PLOS Computational Biology

RESEARCH ARTICLE

Multiplex networks-based directed graph neural network for cancer driver gene identification

Pingting Li, Minzhu Xie*

College of Information Science and Engineering, Hunan Normal University, Changsha, China

* xieminzhu@hunnu.edu.cn

Abstract

Identifying cancer driver genes is crucial in precision oncology. Most existing methods rely on a single interaction network to capture gene relationships. However, with the increasing availability of multi-omics and biological network data, integrating multiplex networks offers a more comprehensive representation of the complex and directional regulatory interactions among genes. Moreover, the number of validated cancer driver genes remains small compared with the vast number of unlabeled genes, leading to label scarcity and class imbalance. To address these limitations, we propose a multiplex networks-based directed graph neural network (MNDGNN). The model learns gene representations on multiplex networks with multi-omics data through directed graph convolution, which integrates neighbor diversity and degree diversity. We also incorporate data augmentation combining positive-sample augmentation with negative-sample inference to mitigate label scarcity. Experimental results show that the proposed method achieves better predictive performance and robustness than existing state-of-the-art methods. The predicted cancer driver genes are significantly enriched in cancer-related pathways and exhibit extensive interactions with known cancer driver genes, offering a new perspective for cancer driver gene discovery and the design of therapeutic strategies.

Author summary

Cancer genomes often contain many mutations, but only a small fraction actively promote tumor growth. Therefore, distinguishing driver mutations from the vast background of passenger mutations is a critical task for understanding disease mechanisms and developing targeted therapies. Although large-scale sequencing has enabled the discovery of hundreds of cancer driver genes, many of these genes remain difficult to interpret because relevant evidence is scattered across different data types and biological interaction networks, and only a limited set has been experimentally validated. In this study, we develop a computational approach that integrates multi-omics data with multiplex biological interaction networks, rather than relying on a single network. We also incorporate

OPEN ACCESS

Citation: Li P, Xie M (2026) Multiplex networks-based directed graph neural network for cancer driver gene identification. PLoS Comput Biol 22(5): e1014275. https://doi.org/10.1371/journal.pcbi.1014275

Editor: Shugang Zhang, Ocean University of China, CHINA

Received: January 13, 2026

Accepted: April 27, 2026

Published: May 14, 2026

Data availability statement: Our code and data are publicly available in the GitHub repository: https://github.com/PTINDEX/MNDGNN.

Funding: This work was supported by grants from the National Natural Science Foundation of China [62172028, 61772197].

Competing interests: The authors have declared that no competing interests exist.

directionality in regulatory relationships to better reflect how signals propagate through gene networks. In addition, we employ a data augmentation strategy to facilitate effective learning under label scarcity. Our method improves predictive performance over existing approaches and prioritizes candidate cancer driver genes that are strongly connected to known cancer driver genes and enriched in cancer-relevant pathways, providing a practical shortlist for downstream experimental validation and therapeutic target discovery.

Introduction

Cancer is a class of diseases characterized by the uncontrolled proliferation of somatic cells. It arises from driver mutations in the human genome, which confer a selective growth advantage to the affected cells and thereby promote tumorigenesis. Genes harboring such mutations are called cancer driver genes (CDGs) [1–4]. As cancer driver genes are under positive selection and contribute to the disruption of key cellular functions, identifying these genes facilitates cancer diagnosis and targeted therapies, and plays a critical role in precision oncology [1].

In recent years, computational methods for identifying cancer driver genes have increased. This growth relies on rich multi-omics resources from databases such as TCGA and biological networks modeling complex molecular interactions. Meanwhile, deep learning has advanced rapidly. Graph neural networks (GNNs) show a strong ability to capture topological information from a network, leading to their widespread application in this field. EMOGI [5] is an explainable method based on graph convolutional networks (GCNs) for cancer driver gene identification through the integration of multi-omics data and protein-protein interaction (PPI) network. MTGCN [6] introduces a multi-task learning graph convolutional network based on a ChebNet [7] variant, enhancing cancer driver gene identification through joint optimization of node classification and link prediction tasks. ECD-CDGI [8] primarily employs an encoder based on energy-constrained diffusion and attention mechanisms, effectively capturing complex gene dependencies. In deepCDG [9], a shared-parameter GCN encoder with an attention layer supports cross-omic integration to improve cancer driver gene identification. DGMP [10] predicts cancer driver genes by integrating a directed graph convolutional network and a multilayer perceptron to learn gene features from a gene regulatory network and multi-omics data.

However, most current methods are limited to analyzing individual networks, capturing only a single type of interaction. This not only overlooks the multifaceted functions of genes across different biological regulatory processes, but may also lead to an overrepresentation of network-specific noise. Recently, several models integrating multiplex networks have been proposed to identify cancer driver genes. MRNGCN [11] leverages three gene relationship networks to propose an identification method integrating a heterogeneous graph convolutional network with a self-attention mechanism. MMGN [12] integrates multiplex networks and multi-omics data through graph neural networks with negative sample inference to identify cancer driver genes.

Both MRNGCN and MMGN represent gene regulatory processes and signaling pathways as undirected networks. However, since these processes are inherently directional, such representations may result in the loss of regulatory logic and reduced prediction accuracy [13].

While some studies, such as DGMP, have employed a directed graph convolutional network to model gene regulatory networks, these approaches generally apply uniform, layer-wise directional weights. This results in equal importance being assigned to both incoming and outgoing messages, disregarding variations in local structure. Furthermore, degree information is often reduced to a simple normalization factor. Consequently, these methods fail to account for node-level differences in the importance of directional neighbors and overlook the structural insights provided by in-degree and out-degree. This ultimately limits the model's ability to capture fine-grained directional weighting dynamics [14].

Regarding the training data, although resources like the Network of Cancer Genes (NCG) [15] provide known cancer driver genes, their number is relatively small compared to the quantity of unlabeled genes. Furthermore, reliable datasets for non-cancer driver genes (NCDGs) are currently unavailable. This label imbalance presents challenges for model training and validation, potentially affecting classification performance [13].

To address these challenges, we propose the MNDGNN model for cancer driver gene identification. Our approach leverages multiplex networks to capture diverse types of molecular interactions and reduce noise associated with any single network. Furthermore, it employs a directed graph neural network that incorporates both neighbor diversity and degree diversity to enable fine-grained modeling of directional information. This enhances node discriminability and allows for more comprehensive use of information from multiplex biological networks. To mitigate label imbalance caused by the limited availability of known cancer driver genes and the absence of confirmed non-cancer driver genes, we introduce a data augmentation strategy that combines positive-sample augmentation with negative-sample inference. Overall, the main contributions of this paper are summarized as follows:

(1) We design a multiplex networks-based directed graph convolutional network (MDGCN). Each convolutional layer of MDGCN captures neighbor diversity and degree diversity to extract directional information of nodes in biological networks, effectively integrating both directed and undirected multiplex networks.

(2) During data augmentation, we employ low information entropy (IE) and radial basis function (RBF) [16] spectral clustering for positive sample enhancement. Additionally, an internal contrastive learning (ICL) [17] model is utilized for negative sample inference to maintain class balance.

(3) The experimental results demonstrate that MNDGNN outperforms other state-of-the-art methods in terms of AUROC, AUPRC, and F1 score on the pan-cancer dataset. It not only identifies known cancer driver genes but also uncovers potential cancer driver genes and candidate therapeutic targets.

Materials and methods

Materials

Multi-omics data. We collected pan-cancer multi-omics data for 16 cancer types from The Cancer Genome Atlas [18] (https://portal.gdc.cancer.gov/), encompassing gene mutation, DNA methylation, and gene expression data. For each cancer type, the mutation frequency of a gene was calculated as the number of non-silent single nucleotide variants in that gene divided by its exonic gene length. The extent of DNA methylation alteration was defined as the average difference in methylation signals between tumor samples and matched normal samples for a specific cancer type. The differential expression level of each gene in a given cancer type was first quantified as the $\log_2$ fold change between its expression values in cancer versus matched normal samples, and then averaged across all available samples. Ultimately, we concatenated these three types of omics data for each gene across all cancer types and performed z-score

normalization. The known cancer driver genes, serving as positive samples, are derived from the following resources: NCG, COSMIC Cancer Gene Census [19], and DigSee [20], comprising a total of 668 genes.

Biological networks. We collected six biological networks. The specific details are as follows:

(1) PPI network: The protein-protein interaction network is an undirected network built from physical or functional interactions between proteins. Sourced from the ConsensusPathDB [21], it comprises 9,691 nodes and 365,138 edges.

(2) Protein complexes network: The network is an undirected graph representation of polypeptide chains connected by disulfide bonds and other protein interactions. Based on the CORUM [22] relationships, the graph is defined by 1,810 nodes and 25,714 edges.

(3) KEGG pathway network: KEGG pathway network describes the functional relationships established through biochemical reactions and signal transduction among gene products within biological pathways. Using the Pathview [23] tool, we integrated pathways from the KEGG [24] database. This process resulted in a directed network containing 3,007 nodes and 58,946 edges.

(4) RegNetwork: Regulatory interactions of transcription factor-transcription factor (TF-TF) and transcription factor-gene (TF-gene) pairs were systematically retrieved from RegNetwork [25] database to construct a directed regulatory network comprising 9,539 TFs and genes and 75,112 regulatory edges.

(5) DawnNet: DawnNet is a directed gene association network derived from DawnRank [26]. It was created by consolidating redundant genes into single entries and combining their corresponding edges. The resulting network contains 6,153 genes and 109,580 directed regulatory relationships.

(6) Kinase-substrate network: The kinase-substrate network primarily describes the intricate relationship where kinases regulate cellular signaling and function through the phosphorylation of their substrates. We obtained the kinase-substrate dataset curated by Wiredja et al. [27] from the PhosphoSitePlus [28] database. This directed network comprises 1,934 nodes and 4,532 edges.

Overview of MNDGNN

As shown in Fig 1, MNDGNN is a deep multi-omics integration framework built on multi-network directed graph convolution for identifying cancer driver genes. The process begins with preprocessing multi-omics data and six types of biological networks, which are subsequently input into a MDGCN. The MDGCN integrates neighbor diversity and degree diversity at each convolutional layer to learn node representations with explicit directional information. At this stage, the node representations are learned only for known CDGs and unknown genes. Subsequently, a two-layer graph convolutional module is employed to compute the IE of each node, followed by RBF spectral clustering on the known CDGs to select a high-confidence pseudo-labeled positive sample set from the unknown genes. In the next step, ICL is used to extract high-confidence non-cancer driver gene representations. Finally, the augmented gene representations are delivered to another MDGCN and a linear classifier for the prediction of cancer driver genes.

Data preprocessing

We first integrate multi-omics data, including gene mutation, DNA methylation, and gene expression, into six original biological networks by assigning these features to the corresponding gene nodes of the networks. This process yields, for each network, a multi-omics feature matrix X of dimensions $n \times f$, where n is the number of genes and the feature dimension is $f = 48$.

To enable multiplex graphs contrastive pretraining, a corrupted multi-omics feature matrix $\tilde{X}$ is generated for each network by randomly permuting the rows of the original matrix X. These two sets, i.e., the original networks and their corrupted counterparts, are subsequently used as paired inputs to the MDGCN for contrastive representation learning.

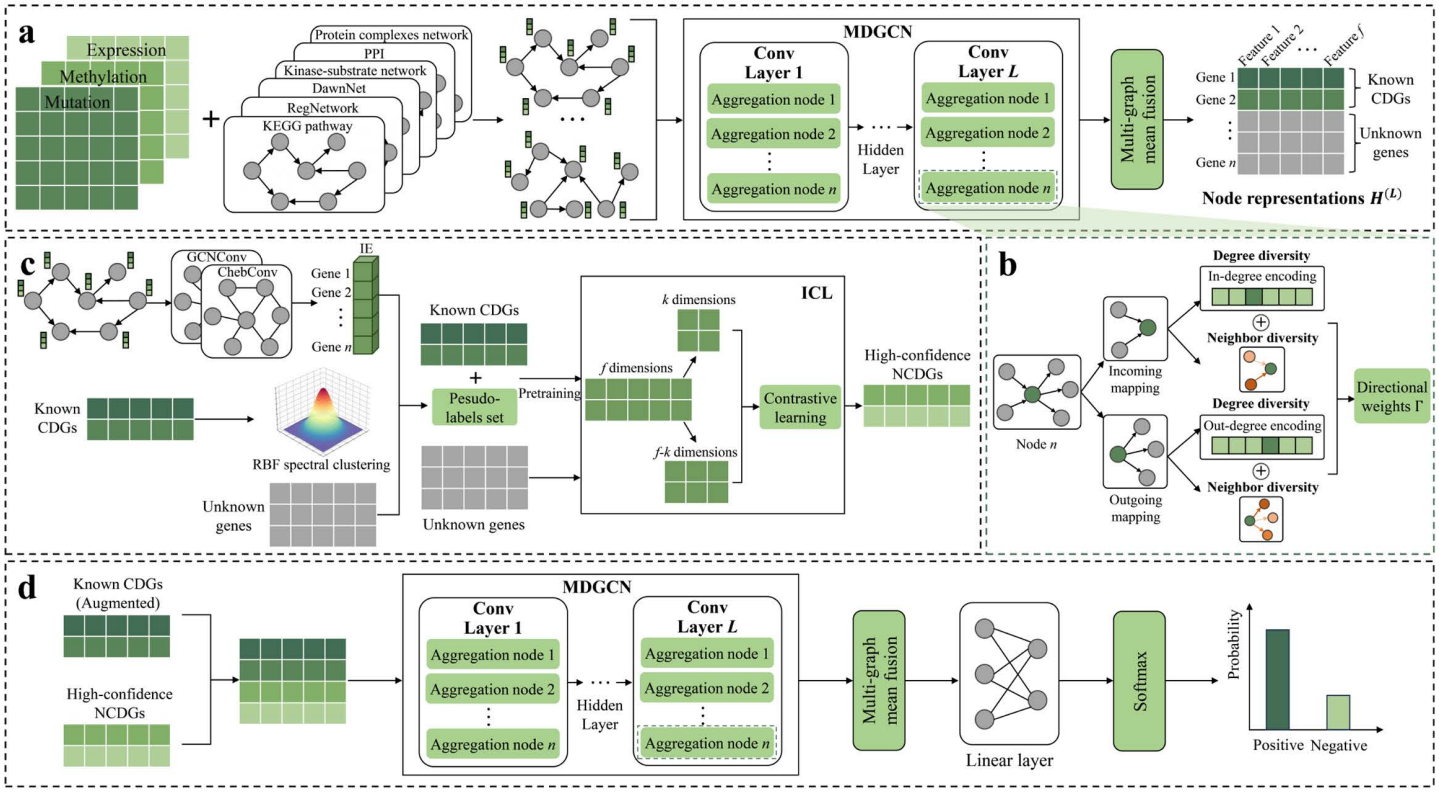

Fig 1. The architecture overview of MNDGNN. (a) Data preprocessing and feature extraction: The six biological networks incorporated with the multi-omics data are preprocessed and subsequently input into the first MDGCN to learn gene representations. **(b) Directional encoding process:** Each convolutional layer encodes directional information for node updates in the MDGCN. **(c) Data augmentation:** The model uses low information entropy and RBF-based spectral clustering to construct a pseudo-labeled set, and then applies an ICL model to identify high-confidence non-cancer driver genes. **(d) Prediction of cancer driver genes:** MNDGNN is trained on the augmented gene representations, and a linear classifier is finally applied to predict cancer driver genes.

https://doi.org/10.1371/journal.pcbi.1014275.g001

MDGCN

A MDGCN consists of several multiplex networks-based directed graph convolutional layers. Each layer takes M graphs (i.e., networks) as input, and undirected edges are treated as bidirectional. For the m-th directed graph $G^{(m)} = (V, E^{(m)}, X)$, $V = \{v_i \mid i = 1, 2, \ldots, n\}$ is the shared node (i.e., gene) set, $E^{(m)} \subseteq V \times V$ is the edge set of the graph, and $X \in \mathbb{R}^{n \times f}$ denotes the multi-omics feature matrix mentioned above. The adjacency matrix of the graph is denoted by $A^{(m)} \in \{0, 1\}^{n \times n}$. The outgoing neighbor set of the i-th node v_i is defined as $N_{i,\rightarrow}^{(m)} = \{v_j \mid (v_i \rightarrow v_j) \in E^{(m)}\}$. For graph $G^{(m)}$, we compute a diagonal out-degree matrix and a diagonal in-degree matrix, where $(D_{\rightarrow}^{(m)})_i = \sum_j (A_{ij}^{(m)})$ and $(D_{\leftarrow}^{(m)})_j = \sum_i (A_{ij}^{(m)})$. Then a symmetric normalized directed out-neighbor matrix is defined as $S_{\rightarrow}^{(m)} = (D_{\rightarrow}^{(m)})^{-\frac{1}{2}} A^{(m)} (D_{\leftarrow}^{(m)})^{-\frac{1}{2}}$. Similarly, we can define $N_{i,\leftarrow}^{(m)}$ and $S_{\leftarrow}^{(m)}$.

In each convolutional layer of MDGCN, the core operation is to aggregate neighborhood information for each node and update its representation based on the aggregated information. This process corresponds to the directional encoding process shown in Fig 1. Specifically, this module characterizes the directional properties of nodes through neighbor diversity and degree diversity.

Neighbor diversity. Under the local homophily assumption, each node is expected to share the class label most common among either its out-neighbors or in-neighbors. Dirichlet energy [29] is a common measure of feature discrepancy between adjacent nodes. To quantify the discrepancy between individual nodes and their respective

neighborhoods, we compute directional Dirichlet energy for each graph m at the node level. Let $H^{(l)} \in \mathbb{R}^{n\times f}$ denotes the node representations at layer l. The in-Dirichlet energy of individual nodes at the l-th layer is defined as follows:

$$e_{\leftarrow}^{(l,m)} = (I + S_{\leftarrow}^{(m)})(H^{(l)} \odot H^{(l)}) - 2\left(((I + S_{\leftarrow}^{(m)})H^{(l)}) \odot H^{(l)} - H^{(l)} \odot H^{(l)}\right), \tag{1}$$

where $e_{\leftarrow}^{(l,m)} \in \mathbb{R}^{n\times f}$, whose i-th row corresponds to the in-Dirichlet energy vector of node i, I is the identity matrix, and $\odot$ denotes the element-wise product. Analogously, the out-Dirichlet energy at the l-th layer is defined as follows:

$$e_{\rightarrow}^{(l,m)} = (I + S_{\rightarrow}^{(m)})(H^{(l)} \odot H^{(l)}) - 2\left(((I + S_{\rightarrow}^{(m)})H^{(l)}) \odot H^{(l)} - H^{(l)} \odot H^{(l)}\right). \tag{2}$$

A larger energy value indicates a greater discrepancy between a node and its neighbors in the corresponding direction, signifying a higher likelihood of their belonging to different classes.

Degree diversity. Degree information can enhance a GNN's ability to distinguish nodes. For each graph m, we first compute the in-degree and out-degree for each node from the adjacency matrix. Then, we use these degree values as indices to retrieve corresponding embeddings from trainable embedding matrices. Specifically, we construct two learnable embeddings for in-degree and out-degree embeddings, denoted $\mathrm{Deg}_{\leftarrow}^{(m)} \in \mathbb{R}^{n\times f}$ and $\mathrm{Deg}_{\rightarrow}^{(m)} \in \mathbb{R}^{n\times f}$.

To better characterize node representations and enhance the discrimination among structurally distinct nodes, we fuse neighbor diversity with degree diversity to obtain an adaptive directional weight for each node:

$$\begin{cases} q_{\leftarrow}^{(l,m)} = \left(-e_{\leftarrow}^{(l,m)} + \mathrm{Deg}_{\leftarrow}^{(m)}\right) w_{\leftarrow}^{(l)} + b_{\leftarrow}^{(l)}, \\ q_{\rightarrow}^{(l,m)} = \left(-e_{\rightarrow}^{(l,m)} + \mathrm{Deg}_{\rightarrow}^{(m)}\right) w_{\rightarrow}^{(l)} + b_{\rightarrow}^{(l)}, \end{cases} \tag{3}$$

where $w_{\leftarrow}^{(l)} \in \mathbb{R}^{f\times 1}$ and $b_{\leftarrow}^{(l)} \in \mathbb{R}$ are learnable parameters for the incoming direction at the l-th layer, and $q_{\leftarrow}^{(l,m)} \in \mathbb{R}^{n\times 1}$. The formulation for the outgoing direction is defined analogously.

Let $\tau^{(l)}$ be a learnable adaptive temperature. The directional weights are normalized by a softmax function and the formula is as follows:

$$\mathrm{diag}\left(\Gamma_{\rightarrow}^{(l,m)}\right) = \frac{\exp\left(q_{\rightarrow}^{(l,m)}/\tau^{(l)}\right)}{\exp\left(q_{\rightarrow}^{(l,m)}/\tau^{(l)}\right) + \exp\left(q_{\leftarrow}^{(l,m)}/\tau^{(l)}\right)}, \tag{4}$$

where $\mathrm{diag}(\Gamma_{\rightarrow}^{(l,m)}) \in \mathbb{R}^{n\times 1}$ denotes the vector of diagonal entries of the diagonal matrix $\Gamma_{\rightarrow}^{(l,m)} \in \mathbb{R}^{n\times n}$. This normalization yields complementary directional weights, i.e., $\Gamma_{\leftarrow}^{(l,m)} = I - \Gamma_{\rightarrow}^{(l,m)}$.

Since multiplex graphs are used as input, the multi-graph mean fusion module in Fig 1 aggregates messages from all M graphs and updates the node representation $H^{(l+1)}$ at the $(l+1)$-th layer by averaging their contributions according to Eq (5).

$$H^{(l+1)} = \alpha H^{(l)} + \frac{1}{M}\sum_{m=1}^{M}\left(\Gamma_{\rightarrow}^{(l,m)} S_{\rightarrow}^{(m)} H^{(l)} W_{\rightarrow}^{(l)} + \Gamma_{\leftarrow}^{(l,m)} S_{\leftarrow}^{(m)} H^{(l)} W_{\leftarrow}^{(l)}\right), \tag{5}$$

where the hyperparameter α retains information from the previous layer. $\Gamma_{\rightarrow}^{(l,m)}$ is diagonal with $\mathrm{diag}(\Gamma_{(1,1)\rightarrow}^{(l,m)}, \Gamma_{(2,2)\rightarrow}^{(l,m)}, \ldots, \Gamma_{(n,n)\rightarrow}^{(l,m)})$, where each diagonal element encodes the fused neighbor and degree diversity. $W_{\rightarrow}^{(l)}$ is a learnable linear transformation matrix at the l-th layer used to transform the aggregation result in the outgoing direction.

In summary, for each node in each network, neighbor diversity is assessed via directional Dirichlet energy, while degree embeddings are constructed from the node's in-degree and out-degree. These two components are fused to compute directional weights for each node. For both the original and corrupted network sets, the neighbor feature embeddings based on outgoing and incoming directions are weighted and aggregated. Finally, the resulting representations are averaged across all graphs to obtain the updated node representations.

Data augmentation

To mitigate label scarcity and overfitting, we apply data augmentation in two stages: positive-sample augmentation and negative-sample inference.

Positive-sample augmentation. We perform positive pseudo-label augmentation by combining entropy-based confidence filtering with a spectral clustering constraint based on the RBF kernel. Specifically, we first obtain node-wise class probabilities using a two-layer graph convolution, and then calculate the prediction entropy for each unlabeled node and retain low-entropy nodes as high-confidence candidates. Next, we apply a spectral clustering constraint based on the RBF kernel in the feature space, thereby filtering out outliers and promoting alignment with the labeled-positive distribution, resulting in a high-confidence set of positive pseudo-labels.

Concretely, we employ a Chebyshev convolutional layer to capture higher-order neighborhoods, followed by a GCN layer for normalized aggregation, and compute the class probabilities as follows:

$$P = \mathrm{softmax}\left(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \left[\sum_{k=0}^{K} T_k(\hat{L}) X \Theta_k \right] W \right), \tag{6}$$

where K denotes the Chebyshev order, and $T_k(\cdot)$ is the Chebyshev polynomial satisfying the recurrence $T_0 = I$, $T_1 = \hat{L}$, and $T_k = 2\hat{L}T_{k-1} - T_{k-2}$ for $k \geq 2$. $\hat{L}$ is the scaled graph Laplacian. $\{\Theta_k\}_{k=0}^{K}$ and W are the learnable weight matrices of the two convolutional layers. We then compute the information entropy for each unlabeled node as follows:

$$\mathcal{H}_i = -\sum_{c=1}^{C} p_{ic} \log p_{ic}, \tag{7}$$

where p_{ic} is the class probability of node i belonging to class c, C is the number of classes, $\hat{y}_i = \mathit{argmax}\, p_{ic}$ denotes a temporary label. The smaller entropy indicates a more reliable prediction. We focus on unlabeled nodes preliminarily assigned to the positive class, rank them by entropy $\mathcal{H}_i$ in ascending order, and retain a fixed number of lowest-entropy nodes to form a size-controlled candidate set.

However, relying solely on probabilities-based selection may lead to the inclusion of outlier candidates. Therefore, we impose the spectral clustering constraint based on the RBF kernel derived from the labeled positive set S_+. Using standardized feature vectors $\phi(q)$ and $\phi(u)$, the RBF kernel with σ, and $\{C_1, C_2\} \subseteq S_+$ denoting the two clusters, we construct $\kappa(q,u) = \exp\left(-\frac{\|\phi(q)-\phi(u)\|_2^2}{2\sigma^2}\right)$ to perform spectral clustering, and compute the cluster centroids:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} \phi(x_i), \quad k = 1, 2. \tag{8}$$

For each candidate unlabeled node i, we define its minimum distance to the positive cluster centroids as $d_i = \min_{k \in \{1,2\}} \|\phi(x_i) - \mu_k\|_2$. By integrating entropy-based confidence filtering with the spectral clustering constraint based on the RBF kernel, the final selection rule for newly added positive samples is defined as:

$$\mathcal{P} = \left\{ i \in \mathcal{U} \mid \hat{y}_i = c^+,\ p_{i,c^+} \geq \theta_p, d_i \leq \text{mean}\{d_i\}_{i \in S_+} + \text{std}\{d_i\}_{i \in S_+} \right\}, \tag{9}$$

where $\mathcal{P}$ is the final pseudo-label set, $\mathcal{U}$ is the set of unlabeled nodes, and c^+ is the positive class. p_{i,c^+} is the class probability that node i belongs to class c^+, and $\theta_p \in (0, 1)$ is the confidence threshold. Here, for a candidate unlabeled node, d_i represents its minimum distance to the positive cluster centroids, while $\text{mean}\{d_i\}_{i \in S_+}$ and $\text{std}\{d_i\}_{i \in S_+}$ denote the mean and standard deviation of the distances from the positive samples to the positive centroids.

Negative-sample inference. Given the absence of an authoritative database of non-cancer driver genes, negatives are inferred from positives using the ICL model. Before negative inference, multiplex graphs contrastive pretraining is performed to obtain a consensus embedding matrix for node representations.

To maximize mutual information between local and global representations and learn more discriminative features, we compute a contrastive loss on each graph while aligning global representations via a cross-graph consensus regularization. We feed all graphs into the MDGCN to obtain the layer-wise node embeddings for the m-th graph as:

$$H^{(m)} = \text{Conv}\left(X, A^{(m)}\right). \tag{10}$$

To aggregate node information at the graph level, we perform mean pooling over the positive nodes to obtain a graph-level vector $g^{(m)}$ for each graph. Then, a shared bilinear discriminator is applied on each graph to construct positive and negative pairs for contrastive learning, and the resulting contrastive loss is given by:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{n} \sum_v \log \sigma\left(\tilde{h}_{v,+}^{(m)} M^{(m)} g^{(m)}\right) - \frac{1}{n} \sum_v \log \sigma\left(1 - \tilde{h}_{v,-}^{(m)} M^{(m)} g^{(m)}\right), \tag{11}$$

where $M^{(m)}$ is a learnable matrix, $\tilde{h}_{v,+}^{(m)}$ and $\tilde{h}_{v,-}^{(m)}$ denote the embeddings of node v in the m-th graph under the positive (i.e., original) and negative (i.e., corrupted) views, respectively. To establish consistent node representations across multiple graphs, we average the node embeddings from all graphs and introduce a learnable consensus embedding matrix $Z \in \mathbb{R}^{n \times f}$. The consensus loss is as follows:

$$\mathcal{L}_{\text{consensus}} = \left\| Z - \frac{1}{M} \sum_{m=1}^{M} H_+^{(m)} \right\|_2^2 - \left\| Z - \frac{1}{M} \sum_{m=1}^{M} H_-^{(m)} \right\|_2^2, \tag{12}$$

where $H_+^{(m)}$ and $H_-^{(m)}$ denote the node embeddings of the m-th graph under the positive and negative views. By combining all contrastive losses with the consensus loss, and using β to balance the two losses, the pretraining objective is:

$$J = \sum_{m=1}^{M} \mathcal{L}_{\text{contrast}} + \beta \mathcal{L}_{\text{consensus}}. \tag{13}$$

In addition, we introduce an ICL to infer high-confidence negatives. ICL is trained on positive samples only to model the internal consistency between each local window and its complementary segment. Samples with larger inconsistency are more likely to be negatives. For each sample $x_i \in \mathbb{R}^f$, we construct sliding window pairs $\Phi(x_i) = \{(a_i^j, b_i^j)\}_{j=1}^m$ with $m = f - k + 1$, where $a_i^j \in \mathbb{R}^k$ is the j-th local window and $b_i^j \in \mathbb{R}^{f-k}$ is its complementary segment. After dual encoders F and G and normalization, we minimize the following objective under a contrastive learning framework with multiple candidates for each sample.

$$\ell(F, G, \Phi(x_i), j) = -\ln \frac{\exp\left(F^N(b_i^j) \cdot G^N(a_i^j)/\tau\right)}{\sum_{j'=1}^{m} \exp\left(F^N(b_i^j) \cdot G^N(a_i^{j'})/\tau\right)}, \tag{14}$$

where F^N and G^N denote normalized embeddings, j' indexes candidate windows, and $\tau > 0$ is the temperature parameter. During inference, we construct $\Phi(x_i)$ for a test sample x in the same manner and define the anomaly score as follows:

$$y(x) = \sum_{j} \ell\left(F, G, \Phi(x_i), j\right). \tag{15}$$

A higher $y(x)$ indicates greater abnormality and therefore a higher probability that the sample is negative. After ranking the anomaly scores in descending order, we select the top-ranked samples in a quantity equal to the number of positives as the high-confidence negative set.

Prediction of cancer driver genes

After L multiplex networks-based directed graph convolutional layers in the second MDGCN, the final representation $H^{(L)}$ is fed into a linear classifier followed by softmax to obtain the predictive distribution:

$$\hat{Y} = \mathrm{softmax}\left(H^{(L)} W_{\mathrm{cls}} + b_{\mathrm{cls}}\right), \tag{16}$$

where W_{cls} and b_{cls} denote learnable classifier parameters.

Supervised loss. Given the labeled node set $V_L \subseteq \{1, \ldots, n\}$ and their labels $y_i \in \{0, 1\}$, the cross-entropy loss is adopted:

$$\mathcal{L}_{\mathrm{sup}} = -\sum_{i \in V_L} \log \hat{Y}_{i,y_i}. \tag{17}$$

Consistency regularization loss. To mitigate learning abnormal weights that deviate from the distribution, we first average all weights associated with incoming and outgoing messages:

$$\bar{\gamma}_{\rightarrow} = \frac{1}{n} \sum_{i=0}^{n-1} \Gamma_{(i,i),\rightarrow}, \qquad \bar{\gamma}_{\leftarrow} = \frac{1}{n} \sum_{i=0}^{n-1} \Gamma_{(i,i),\leftarrow}, \tag{18}$$

where $\Gamma_{(i,i),\rightarrow}$ is obtained by averaging the i-th diagonal entry of the diagonal matrix $\Gamma_{\rightarrow}^{(l,m)}$ over all graphs and layers. Then we minimize the distances between $\bar{\gamma}_{\rightarrow}$ and $\Gamma_{(i,i),\rightarrow}$ and between $\bar{\gamma}_{\leftarrow}$ and $\Gamma_{(i,i),\leftarrow}$ using the following method.

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{n} \sum_{i=0}^{n-1} \left\| \bar{\gamma}_{\rightarrow} - \Gamma_{(i,i),\rightarrow} \right\|_2^2 + \frac{1}{n} \sum_{i=0}^{n-1} \left\| \bar{\gamma}_{\leftarrow} - \Gamma_{(i,i),\leftarrow} \right\|_2^2. \tag{19}$$

Objective function. Finally, we combine the two losses above and use the hyperparameter λ to regulate the balance between them.

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \lambda \mathcal{L}_{\mathrm{reg}}. \tag{20}$$

Implementation details of MNDGNN

Our model was implemented in Python 3.9.19 with PyTorch 2.1.2. We chose AdamW as the optimizer with a learning rate of 0.01. The MDGCN module comprised three layers ($L = 3$) with a hidden dimension of 256, a dropout rate of 0.5, and a

weight decay of 0.0. We set α = 0.5 to preserve information from the previous layer, and chose β = 0.001 and λ = 0.0003 to balance the contributions of the loss terms.

Results

All the following experiments were conducted on a computer with an Nvidia RTX 4060 GPU. Each experiment was repeated with 10 different random seeds, with early stopping applied based on validation loss. The final reported results represent the average across all runs.

Performance of MNDGNN in pan-cancer driver gene prediction

To evaluate the performance of MNDGNN, we compared our method with baseline models (GCN, GAT, SAGE, ChebNet, Dir-GNN, EMOGI, DGMP and MMGN) under ten runs of five-fold cross-validation (5CV) with area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC) and F1 score metrics. For a fair comparison, all methods use the same multiplex biological networks and the same feature matrix. Methods limited to a single graph input operate on a merged graph constructed from the multiplex networks. Methods without negative sample inference are trained with the same positive and negative samples as ours. In addition, the hyperparameters of all baselines follow those specified in their original implementations.

As shown in Table 1, MNDGNN outperforms all competing methods across all evaluation metrics, indicating its superior accuracy in identifying cancer driver genes.

Ablation experiments

To validate the contribution of different biological networks, we first conducted ablations of input networks on the pan-cancer dataset under 5CV. We removed one network at a time and trained the model using the remaining biological networks. To analyze the contributions of different components in MNDGNN, we subsequently performed directionality and data augmentation ablations. “Without directionality” means all graphs are treated as undirected. “Without positive-sample augmentation” means the positive-sample augmentation module is disabled. “Without negative-sample inference” means that we use the EMOGI negative set and randomly sample an equal number of negative genes to match the positives during training.

As shown in Table 2, our full model achieves the best performance in AUROC, AUPRC and F1 score compared with its variants. Removing any single biological network consistently degrades performance, suggesting that each network provides useful information for cancer driver gene identification. In particular, excluding the PPI network or DawnNet leads to the largest drops, indicating these two networks capture especially important informative relationships. When edge

Table 1. Predictive performance of MNDGNN compared to other baseline models on multiplex networks.

Methods	AUROC	AUPRC	F1
GCN	0.8272	0.8494	0.7625
GAT	0.7729	0.7526	0.7545
SAGE	0.7959	0.8201	0.7393
ChebNet	0.8136	0.8292	0.7549
Dir-GNN	0.8202	0.8566	0.7614
EMOGI	0.8360	0.8567	0.7766
DGMP	0.8396	0.8585	0.7911
MMGN	0.8634	0.8845	0.7981
MNDGNN	0.8780	0.8962	0.8238

https://doi.org/10.1371/journal.pcbi.1014275.t001

Table 2. The ablation experimental results of MNDGNN on multiplex networks in 5CV test.

Methods	AUROC	AUPRC	F1
Without PPI network	0.8447	0.8487	0.7947
Without protein complexes network	0.8654	0.8839	0.8137
Without KEGG pathway network	0.8661	0.8837	0.7969
Without RegNetwork	0.8703	0.8872	0.8042
Without DawnNet	0.8566	0.8756	0.8001
Without kinase-substrate network	0.8654	0.8781	0.8113
Without directionality	0.8609	0.8778	0.7944
Without positive-sample augmentation	0.8767	0.8890	0.8062
Without negative-sample inference	0.8544	0.8625	0.8018
MNDGNN	0.8780	0.8962	0.8238

https://doi.org/10.1371/journal.pcbi.1014275.t002

directionality is removed, AUROC decreases by approximately 1.7%, AUPRC by approximately 1.8%, and F1 score by nearly 3%, underscoring the importance of directionality in networks such as gene regulation. Disabling either positive-sample augmentation or negative-sample inference also reduces performance, further supporting the utility of the proposed data augmentation strategy. Overall, these ablations show that both diverse biological network inputs and the components of MNDGNN contribute to performance, validating the effectiveness of our framework for cancer driver gene identification.

Performance on independent test sets

To assess whether the model is biased toward a specific data source, we evaluated its generalization on two independent test sets derived from OncoKB [30] and ONGene [31]. For each independent set, we first removed genes that overlap with the training positives. Under this setup, genes in the independent set were treated as true positives, and all remaining genes outside this set were treated as negatives. We then computed the AUPRC for each method and did not perform any hyperparameter tuning on the independent set.

All methods exhibit relatively low AUPRC on the independent test sets due to the limited number of true positives. Nevertheless, MNDGNN achieves the highest AUPRC on both OncoKB and ONGene (Fig 2), validating its robustness and ability to generalize beyond any single curated dataset.

Performance under class imbalance

During the data augmentation stage, we used the same number of negative and positive samples to mitigate class imbalance. To further evaluate the effect of the positive-to-negative sample ratio on model performance, we set three ratios, namely 1:1, 1:1.5, and 1:2, while keeping all the other experimental data and parameters unchanged. All comparative methods were evaluated under these settings. We selected AUPRC as the evaluation metric and plotted a line chart (Fig 3) to directly compare the performance changes of different models under different sample ratios.

The results show a clear downward trend in AUPRC for all models as the proportion of negative samples increases. This indicates that class imbalance weakens the model's identification capability. Specifically, each model achieves its best performance at the 1:1 ratio. This suggests that a relatively balanced sample distribution is more conducive to learning discriminative features of the positive class. As the number of negative samples increases, the models become more likely to be dominated by the majority class, thereby reducing their ability to identify positive samples. In addition, although performance declines under class-imbalanced settings, our model still achieves the best results, demonstrating superior stability and robustness.

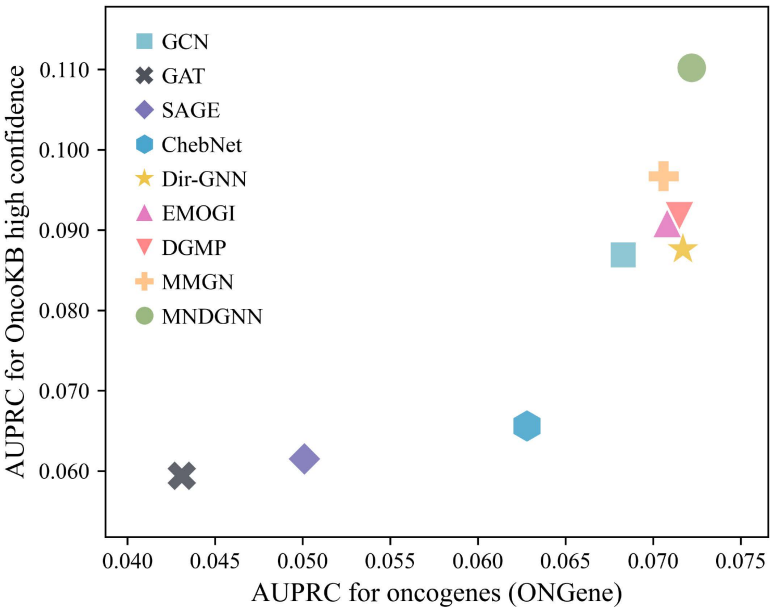

Fig 2. Performance comparison of different methods on OncoKB and ONGene.

https://doi.org/10.1371/journal.pcbi.1014275.g002

Enrichment analysis

We applied the trained model to the remaining unlabeled genes after removing training genes, ranked the candidates by their predicted probability of being CDGs, and selected the top 50 newly predicted CDGs for GO and KEGG pathway enrichment analysis. Fig 4 summarizes the enrichment results. In each bubble plot, the x-axis denotes the proportion of genes annotated to a pathway, bubble color indicates statistical significance, and bubble size reflects the number of hit genes.

Regarding biological processes (BP), the genes exhibit significant enrichment in the positive regulation of integrin-mediated signaling pathway, cell-substrate junction assembly, cell adhesion mediated by integrin, TRAIL-activated apoptotic signaling pathway, and non-canonical NF-kappaB signaling transduction. In the context of cellular components (CC), the genes are predominantly localized to migratory and adhesive structures such as focal adhesion, lamellipodium, and ruffle; receptor platforms including membrane raft and receptor complex; as well as matrix and barrier components like the basement membrane, collagen-containing extracellular matrix, extracellular matrix (ECM) and external side of the plasma membrane. For molecular functions (MF), the functional activities are concentrated on ECM-receptor interfaces and transcriptional or epigenetic regulation, encompassing integrin binding, fibronectin binding, cell adhesion molecule binding, protein tyrosine kinase binding, tumor necrosis factor receptor binding, ubiquitin protein ligase binding, histone acetyltransferase binding, histone deacetylase binding, and transcription coactivator binding. In terms of KEGG pathway, the genes are primarily enriched in ECM-receptor interaction, focal adhesion, regulation of actin cytoskeleton, apoptosis, necroptosis, proteoglycans in cancer, TNF signaling pathway, PI3K-Akt signaling pathway and pathways in cancer.

Collectively, these results indicate the identified driver genes are closely associated with known cancer pathways and may play key roles in tumor initiation and progression.

Interaction analysis between newly predicted CDGs and known CDGs

We conducted an in-depth analysis of the newly predicted CDGs. The model score, interpreted as the predicted probability that a gene is a cancer driver gene, shows a significant association with the number of interactions with known CDGs,

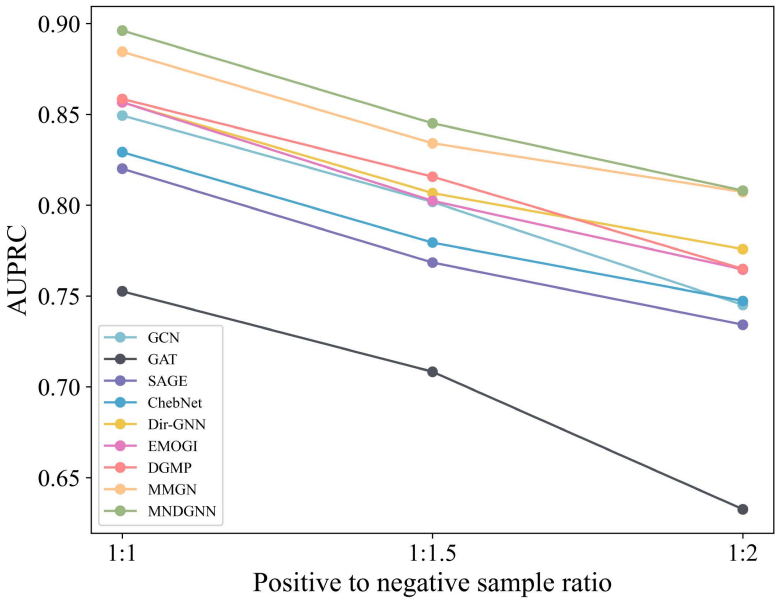

Fig 3. Performance comparison of different methods under different positive-to-negative sample ratios.

https://doi.org/10.1371/journal.pcbi.1014275.g003

as quantified by Spearman's rank correlation. Since our dataset contains both undirected and directed graphs, we consider three interaction types for each gene: **in** (in-degree interactions), **out** (out-degree interactions), and **total**. Specifically, **total** denotes the count of unique neighbors among known CDGs for a given gene, regardless of edge direction. It is computed by collecting all known CDGs linked to the gene in the graph and counting the unique entries after deduplication. For undirected graphs, these three counts are identical. For directed graphs, **total** is defined as the size of the union of in-neighbors and out-neighbors. If the same known CDG appears in both directions, it is counted only once.

For total interaction counts, we plot rank correlations between each gene's score (predicted probability) and its number of interactions with known CDGs across the six networks. Genes with higher scores tend to interact with more known CDGs. As shown in Fig 5, the strongest correlation appears in the DawnNet, followed by the KEGG pathway network and kinase-substrate network. The PPI network is dense and undirected, exhibiting a significant yet broadly dispersed correlation. The RegNetwork is weaker but still significant. The protein complexes network is the weakest, likely limited by noisier data. Overall, the density contour lines shift toward the upper right as the score increases, indicating high-scoring genes interact with more known CDGs.

For directed graphs, we compute not only total interactions, but also incoming and outgoing interaction counts. In Fig 6, the DawnNet shows strong correlations in both directions, indicating enrichment for both upstream regulation by known CDGs and downstream effects. The KEGG pathway network yields similar values in the two directions. Pathway edges are directed reactions, and both directions capture path proximity to known CDGs, so the correlations are comparable. The kinase-substrate network is asymmetric, with the incoming directions stronger than the outgoing, suggesting that kinases pointing to substrates drive the trend more. The RegNetwork is weaker in both directions yet remains significant. Directionality reveals concordance or asymmetry between incoming and outgoing interactions, which strengthens the biological interpretability of the model's predictions.

Drug sensitivity analysis

We selected the top 15 newly predicted cancer driver genes and conducted drug sensitivity analysis on the GDSC via Gene Set Cancer Analysis [32]. As shown in Fig 7, most predicted genes exhibit significant

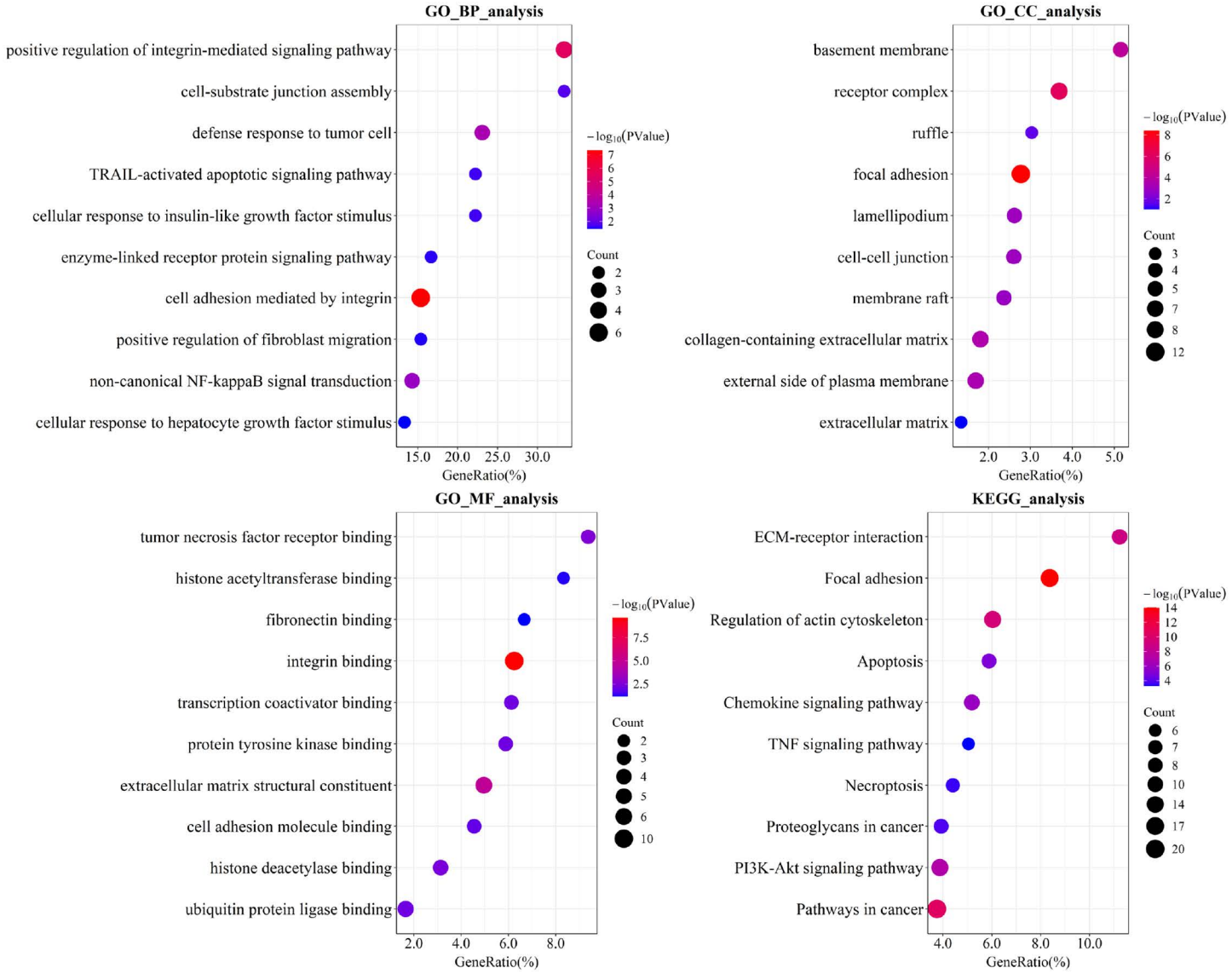

Fig 4. Enrichment analysis of the top 50 newly predicted CDGs included GO categories (BP, CC, MF) and KEGG pathway enrichment analysis.

https://doi.org/10.1371/journal.pcbi.1014275.g004

associations with sensitivity to multiple classes of clinically relevant targeted therapies. This suggests that most predicted genes indeed occupy key signaling nodes in mediating drug response. The GDSC drug sensitivity analysis supports, from a pharmacological perspective, the biological plausibility and accuracy of the cancer driver genes predicted by our model, and provides a credible basis for subsequent target validation and therapeutic strategy design.

For example, GSK1070916 selectively inhibits Aurora B/C, thereby blocking mitosis and inducing apoptosis, and exhibits broad antitumor activity [33]. YM201636 promotes epidermal growth factor receptor expression by inducing autophagy, thereby suppressing tumor growth [34]. AT-7519 induces cell death through multiple pathways and inhibits glioblastoma progression [35]. CAY10603 ameliorates diabetic nephropathy by suppressing NLRP3 inflammasome activation in renal tubular cells and macrophages [36]. OSI-027 overcomes rapamycin insensitivity via dual

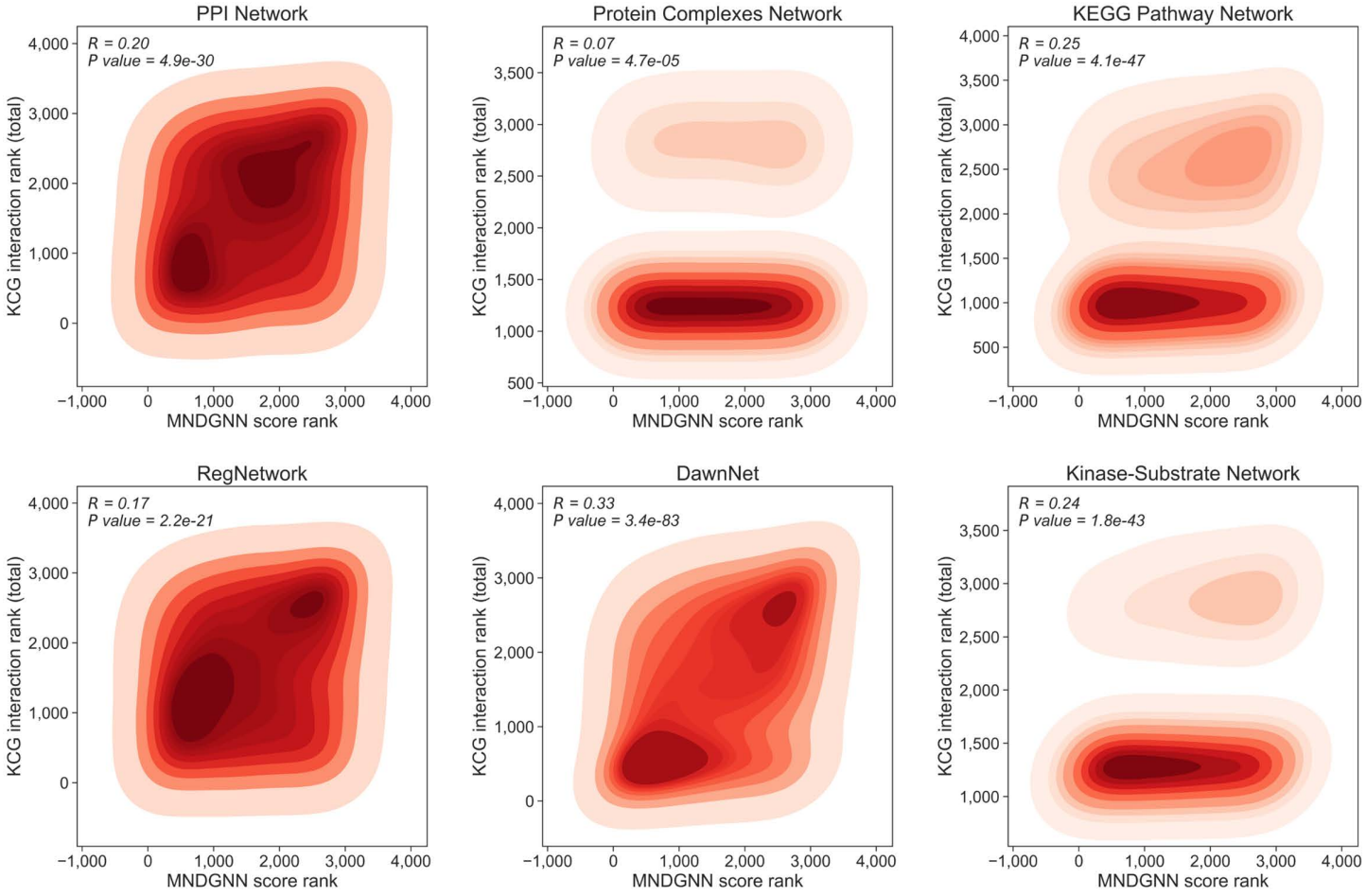

Fig 5. Spearman correlation and bivariate kernel density between the MNDGNN prediction score rank and the rank of the total number of interactions with known CDGs across different networks.

https://doi.org/10.1371/journal.pcbi.1014275.g005

inhibition of mTORC1/2 and induces tumor cell death through activation of the PI3K-AKT signaling pathway [37]. PIK-93 promotes PD-L1 ubiquitination, and in combination with anti-PD-L1 antibodies enhances T-cell activation to inhibit tumor growth [38]. QL-X-138 suppresses B-cell malignancies through dual inhibition of BTK/MNK, arresting lymphoma and leukemia cells in G0-G1 phase [39]. Finally, TPCA-1 concurrently inhibits NF-κB and STAT3 signaling in lung cancer and, in combination with tyrosine kinase inhibitors, synergistically treats non-small cell lung cancer [40].

Discussion and conclusion

In this study, we propose a model named MNDGNN for identifying cancer driver genes. Specifically, we design the MDGCN module as a stack of multiplex networks-based directed graph convolutional layers that integrate neighbor diversity and degree diversity. In each layer, edge directionality is modeled through learnable node-level weights, enabling the network to capture subtle differences among gene nodes. Furthermore, each layer fuses information from multiplex graphs, allowing the module to fully exploit the rich information contained in diverse biological networks. In addition, we augment positives using low information entropy and RBF spectral clustering, and

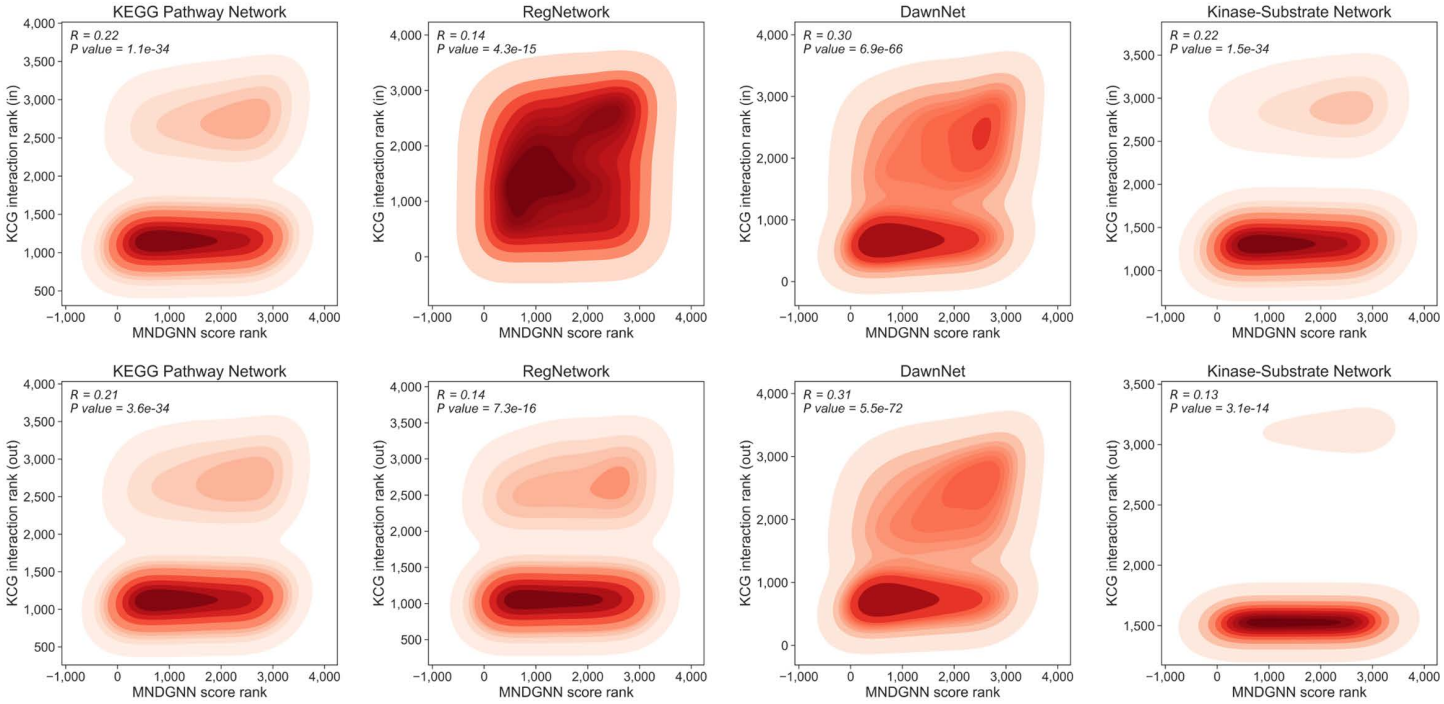

Fig 6. Spearman correlation and bivariate kernel density between the MNDGNN prediction score rank and the ranks of incoming and outgoing interaction counts with known CDGs in directed networks.

https://doi.org/10.1371/journal.pcbi.1014275.g006

we introduce the ICL to infer high-confidence negatives from positives, which mitigates label scarcity and overfitting. Finally, we train the model on the augmented data and use a linear classifier to predict candidate cancer driver genes.

The results demonstrated superior performance and stable effectiveness of the proposed model, as evidenced by comparative, ablation, and independent test set experiments on multiplex biological networks. Model predictions were cross-validated at multiple levels, highlighting its robustness across hierarchies. The top 50 newly predicted cancer driver genes were subjected to GO and KEGG enrichment analyses, revealing mechanistic associations between these predicted genes and cancer pathways. We further validated biological interpretability by examining the relationship between model scores and interactions with known CDGs, showing that high-scoring genes exhibited stronger topological proximity and functional relatedness in the relevant networks. In addition, drug sensitivity analyses based on the top 15 newly predicted cancer driver genes revealed significant associations between the majority of these genes and multiple compounds, supporting their biological plausibility and predictive accuracy and informing subsequent therapeutic strategy design.

Despite its strong performance in pan-cancer driver gene prediction, our model still has several limitations. Incomplete networks and noisy edges may affect performance. Moreover, different biological networks are likely to contribute unequally, yet in this work we adopt simple averaging to combine them, implicitly treating all networks as equally informative. Future work will focus on improving robustness to network noise and learning adaptive cross-network integration, for example by incorporating attention-based fusion or related weighting mechanisms to assign specific importance to each network and thereby further enhance model performance.

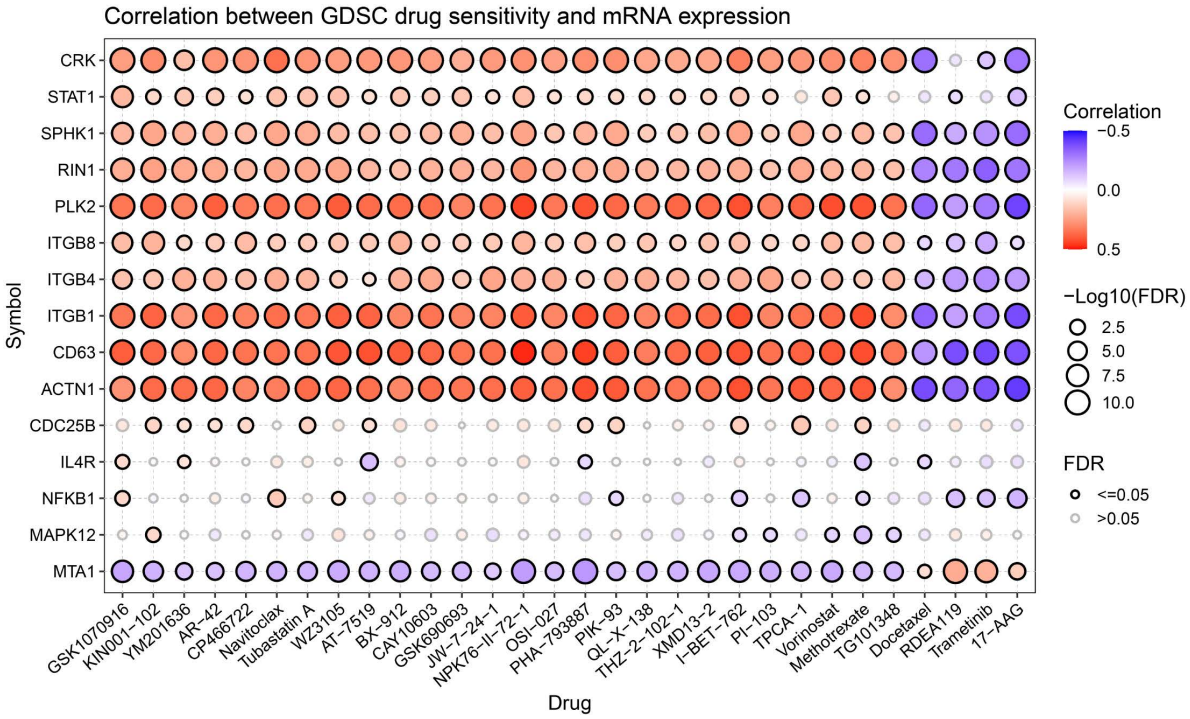

Fig 7. Correlation between GDSC drug sensitivity and mRNA expression for the top 15 newly predicted CDGs.

https://doi.org/10.1371/journal.pcbi.1014275.g007

Supporting information

S1 File. The list of the top 50 predicted cancer driver genes.
(DOCX)

Author contributions

Data curation: Pingting Li.

Investigation: Pingting Li.

Methodology: Pingting Li, Minzhu Xie.

Supervision: Minzhu Xie.

Writing – original draft: Pingting Li, Minzhu Xie.

Writing – review & editing: Pingting Li, Minzhu Xie.

References

1. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555–72. https://doi.org/10.1038/s41568-020-0290-x PMID: 32778778
2. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58. https://doi.org/10.1126/science.1235122 PMID: 23539594
3. Karnwal A, Dutta J, Al-Tawaha ARMS, et al. Genetic landscape of cancer: mechanisms, key genes, and therapeutic implications. Clin Transl Oncol. 2025;1–22. https://doi.org/10.1007/s12094-025-04019-4
4. Porta-Pardo E, Valencia A, Godzik A. Understanding oncogenicity of cancer driver genes and mutations in the cancer genomics era. FEBS Lett. 2020;594(24):4233–46. https://doi.org/10.1002/1873-3468.13781 PMID: 32239503

5. Schulte-Sasse R, Budach S, Hnisz D, Marsico A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nat Mach Intell. 2021;3(6):513–26. https://doi.org/10.1038/s42256-021-00325-y
6. Peng W, Tang Q, Dai W, Chen T. Improving cancer driver gene identification using multi-task learning on graph convolutional network. Brief Bioinform. 2022;23(1):bbab432. https://doi.org/10.1093/bib/bbab432 PMID: 34643232
7. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. Adv Neural Inform Process Syst. 2016;29. Available from: https://arxiv.org/abs/1606.09375
8. Wang T, Zhuo L, Chen Y, Fu X, Zeng X, Zou Q. ECD-CDGI: An efficient energy-constrained diffusion model for cancer driver gene identification. PLoS Comput Biol. 2024;20(8):e1012400. https://doi.org/10.1371/journal.pcbi.1012400 PMID: 39213450
9. Wu Y, Xu J, Li J, Gu J, Shang X, Li X. Deep graph convolutional network-based multi-omics integration for cancer driver gene identification. Brief Bioinform. 2025;26(4):bbaf364. https://doi.org/10.1093/bib/bbaf364 PMID: 40716043
10. Zhang S-W, Xu J-Y, Zhang T. DGMP: Identifying Cancer Driver Genes by Jointing DGCN and MLP from Multi-omics Genomic Data. Genomics Proteomics Bioinformatics. 2022;20(5):928–38. https://doi.org/10.1016/j.gpb.2022.11.004 PMID: 36464123
11. Peng W, Wu R, Dai W, Yu N. Identifying cancer driver genes based on multi-view heterogeneous graph convolutional network and self-attention mechanism. BMC Bioinformatics. 2023;24(1):16. https://doi.org/10.1186/s12859-023-05140-3 PMID: 36639646
12. Li X, Li J, Hao J, Liao X, Li M, Shang X. Multiplex Networks and Pan-Cancer Multiomics-Based Driver Gene Identification Using Graph Neural Networks. Big Data Min Anal. 2024;7(4):1262–72. https://doi.org/10.26599/bdma.2024.9020043
13. Zhang H, Lin C, Chen Y, Shen X, Wang R, Chen Y, et al. Enhancing Molecular Network-Based Cancer Driver Gene Prediction Using Machine Learning Approaches: Current Challenges and Opportunities. J Cell Mol Med. 2025;29(1):e70351. https://doi.org/10.1111/jcmm.70351 PMID: 39804102
14. Huang J, Mo Y, Hu P, et al. Exploring the role of node diversity in directed graph representation learning. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI). 2024. p. 2072-80.
15. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 2019;20(1):1. https://doi.org/10.1186/s13059-018-1612-0 PMID: 30606230
16. Buhmann MD. Radial basis functions. Acta Numerica. 2000;9:1–38. https://doi.org/10.1017/s0962492900000015
17. Shenkar T, Wolf L. Anomaly detection for tabular data with internal contrastive learning. In: International Conference on Learning Representations (ICLR). 2022. Available from: https://openreview.net/forum?id=_hszZbt46bT
18. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20. https://doi.org/10.1038/ng.2764 PMID: 24071849
19. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18(11):696–705. https://doi.org/10.1038/s41568-018-0060-1 PMID: 30293088
20. Kim J, So S, Lee H-J, Park JC, Kim J-J, Lee H. DigSee: Disease gene search engine with evidence sentences (version cancer). Nucleic Acids Res. 2013;41(Web Server issue):W510–7. https://doi.org/10.1093/nar/gkt531 PMID: 23761452
21. Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. Nat Protoc. 2016;11(10):1889–907. https://doi.org/10.1038/nprot.2016.117 PMID: 27606777
22. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. 2019;47(D1):D559–63. https://doi.org/10.1093/nar/gky973 PMID: 30357367
23. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013;29(14):1830–1. https://doi.org/10.1093/bioinformatics/btt285 PMID: 23740750
24. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173
25. Liu Z-P, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database. 2015;2015:bav095. https://doi.org/10.1093/database/bav095
26. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med. 2014;6(7):56. https://doi.org/10.1186/s13073-014-0056-8 PMID: 25177370
27. Wiredja DD, Koyutürk M, Chance MR. The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. Bioinformatics. 2017;33(21):3489–91. https://doi.org/10.1093/bioinformatics/btx415 PMID: 28655153
28. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 2012;40(Database issue):D261–70. https://doi.org/10.1093/nar/gkr1122 PMID: 22135298
29. Xu L, Chen L, Wang R, et al. Joint Feature and Differentiable k-NN Graph Learning using Dirichlet Energy. Adv Neural Inf Process Syst. 2023;36:24497–518. https://doi.org/10.48550/arXiv.2305.12396
30. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017;2017:PO.17.00011. https://doi.org/10.1200/PO.17.00011 PMID: 28890946

31. Liu Y, Sun J, Zhao M. ONGene: A literature-based database for human oncogenes. J Genet Genomics. 2017;44(2):119–21. https://doi.org/10.1016/j.jgg.2016.12.004 PMID: 28162959

32. Liu C-J, Hu F-F, Xie G-Y, Miao Y-R, Li X-W, Zeng Y, et al. GSCA: an integrated platform for gene set cancer analysis at genomic, pharmacogenomic and immunogenomic levels. Brief Bioinform. 2023;24(1):bbac558. https://doi.org/10.1093/bib/bbac558 PMID: 36549921

33. Hardwicke MA, Oleykowski CA, Plant R, Wang J, Liao Q, Moss K, et al. GSK1070916, a potent Aurora B/C kinase inhibitor with broad antitumor activity in tissue culture cells and human tumor xenograft models. Mol Cancer Ther. 2009;8(7):1808–17. https://doi.org/10.1158/1535-7163.MCT-09-0041 PMID: 19567821

34. Hou J-Z, Xi Z-Q, Niu J, Li W, Wang X, Liang C, et al. Inhibition of PIKfyve using YM201636 suppresses the growth of liver cancer via the induction of autophagy. Oncol Rep. 2019;41(3):1971–9. https://doi.org/10.3892/or.2018.6928 PMID: 30569119

35. Zhao W, Zhang L, Zhang Y, Jiang Z, Lu H, Xie Y, et al. The CDK inhibitor AT7519 inhibits human glioblastoma cell growth by inducing apoptosis, pyroptosis and cell cycle arrest. Cell Death Dis. 2023;14(1):11. https://doi.org/10.1038/s41419-022-05528-8 PMID: 36624090

36. Hou Q, Kan S, Wang Z, Shi J, Zeng C, Yang D, et al. Inhibition of HDAC6 With CAY10603 Ameliorates Diabetic Kidney Disease by Suppressing NLRP3 Inflammasome. Front Pharmacol. 2022;13:938391. https://doi.org/10.3389/fphar.2022.938391 PMID: 35910382

37. Bhagwat SV, Gokhale PC, Crew AP, Cooke A, Yao Y, Mantis C, et al. Preclinical characterization of OSI-027, a potent and selective inhibitor of mTORC1 and mTORC2: distinct from rapamycin. Mol Cancer Ther. 2011;10(8):1394–406. https://doi.org/10.1158/1535-7163.MCT-10-1099 PMID: 21673091

38. Lin C-Y, Huang K-Y, Kao S-H, Lin M-S, Lin C-C, Yang S-C, et al. Small-molecule PIK-93 modulates the tumor microenvironment to improve immune checkpoint blockade response. Sci Adv. 2023;9(14):eade9944. https://doi.org/10.1126/sciadv.ade9944 PMID: 37027467

39. Wu H, Hu C, Wang A, Weisberg EL, Chen Y, Yun C-H, et al. Discovery of a BTK/MNK dual inhibitor for lymphoma and leukemia. Leukemia. 2016;30(1):173–81. https://doi.org/10.1038/leu.2015.180 PMID: 26165234

40. Nan J, Du Y, Chen X, Bai Q, Wang Y, Zhang X, et al. TPCA-1 is a direct dual inhibitor of STAT3 and NF-κB and regresses mutant EGFR-associated human non-small cell lung cancers. Mol Cancer Ther. 2014;13(3):617–29. https://doi.org/10.1158/1535-7163.MCT-13-0464 PMID: 24401319