* Attention-based GCN integrates multi-omics data for breast cancer subtype classification and patient-specific gene marker identification. *Brief Funct Genomics* 2023;**22**:463–74. https://doi.org/10.1093/bfgp/elad013.
30. Holderfield M, Deuker MM, McCormick F. *et al.* Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. *Nat Rev Cancer* 2014;**14**:455–67. https://doi.org/10.1038/nrc3760.
31. Velghe AI, Van Cauwenberghe S, Polyansky AA. *et al.* PDGFRA alterations in cancer: Characterization of a gain-of-function

V536E transmembrane mutant as well as loss-of-function and passenger mutations. *Oncogene* 2014;**33**:2568–76. https://doi.org/10.1038/onc.2013.218.

32. Tharin Z, Richard C, Derangere V. *et al.* PIK3CA and PIK3R1 tumor mutational landscape in a pan-cancer patient cohort and its association with pathway activation and treatment efficacy. *Sci Rep* 2023;**13**:4467. https://doi.org/10.1038/s41598-023-31593-w.
33. Chen X, Zhang T, Su W. *et al.* Mutant p53 in cancer: From molecular mechanism to therapeutic modulation. *Cell Death Dis* 2022;**13**:974. https://doi.org/10.1038/s41419-022-05408-1.
34. Ma F, Laster K, Dong Z. The comparison of cancer gene mutation frequencies in Chinese and U.S. patient populations. *Nat Commun* 2022;**13**:5651. https://doi.org/10.1038/s41467-022-33351-4.
35. Szklarczyk D, Gable AL, Lyon D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**:D607–13. https://doi.org/10.1093/nar/gky1131.
36. Elefsinioti A, Sarac OS, Hegele A. *et al.* Large-scale de novo prediction of physical protein-protein association. *Mol Cell Proteomics* 2011;**10**:M111.010629. https://doi.org/10.1074/mcp.M111.010629.
37. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559. https://doi.org/10.1186/1471-2105-9-559.
38. Yin L, Cai Z, Zhu B. *et al.* Identification of key pathways and genes in the dynamic progression of HCC based on WGCNA. *Genes (Basel)* 2018;**9**:92. https://doi.org/10.3390/genes9020092.
39. Jha K, Saha S, Singh H. Prediction of protein-protein interaction using graph neural networks. *Sci Rep* 2022;**12**:8360. https://doi.org/10.1038/s41598-022-12201-9.
40. Colaprico A, Olsen C, Bailey MH. *et al.* Interpreting pathways to discover cancer driver genes with moonlight. *Nat Commun* 2020;**11**:69. https://doi.org/10.1038/s41467-019-13803-0.
41. Zhang W, Shang X, Liu N. *et al.* ANK2 as a novel predictive biomarker for immune checkpoint inhibitors and its correlation with antitumor immunity in lung adenocarcinoma. *BMC Pulm Med* 2022;**22**:483. https://doi.org/10.1186/s12890-022-02279-2.
42. Gu Y, Tang S, Wang Z. *et al.* A pan-cancer analysis of the prognostic and immunological role of beta-actin (ACTB) in human cancers. *Bioengineered* 2021;**12**:6166–85. https://doi.org/10.1080/21655979.2021.1973220.
43. Murphy SJ, Karnes RJ, Kosari F. *et al.* Integrated analysis of the genomic instability of PTEN in clinically insignificant and significant prostate cancer. *Mod Pathol* 2016;**29**:143–56. https://doi.org/10.1038/modpathol.2015.136.
44. Cheng SM, Lin TY, Chang YC. *et al.* YM155 and BIRC5 down-regulation induce genomic instability via autophagy-mediated ROS production and inhibition in DNA repair. *Pharmacol Res* 2021;**166**:105474. https://doi.org/10.1016/j.phrs.2021.105474.

Briefings in Bioinformatics, 2025, **26(4)**, bbaf412
https://doi.org/10.1093/bib/bbaf412
Problem Solving Protocol