Improving Cancer Gene Identification by Enhancing Common Information Between the PPI Network and Gene Functional Association

Chao Deng[1,2], **Hongdong Li**[1,2*], **Jianxin Wang**[1,2,3]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[2]Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha, 410083, China
[3]Xinjiang Engineering Research Center of Big Data and Intelligent Software, School of Software, Xinjiang University, Urumqi 830091, China
{deng_chao, hongdong}@csu.edu.cn, jxwang@mail.csu.edu.cn

Abstract

Identifying cancer genes is crucial for treatment and understanding pathogenesis. Recent methods typically leverage protein-protein interaction (PPI) networks or gene functional association data from annotated gene sets. There may be some shared neighborhood structure information between these two types of gene association data. While this common information may contain more accurate gene association information, existing methods often overlook this potential. To address this gap, we introduce DISFusion, which integrates multi-omics cancer data, PPI networks, and gene functional associations to identify cancer genes. A key innovation of DISFusion is the cross-view decorrelation loss, which enhances the common information between PPI networks and gene functional associations, thereby improving prediction accuracy. Extensive experiments indicate that DISFusion outperforms state-of-the-art methods and exhibits greater generalization ability. Moreover, analysis of CPTAC pan-cancer proteomic data highlights significant associations between the 30 novel cancer genes predicted by DISFusion and multiple cancer types, underscoring its practical utility. These findings validate the effectiveness of enhancing common information and provide new insights into cancer gene identification.

Code —
https://github.com/CharlesDeng0814/DISFusion.git

Introduction

Cancer arises from the accumulation of alterations in critical genes, leading to dysregulation of the balance between cell division and apoptosis. Therefore, a key goal of cancer genomic research is to identify the genes that drive cancer initiation and progression (Garraway and Lander 2013; Lawrence et al. 2014; Alexandrov et al. 2013). The identification of cancer genes is the key to exploring the complex pathogenesis and targeted therapy of cancer (Garraway and Lander 2013; Lawrence et al. 2014; Alexandrov et al. 2013; Nofech-Mozes et al. 2023). Although the Network of Cancer Genes (NCG) (Repana et al. 2019) and the COSMIC Cancer Gene Census (CGC) (Sondka et al. 2018) have been used to annotate some cancer genes. However, the current catalog of known cancer genes is insufficient to fully account for the characteristics observed in the majority of cancer samples. Accurate identification of cancer genes from many candidate genes remains a critical challenge.

Recent computational methods for cancer gene identification can be categorized into manual feature-based, biological network-based, and functional association-based approaches. Manual feature-based methods typically use statistical results from sequencing data as rules for identifying cancer genes or as features of genes to train machine learning models, such as MutSigCV (Lawrence et al. 2013), 20/20+ (Tokheim et al. 2016) and DORGE (Lyu et al. 2020). Biological network-based methods use the association information between genes in biological networks to predict cancer genes (Reyna, Leiserson, and Raphael 2018; Barel and Herwig 2020; Schulte-Sasse et al. 2021; Peng et al. 2022). Among those methods, one of the most representative is EMOGI (Schulte-Sasse et al. 2021); it proposes integrating multi-omics features of genes and the protein-protein interaction (PPI) network with the graph neural network to identify cancer genes. Functional association-based methods use machine learning or hypergraph neural networks to extract higher-order functional association information among genes in annotated gene sets to predict cancer genes, such as predCAN (Althubaiti et al. 2019) and DISHyper (Deng et al. 2024).

Biological network and functional association-based methods typically focus on binary gene associations (e.g., PPI) or higher-order functional associations (e.g., signaling pathways). Previous studies have demonstrated that cancer genes tend to cluster in specific biological processes, hallmark signaling pathways, and interacting subnetworks (Vogelstein et al. 2013; Creixell et al. 2015; Reyna et al. 2020). This observation suggests that integrating binary and functional gene association information could enhance the accuracy of cancer gene identification. Indeed, methods like MODIG (Zhao et al. 2022) have begun to explore this integration through the use of multiple gene networks and graph attention networks. However, a significant limitation of these methods is that they ignore the information interactions between different types of gene association data. Specifically, there may be shared neighborhood structure information between PPI networks and gene functional associations. This common information could provide more accurate and im-

*Corresponding author

portant insights into gene associations, yet it is often overlooked. Therefore, our work focuses on enhancing this common information between PPI networks and gene functional associations to more accurately and comprehensively characterize gene association patterns, ultimately improving the identification of cancer genes.

We present DISFusion, which improves the performance of identifying cancer genes by enhancing the common information between PPI networks and gene functional associations. In DISFusion, we first employ graph convolutional networks (GCN) (Kipf and Welling 2017; Defferrard, Bresson, and Vandergheynst 2016) and hypergraph neural networks (HGNN) (Feng et al. 2019) to extract gene association information and generate gene embeddings from PPI networks and annotated gene sets. Then, we employ the cross-view decorrelation loss to enhance the common information between the PPI network and gene functional association of annotated gene sets, and to reduce the redundancy between gene embeddings. Finally, we fuse the gene embedding of these two association information and identify cancer genes.

The main contributions of our work are listed as follows:

- A new method named DISFusion is proposed to identify cancer genes by integrating multi-omics features, PPI networks, and gene function associations.
- DISFusion considers the common information between gene association data and designs cross-view decorrelation loss to enhance the common information and improve gene representation quality.
- DISFusion is compared with other advanced methods on pan-cancer gene data, two independent test sets, and six single cancer types. Experimental results show that DISFusion outperforms all other baseline methods.
- The analysis of pan-cancer proteomic data highlights significant associations between the 30 novel cancer genes predicted by DISFusion and multiple cancer types. This real-world validation underscores the practical utility of our approach in cancer research.

Related Works

The cancer gene identification task can be viewed as a classification problem, which aims to screen out cancer genes from tens of thousands of candidate genes using gene multi-omics data or gene association data.

Recently, there have been many computational methods for cancer gene identification. MutSigCV (Lawrence et al. 2013) identifies cancer genes by looking for genes with mutation rates that are significantly different from the background frequency distribution. The 20/20+ (Tokheim et al. 2016) method constructs multiple gene mutation features and proposes to use machine learning models to find genes with similar mutation patterns to known cancer genes. In addition, DORGE proposes to identify oncogenes and tumor suppressor genes based on elastic networks integrating multiple features such as mutations and epigenetics (Lyu et al. 2020). The above methods usually only consider the characteristics of genes but not the association between genes. To this end, many cancer gene identification methods that utilize gene networks or gene functional associations have been proposed. Methods such as Hierarchical HotNet (Reyna, Leiserson, and Raphael 2018) and NetCore (Barel and Herwig 2020) use PPI networks to partition submodules with cancer mutation signatures and identify cancer candidate genes. MTGCN (Peng et al. 2022) introduces the link prediction auxiliary task of the PPI network to the EMOGI and enhances gene feature learning through a multi-task learning framework. HGDC (Zhang et al. 2023) introduces graph diffusion to generate auxiliary networks and designs an improved message aggregation and propagation scheme to adapt to the heterophilic setting of biological networks. EMGNN (Chatzianastasis, Vazirgiannis, and Zhang 2023) designs an interpretable multi-layer graph neural network approach to identify cancer genes by leveraging multiple gene-gene interaction networks and pan-cancer multi-omics data. Meanwhile, predCAN (Althubaiti et al. 2019) proposes to predict cancer genes based on functional association data such as Gene Ontology (GO) and Mammalian Phenotype Ontology (MPO). DISHyper (Deng et al. 2024) proposes to extract higher-order functional association information of genes in different types of annotated gene sets based on hypergraph neural networks to predict genes with similar functional association patterns to cancer genes. Moreover, methods such as MODIG (Zhao et al. 2022) also attempt to integrate multiple gene association information to identify cancer genes.

Methodology

In this section, we introduce the DISFusion method for cancer gene identification tasks, designed to improve the performance of identifying cancer genes by enhancing the common information between PPI networks and annotated gene sets. As illustrated in Figure 1, the following parts will detail the procedure of DISFusion.

Extract Feature Representation of PPI Network Based on Graph Model

The PPI network represents binary association relationships between genes(Schulte-Sasse et al. 2021; Liu et al. 2023). To make full use of the binary association in the PPI network, we use the graph $G(V, E)$ to represent it, where the node $V = \{v_1, v_2, v_3, ..., v_n\}$ in the graph represents n genes, and the edge E represents m protein interactions. Then, we use the adjacency matrix of the PPI network and the pan-cancer feature matrix of genes as the input of the graph neural network to extract the binary association information and multi-omics features of the gene. In DISFusion, we use the Chebyshev graph convolutional network (Defferrard, Bresson, and Vandergheynst 2016). Compared with typical GCN, Chebyshev GCN has higher flexibility in aggregating high-order neighbor information. The single-layer Chebyshev GCN formula is as follows:

$$\mathbf{Z} = \sigma(\sum_{k=1}^{K} \mathbf{T}^{(k)} \cdot \Theta^{(k)}) \tag{1}$$

where $\mathbf{Z}$ denotes the gene feature learned by the Chebyshev GCN layer. σ denotes the activation function, where we use ReLU. Θ is the learnable parameter matrix. $\mathbf{T}^{(k)}$ represents

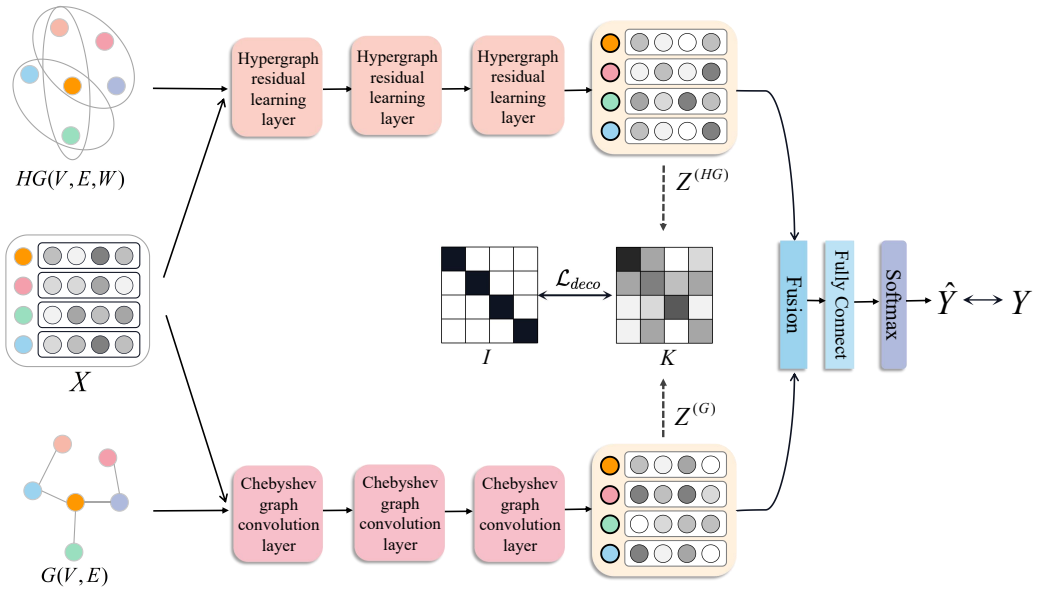

Figure 1: Illustration of DISFusion. DISFusion integrates functional association hypergraphs, pan-cancer multi-omics features, and PPI networks as inputs. It uses a three-layer Chebyshev GCN and a three-layer hypergraph residual learning module to extract gene binary association embeddings $\mathbf{Z}^{(G)}$ and gene functional association embeddings $\mathbf{Z}^{(HG)}$. Then, it applies a cross-view decorrelation loss $\mathcal{L}_{deco}$ to enhance shared information between the PPI network and functional associations. Finally, DISFusion fuses the gene embeddings and predicts the cancer risk score for each gene.

the Chebyshev polynomial of order k, which may be computed recursively by $\mathbf{T}^{(k)} = 2 \cdot \hat{L} \cdot \mathbf{T}^{(k-1)} - \mathbf{T}^{(k-2)}$ with $\mathbf{T}^{(1)} = \mathbf{X}$ and $\mathbf{T}^{(2)} = \hat{L} \cdot \mathbf{X}$. The K controls the order of the Chebyshev filter, and we set the K as 2 in our model. $\hat{L} = \frac{2L}{\lambda_{max}} - \mathbf{I}$ represents the scaled Laplacian, where L denotes the Laplacian matrix of G and λ_{max} denotes the largest eigenvalue of L. We stack three layers of Chebyshev GCN layers in DISFusion to learn the association information of genes in the PPI network and finally output the gene embedding $\mathbf{Z}^{(G)} \in \mathbb{R}^{n \times d}$, where the input of each layer is the output of the previous layer.

Extract Feature Representation of Functional Association Based on Hypergraph Model

Annotated gene sets encode higher-order functional associations between multiple genes (Luo 2022; Deng et al. 2024). To exploit higher-order gene functional association information in annotated gene sets, we represent and integrate multiple types of annotated gene sets with the hypergraph $HG(V, E, \mathbf{W})$. The node $V = \{v_1, v_2, v_3, ..., v_n\}$ in the hypergraph represents n genes, and the hyperedge $E = \{e_1, e_2, e_3, ..., e_m\}$ represents m annotated gene sets. The $\mathbf{W}$ denotes the diagonal matrix of the hyperedge weights. For the incidence matrix $\mathbf{H} \in \Re^{n \times m}$ of the hypergraph, we define that if the a-th gene v_a belongs to the b-th annotated gene set e_b, then $\mathbf{H}(v_a, e_b) = 1$, otherwise 0. Then, we use the incidence matrix $\mathbf{H}$ of the hypergraph and the pan-cancer feature matrix as the input of the hypergraph neural network method to extract the higher-order functional association information and multi-omics feature of the gene. Referring to previous research, we use the disease-specific hypergraph neural network (Deng et al. 2024), whose formula is as follows:

$$\mathbf{X}^{(l+1)} = \sigma(\mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathbf{W_d}\mathbf{D}_e^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{D}_v^{-\frac{1}{2}}\mathbf{X}^{(l)}\Theta^{(l)} + \mathbf{X}^{(l)}) \tag{2}$$

$$w_i = \frac{\sum_{v \in V} \mathbf{H}(v, e_i) f(v, V_d)}{\sum_{v \in V} \mathbf{H}(v, e_i)} \tag{3}$$

$$f(v, V_d) = \begin{cases} 1 & \text{if } v \in V_d \\ 0 & \text{if } v \notin V_d \end{cases} \tag{4}$$

where the $\mathbf{X}^{(l+1)}$ and $\mathbf{X}^{(l)}$ represent the output of the $l+1$ layer and the l layer. We define the degree of the node v_a is $d(v_a) = \sum_{e \in E} w(e)\mathbf{H}(v_a, e)$ and the degree of the hyperedge e_b is $\delta(e_b) = \sum_{v \in V} \mathbf{H}(v, e_b)$. $\mathbf{W_d}$ is the weighted hyperedge weight matrix and w_i denotes the weight of the i-th ($i \in [1, m]$) hyperedge, and V_d denotes the set of known cancer genes. The $f(v, V_d)$ is used to indicate whether gene v belongs to V_d. The $\mathbf{D_v} \in \mathbb{R}^{n \times n}$ and $\mathbf{D_e} \in \mathbb{R}^{m \times m}$ denote the diagonal matrices of the node degrees and the hyperedge degrees, respectively. The final output of this module is the gene embedding $\mathbf{Z}^{(HG)} \in \mathbb{R}^{n \times d}$.

Enhancing Common Information Based on Cross-View Decorrelation Loss

DISFusion aims to learn the output embedding $\mathbf{Z}$ based on the input, that is, the multi-omics feature $\mathbf{X}$, the graph G, and the hypergraph HG. Integrating two types of association information may help us describe gene association patterns

more comprehensively. For example, genes like *STIM1* and *TRPC1* exhibit PPI without functional associations, while *STIM1* and *GCHFR* show functional associations without direct PPI. This complementarity highlights how PPI and functional data offer distinct yet valuable insights. Additionally, genes like *NOS1* and *BDKRB2*, which share both PPI and functional associations, often indicate stronger, more biologically relevant connections. Such shared neighborhood structure information may be more accurate or important in gene association research. Therefore, it is necessary to maintain the consistency of gene embedding in graphs and hypergraphs to enhance common information across views and improve the quality of gene embeddings.

In DISFusion, we employ the cross-view decorrelation loss to enhance common information from gene embeddings $\mathbf{Z}^{(G)}$ and $\mathbf{Z}^{(HG)}$ across the graph and hypergraph. Specifically, we first generate the cross-view embedding correlation matrix $\mathbf{K} \in \mathbb{R}^{d\times d}$ for $\mathbf{Z}^{(G)}$ and $\mathbf{Z}^{(HG)}$ by defining its element as:

$$k_{ij} = \frac{\sum_n z_{n,i}^{(G)} z_{n,j}^{(HG)}}{\sqrt{\sum_n (z_{n,i}^{(G)})^2}\sqrt{\sum_n (z_{n,j}^{(HG)})^2}} \tag{5}$$

We then enforce $\mathbf{K}$ to converge to the identity matrix $\mathbf{I} \in \mathbb{R}^{d\times d}$ by:

$$\mathcal{L}_{deco} = \sum_{i=1}^{d}(k_{ii}-1)^2 + \lambda \sum_{i=1}^{d}\sum_{j\neq i}(k_{ij})^2 \tag{6}$$

where λ is a non-negative parameter to have a trade-off between the first term and the second term of Eq. (6).

In Eq. (6), the first term makes each element in $\mathbf{Z}^{(G)}$ tend to be equivalent to the corresponding element in $\mathbf{Z}^{(HG)}$ (Zbontar et al. 2021). This consistency constraint enables Eq. (6) to enhance cross-view common information between graph and hypergraph. In addition, the second term eliminates feature redundancy in different dimensions between $\mathbf{Z}^{(G)}$ and $\mathbf{Z}^{(HG)}$, which can enhance the quality of feature embedding (Zbontar et al. 2021).

Finally, we fuse the features of the graph and the hypergraph, and use the fully connected layer and softmax function to obtain the predicted risk score for each gene:

$$\mathbf{Z} = concat(\mathbf{Z}^{(G)}, \mathbf{Z}^{(HG)}) \tag{7}$$

$$\hat{Y} = Softmax(\mathbf{Z}\Theta + \mathbf{b}) \tag{8}$$

We use the cross-entropy loss as the loss function for the cancer gene identification task and optimize it together with the cross-view decorrelation loss.

$$\mathcal{L} = -\sum_{l\in\mathcal{Y}_L}\sum_{i=1}^{2} Y_{li}\hat{Y}_{li} + \alpha\mathcal{L}_{deco} \tag{9}$$

where $Y \in \mathbb{R}^{n\times 2}$ denotes the label of each gene, and $\mathcal{Y}_L$ denotes the labeled nodes set. The α represents the weight of the cross-view decorrelation loss, and we use this weight to adjust the impact of the cross-view decorrelation loss on model training.

Experiment

Data Source and Baselines

Data Source. DISFusion takes three main inputs: the PPI network, pan-cancer multi-omics features, and annotated gene sets. Known cancer and non-cancer genes are used as labeled data during model training.

We use the same PPI network and pan-cancer multi-omics features as EMOGI (Schulte-Sasse et al. 2021). The PPI data, sourced from ConsensusPathDB (CPDB) (Kamburov et al. 2013), filters interactions with scores below 0.5, resulting in a network of 13,627 genes and 504,378 interactions. Pan-cancer features include gene mutation, DNA methylation, and gene expression data from 8,000 samples across 16 cancer types in The Cancer Genome Atlas (TCGA) (Weinstein et al. 2013). The annotated gene data is obtained from the Molecular Signatures Database (MSigDB) (Liberzon et al. 2015), utilizing C2 and C5 gene sets comprising signaling pathway and ontology gene data from expert databases like BioCarta (Rouillard et al. 2016), Reactome (Fabregat et al. 2018), KEGG (Kanehisa and Goto 2000), GO (Consortium 2004), and human phenotype ontology (HPO) (Köhler et al. 2021). To ensure accurate model evaluation, we exclude gene sets directly linked to cancer during training and testing. To align with the previous data, we retained only the functional associations of 13,627 genes in the annotated gene set. In total, we compile 20,647 annotated gene sets representing functional associations among 13,627 genes.

The known cancer driver genes are gathered from NCG(v6.0) (Repana et al. 2019), COSMIC CGC (v91) (Sondka et al. 2018), and DigSEE databases (Kim et al. 2013), serving as positive samples. The non-cancer genes are obtained by iteratively removing genes from NCG, COSMIC CGC, OMIM, DigSEE, and KEGG cancer pathway gene set databases. Therefore, during our model training and testing process, we have a total of 796 positive sample genes and 2,187 negative sample genes.

We also collect six single cancer types datasets: cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), bladder urothelial carcinoma (BLCA), esophageal carcinoma (ESCA), head and neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD), and lung squamous cell carcinoma (LUSC). The positive sample genes for these single cancer types are sourced from DriverDB (v.3) (Liu et al. 2019), with 291, 522, 555, 214, 135, and 396 samples, respectively. The negative samples remain the same 2,187 genes used in the pan-cancer dataset.

Baseline and Implementation. We compare our method with the following six advanced methods: 20/20+ (Tokheim et al. 2016), DORGE (Lyu et al. 2020), EMOGI (Schulte-Sasse et al. 2021), MTGCN (Peng et al. 2022), MODIG (Zhao et al. 2022) and DISHyper (Deng et al. 2024). The key distinctions of DISFusion compared to baseline methods include: (1) it integrates multi-omics data, PPI networks, and functional associations for cancer gene identification, whereas many baselines use only one or two data types; (2) it employs a cross-view decorrelation loss to capture and enhance shared information between PPI and functional associations, thus improving gene representation quality. All

(A)

Method	AUROC	AUPRC
20/20+	0.821±0.025	0.733±0.030
DORGE	0.896±0.017	0.841±0.020
EMOGI	0.893±0.015	0.799±0.028
MTGCN	0.899±0.013	0.813±0.020
MODIG	0.903±0.014	0.819±0.022
DISHyper	0.930±0.010	0.870±0.017
DISFusion	0.948±0.008	0.897±0.013

(B)

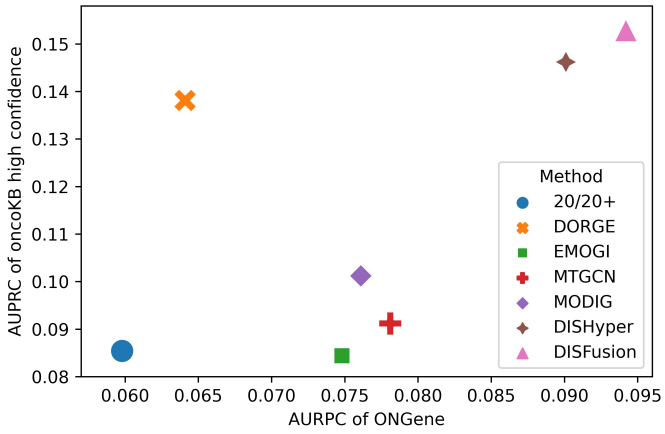

Figure 2: Benchmarking results of DISFusion. **(A)** Performance comparison of DISFusion and six state-of-the-art cancer gene identification methods. **(B)** Performance comparison of DISFusion and other methods on two independent test sets.

experiments are conducted on an Ubuntu server equipped with 256GB of memory and six RTX 2080 GPUs. Detailed implementation information can be found in our code and Appendix A.

Performance on Pan-Cancer Gene Identification

We demonstrate the performance superiority of DISFusion by comparing it with six advanced cancer gene identification methods. The area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) are evaluation metrics for the model performance. As shown in Figure 2A, DISFusion significantly outperforms other advanced cancer gene identification methods. Compared with 20/20+ and DORGE, two methods based on manual features, DISFusion achieved significant improvements of 5.2% and 5.6% on AUROC and AUPRC, respectively (P-value < 0.05 by one-sided Wilcoxon rank-sum test). Compared with EMOGI and MTGCN, two deep learning methods based on biological networks, DISFusion also achieved significant improvements, improving by 4.9% and 8.4% on AUROC and AUPRC, respectively (P-value < 0.05). Compared with the functional association-based method DISHyper, DISFusion also significantly improves AUROC and AUPRC by 1.8% and 2.7%, respectively (P-value < 0.05). Compared with MODIG, a method that integrates multiple gene networks, DISFusion also significantly improves AUROC and AUPRC by 4.5% and 7.8%, respectively (P-value < 0.05). The above results show the performance superiority of DISFusion and the significance of integrating multiple association information.

To further evaluate the generalization performance of DISFusion and its ability to identify cancer genes, we use two sets of curated cancer genes as independent test sets. The two independent test sets are from the OncoKB (Chakravarty et al. 2017) database and the ONGene (Liu, Sun, and Zhao 2017) database. To ensure the fairness of the comparison, we screened 313 and 382 cancer genes from these two databases for performance comparison of independent test sets. We use AUPRC as the evaluation metrics and we treat the genes in the independent test set as true positives and all other genes not included in the independent test set as false positives. Since the baseline size of AUPRC depends on the ratio of positive and negative samples, the AUPRC value of each method is low. As shown in Figure 2B, we find that DISFusion achieves the best performance on both independent test sets. DISHyper is the best-performing method among the compared methods. Compared with DISHyper, DISFusion has a relative improvement of 25% and 13% on two independent test sets, respectively. This result shows that DISFusion has a stronger generalization ability in predicting new cancer genes.

Performance on Cancer Type-Specific Driver Gene Identification

To further evaluate the effectiveness of DISFusion, we test its ability to identify cancer genes in six single cancer types. We use the same PPI network and functional association data as the pan-cancer data to train individual cancer gene identification models.

Table 1 lists the AUPRC of DISFusion and other advanced methods for these six single cancer types. The results show that DISFusion significantly outperforms other advanced methods across all six cancer types. For instance, DISFusion improves by 3.4% over the second-best method DORGE, in CESC. In HNSC and LUAD, DISFusion improves by 6.8% and 9.8%, respectively, over the second-best method DISHyper. These results demonstrate the effectiveness of DISFusion for individual cancer types and highlight its strong generalization capability, indicating its potential as a robust tool for identifying cancer genes.

Significance of Cross-View Decorrelation Loss in DISFusion

DISFusion employs cross-view decorrelation loss to enhance the common information between the PPI network

Method	CESC	BLCA	ESCA	HNSC	LUAD	LUSC
20/20+	0.689±0.049	0.438±0.036	0.647±0.036	0.417±0.069	0.518±0.064	0.485±0.042
DORGE	0.695±0.051	0.512±0.036	0.658±0.028	0.427±0.063	0.537±0.068	0.543±0.048
EMOGI	0.676±0.046	0.521±0.032	0.680±0.042	0.424±0.066	0.499±0.071	0.493±0.050
MTGCN	0.686±0.053	0.537±0.040	0.691±0.039	0.445±0.067	0.546±0.083	0.499±0.048
MODIG	0.672±0.056	0.543±0.047	0.672±0.048	0.462±0.073	0.512±0.076	0.512±0.059
DISHyper	0.658±0.052	0.605±0.048	0.695±0.040	0.474±0.063	0.553±0.062	0.538±0.046
DISFusion	**0.729±0.039**	**0.618±0.053**	**0.729±0.041**	**0.542±0.068**	**0.651±0.065**	**0.574±0.043**

Table 1: The performance of cancer type-specific driver gene identification: Mean AUPRC ± standard deviation. Boldfaced letters are used to indicate the best mean AUPRC and underline is for the second.

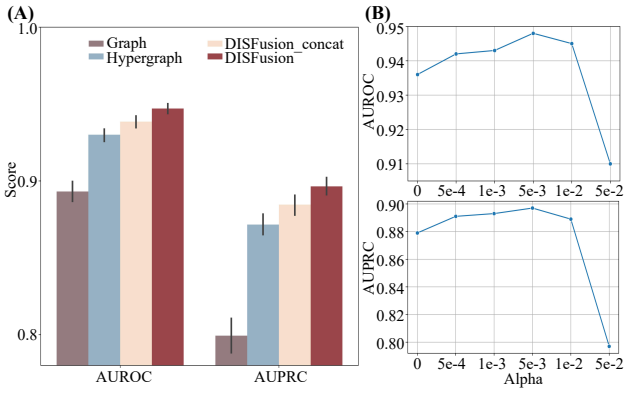

Figure 3: **(A)** Ablation analysis experimental results of DISFusion. The error bars in the figure represent 95% confidence intervals. **(B)** Impact of α on DISFusion performance.

and gene function association. We evaluate the importance and effectiveness of cross-view decorrelation loss through ablation analysis, parameter analysis, and case studies.

Firstly, in ablation analysis, we compare DISFusion with the following variants: (1) Graph: Utilizes only the binary association information extraction module in DISFusion, which uses the PPI network and multi-omics features to predict cancer genes. (2) Hypergraph: Only using the gene functional association information extraction module in DISFusion to predict cancer genes. (3) DISFusion w/o $\mathcal{L}_{deco}$: It is only fusing feature representation without cross-view decorrelation loss. We investigate the effectiveness of fuse two gene association information by comparing DISFusion w/o $\mathcal{L}_{deco}$ with the previous two variants. As shown in Figure 3A, DISFusion w/o $\mathcal{L}_{deco}$ outperforms other variants. Compared with Graph and Hypergraph, DISFusion w/o $\mathcal{L}_{deco}$ improves AUROC and AUPRC by 0.9% and 1.2%, respectively. This result shows that fusing these two types of gene association information can improve the performance of the model. We also investigate the effectiveness of cross-view decorrelation loss by comparing DISFusion with DISFusion w/o $\mathcal{L}_{deco}$. As shown in Figure 3A, adding the cross-view decorrelation loss significantly improved the AUROC and AUPRC of DISFusion by 1.1% and 1.5% (P-value < 0.05), respectively. The results show that using cross-view decorrelation loss can effectively improve the performance of the model in identifying cancer genes.

Secondly, we analyze the impact of the cross-view decorrelation loss weight. Figure 3B shows the change in model performance as the α value changes. We find that the model performance first increase and then decrease as α changes and achieves optimal performance when α is 0.005. When α is 0, the cross-view decorrelation loss has no effect, and the performance drops significantly compared to when α is 0.005. As α increases, cross-view decorrelation loss comes into play, improving model performance by extracting common information across views and reducing redundancy between embeddings. But when $\alpha > 0.01$, the performance of the model drops significantly. This may be because excessively high values of α impose overly stringent constraints on model parameters, leading the loss function to overemphasize the extraction of common information at the expense of optimizing the classification task. These results highlight the critical role of the cross-view decorrelation loss in improving model performance.

Finally, we illustrate the importance of cross-view decorrelation loss through the case study of DISFusion prediction results. In the top-ranked 200 genes of DISFusion's prediction (Appendix Table 1), we find that some genes would be ranked lower in the prediction results of biological network-based or functional association-based methods. For instance, *TTN* and *DMD* are highly ranked by biological network-based approaches but low in functional association-based methods, while the opposite is true for *WNT5A* and *CDC42*. These genes have been extensively documented in the literature as cancer-driver genes or as being closely associated with various cancers (Zheng et al. 2021; Jones et al. 2021; Prasad et al. 2018; Xiao et al. 2018). This finding underscores that integrating multiple types of association information enables more precise identification of cancer genes. Moreover, we observe that *ANXA1* and *TLR2*, which are ranked low by both biological network-based and functional association-based methods, achieve higher rankings in DISFusion. Notably, *ANXA1* has been linked to survival in can-

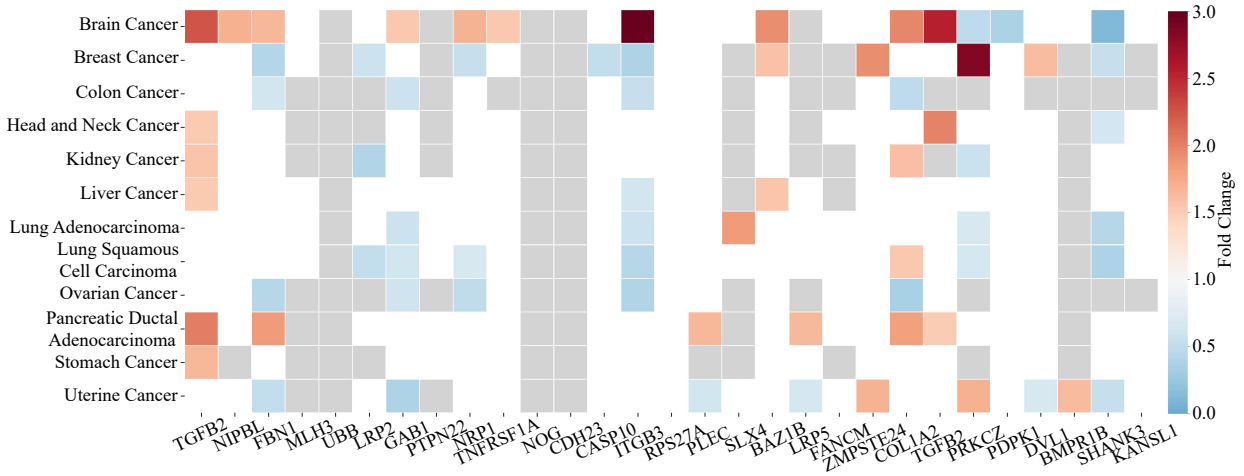

Figure 4: Evaluation of DISFusion-predicted novel cancer genes (novelCG) using CPTAC proteomic data. The figure displays protein differential expression results for 30 novelCG across 12 cancer proteome datasets from the CPTAC study. The brown in the figure represents significant up-regulation in cancer, blue indicates significant down-regulation, gray denotes missing data, and white indicates no significant difference.

cers such as pancreatic cancer and triple-negative breast cancer (Al-Ali et al. 2024), while *TLR2* has been identified as a tumor suppressor gene in non-small cell lung cancer (Millar et al. 2022). This result highlights the critical role of cross-view decorrelation loss, which can help us reveal cancer genes that were not discovered by previous methods to obtain more comprehensive and accurate predictions.

Validation of DISFusion-Predicted Novel Cancer Genes Using Proteomic Data

To demonstrate the practical value of DISFusion, we conducted a comprehensive analysis of the top-ranked 200 genes of DISFusion's prediction (see details in Appendix B and C). Moreover, we find that 170 of the top-ranked 200 genes of DISFusion's prediction have been annotated as cancer genes by the CancerMine database (Lever et al. 2019), and the remaining 30 genes are neither in CancerMine nor in NCG and COSMIC CGC. We regard these 30 genes as novel cancer genes (novelCG, Appendix Table 2) discovered by our model and perform validation analysis on these 30 genes based on the proteomic data in the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Li et al. 2023).

Specifically, we assess the differential protein expression of these 30 novelCG across the 12 tumor types. As shown in Figure 4, we find that 21 genes in novelCG have significant protein differential expression in at least one tumor (fold change > 1.5 or fold change $< 1/1.5$, *P*-value < 0.05). In addition, we also find that *TGFB2*, *TGFB3*, *FBN1*, *GAB1*, *ITGB3*, and other genes have significant differential protein expression in multiple tumors. *TGFB2* and *TGFB3* encode secreted ligands of the transforming growth factor-β (TGF-β) protein superfamily. Both genes are integral components of the TGF-β signaling pathway, regulating processes such as cell growth, proliferation, differentiation, and movement, and are closely associated with apoptosis. In addition, FBN1 encodes the fibrillin-1 protein, which binds to each other and other proteins in the extracellular matrix to form microfibrils. Microfibrils store transforming growth factor-β and help regulate TGF-β availability. Therefore, *FBN1*, along with *TGFB2* and *TGFB3*, may be closely associated with cancer and potentially cooperatively regulate the occurrence and development of various tumors. Among the genes where no protein differential expression is observed, *MLH3* is found to have mutations in cancer patients (Korhonen, Vuorenmaa, and Nyström 2008), while *FANCM* is similarly observed to have significant nonsense mutations in breast cancer (Kiiski et al. 2014). The above results demonstrate the practical value and prediction reliability of DISFusion and suggest that novelCGs may contain potential cancer genes.

Conclusion

In this paper, we propose a new method named DISFusion, which integrates multi-omics features, PPI networks, and gene functional associations to improve the identification of cancer genes. In DISFusion, it first employs GCN and HGNN to extract gene association information and generate gene embeddings from PPI networks and annotated gene sets. It then employs the cross-view decorrelation loss to enhance the common information between the PPI network and gene functional association of annotated gene sets to improve the quality of gene embeddings. Finally, it fuses the gene embedding of these two association information and identifies cancer genes. Extensive experiments indicate that DISFusion outperforms state-of-the-art methods and exhibits greater generalization ability. Our analysis highlights the critical role of cross-view decorrelation loss in improving model performance and identifying cancer genes. Moreover, analysis of CPTAC pan-cancer proteomic data highlights the practical utility of DISFusion and its potential in advancing cancer gene discovery.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 62350004, 62332020, U22A2041), the Project of Xiangjiang Laboratory (No. 23XJ01011), the Science and Technology Major Project of Changsha (No.kh2402004), Hunan Provincial Postgraduate Scientific Research Innovation Project (CX20230234). This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

References

Al-Ali, H. N.; Crichton, S. J.; Fabian, C.; Pepper, C.; Butcher, D. R.; Dempsey, F. C.; and Parris, C. N. 2024. A therapeutic antibody targeting annexin-A1 inhibits cancer cell growth in vitro and in vivo. *Oncogene*, 1–7.

Alexandrov, L. B.; Nik-Zainal, S.; Wedge, D. C.; Aparicio, S. A.; Behjati, S.; Biankin, A. V.; Bignell, G. R.; Bolli, N.; Borg, A.; Børresen-Dale, A.-L.; et al. 2013. Signatures of mutational processes in human cancer. *Nature*, 500(7463): 415–421.

Althubaiti, S.; Karwath, A.; Dallol, A.; Noor, A.; Alkhayyat, S. S.; Alwassia, R.; Mineta, K.; Gojobori, T.; Beggs, A. D.; Schofield, P. N.; et al. 2019. Ontology-based prediction of cancer driver genes. *Scientific Reports*, 9(1): 1–9.

Barel, G.; and Herwig, R. 2020. NetCore: a network propagation approach using node coreness. *Nucleic Acids Research*, 48(17): e98–e98.

Chakravarty, D.; Gao, J.; Phillips, S.; Kundra, R.; Zhang, H.; Wang, J.; Rudolph, J. E.; Yaeger, R.; Soumerai, T.; Nissan, M. H.; et al. 2017. OncoKB: a precision oncology knowledge base. *JCO Precision Oncology*, 1: 1–16.

Chatzianastasis, M.; Vazirgiannis, M.; and Zhang, Z. 2023. Explainable multilayer graph neural network for cancer gene prediction. *Bioinformatics*, 39(11): btad643.

Consortium, G. O. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(suppl_1): D258–D261.

Creixell, P.; Gonzalez-Perez, A.; López Bigas, N.; et al. 2015. Pathway and network analysis of cancer genomes. *Nature Methods*, 12(7): 615–621.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems*, 29: 3844–3852.

Deng, C.; Li, H.-D.; Zhang, L.-S.; Liu, Y.; Li, Y.; and Wang, J. 2024. Identifying new cancer genes based on the integration of annotated gene sets via hypergraph neural networks. *Bioinformatics*, 40(Supplement 1): i511–i520.

Fabregat, A.; Jupe, S.; Matthews, L.; Sidiropoulos, K.; Gillespie, M.; Garapati, P.; Haw, R.; Jassal, B.; Korninger, F.; May, B.; et al. 2018. The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1): D649–D655.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3558–3565.

Garraway, L. A.; and Lander, E. S. 2013. Lessons from the cancer genome. *Cell*, 153(1): 17–37.

Jones, L.; Naidoo, M.; Machado, L. R.; and Anthony, K. 2021. The Duchenne muscular dystrophy gene and cancer. *Cellular Oncology*, 44: 19–32.

Kamburov, A.; Stelzl, U.; Lehrach, H.; and Herwig, R. 2013. The ConsensusPathDB interaction database: 2013 update. *Nucleic acids research*, 41(D1): D793–D800.

Kanehisa, M.; and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1): 27–30.

Kiiski, J. I.; Pelttari, L. M.; Khan, S.; Freysteinsdottir, E. S.; Reynisdottir, I.; Hart, S. N.; Shimelis, H.; Vilske, S.; Kallioniemi, A.; Schleutker, J.; et al. 2014. Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proceedings of the National Academy of Sciences*, 111(42): 15172–15177.

Kim, J.; So, S.; Lee, H.-J.; Park, J. C.; Kim, J.-j.; and Lee, H. 2013. DigSee: disease gene search engine with evidence sentences (version cancer). *Nucleic Acids Research*, 41(W1): W510–W517.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Köhler, S.; Gargano, M.; Matentzoglu, N.; Carmody, L. C.; Lewis-Smith, D.; Vasilevsky, N. A.; Danis, D.; Balagura, G.; Baynam, G.; Brower, A. M.; et al. 2021. The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1): D1207–D1217.

Korhonen, M. K.; Vuorenmaa, E.; and Nyström, M. 2008. The first functional study of MLH3 mutations found in cancer patients. *Genes, Chromosomes and Cancer*, 47(9): 803–809.

Lawrence, M. S.; Stojanov, P.; Mermel, C. H.; Robinson, J. T.; Garraway, L. A.; Golub, T. R.; Meyerson, M.; Gabriel, S. B.; Lander, E. S.; and Getz, G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484): 495–501.

Lawrence, M. S.; Stojanov, P.; Polak, P.; Kryukov, G. V.; Cibulskis, K.; Sivachenko, A.; Carter, S. L.; Stewart, C.; Mermel, C. H.; Roberts, S. A.; et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457): 214–218.

Lever, J.; Zhao, E. Y.; Grewal, J.; Jones, M. R.; and Jones, S. J. 2019. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nature Methods*, 16(6): 505–507.

Li, Y.; Dou, Y.; Leprevost, F. D. V.; Geffen, Y.; Calinawan, A. P.; Aguet, F.; Akiyama, Y.; Anand, S.; Birger, C.; Cao, S.; et al. 2023. Proteogenomic data and resources for pan-cancer analysis. *Cancer cell*, 41(8): 1397–1406.

Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J. P.; and Tamayo, P. 2015. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425.

Liu, S.-H.; Shen, P.-C.; Chen, C.-Y.; Hsu, A.-N.; Cho, Y.-C.; Lai, Y.-L.; Chen, F.-H.; Li, C.-Y.; Wang, S.-C.; Chen, M.; Chung, I.-F.; and Cheng, W.-C. 2019. DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Research*, 48(D1): D863–D870.

Liu, Y.; Sun, J.; and Zhao, M. 2017. ONGene: a literature-based database for human oncogenes. *Journal of Genetics and Genomics*, 44(2): 119–121.

Liu, Y.; Yang, C.; Li, H.-D.; and Wang, J. 2023. IsoFrog: a reversible jump Markov Chain Monte Carlo feature selection-based method for predicting isoform functions. *Bioinformatics*, 39(9): btad530.

Luo, Y. 2022. SHINE: SubHypergraph Inductive Neural nEtwork. In *Advances in Neural Information Processing Systems*.

Lyu, J.; Li, J. J.; Su, J.; Peng, F.; Chen, Y. E.; Ge, X.; and Li, W. 2020. DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features. *Science Advances*, 6(46): eaba6784.

Millar, F. R.; Pennycuick, A.; Muir, M.; Quintanilla, A.; Hari, P.; Freyer, E.; Gautier, P.; Meynert, A.; Grimes, G.; Coll, C. S.; et al. 2022. Toll-like receptor 2 orchestrates a tumor suppressor response in non-small cell lung cancer. *Cell reports*, 41(6).

Nofech-Mozes, I.; Soave, D.; Awadalla, P.; and Abelson, S. 2023. Pan-cancer classification of single cells in the tumour microenvironment. *Nature Communications*, 14(1): 1615.

Peng, W.; Tang, Q.; Dai, W.; and Chen, T. 2022. Improving cancer driver gene identification using multi-task learning on graph convolutional network. *Briefings in Bioinformatics*, 23(1): bbab432.

Prasad, C. P.; Manchanda, M.; Mohapatra, P.; and Andersson, T. 2018. WNT5A as a therapeutic target in breast cancer. *Cancer and Metastasis Reviews*, 37(4): 767–778.

Repana, D.; Nulsen, J.; Dressler, L.; Bortolomeazzi, M.; Venkata, S. K.; Tourna, A.; Yakovleva, A.; Palmieri, T.; and Ciccarelli, F. D. 2019. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biology*, 20(1): 1–12.

Reyna, M. A.; Haan, D.; Paczkowska, M.; Verbeke, L. P.; Vazquez, M.; Kahraman, A.; Pulido-Tamayo, S.; Barenboim, J.; Wadi, L.; Dhingra, P.; et al. 2020. Pathway and network analysis of more than 2500 whole cancer genomes. *Nature Communications*, 11(1): 1–17.

Reyna, M. A.; Leiserson, M. D.; and Raphael, B. J. 2018. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics*, 34(17): i972–i980.

Rouillard, A. D.; Gundersen, G. W.; Fernandez, N. F.; Wang, Z.; Monteiro, C. D.; McDermott, M. G.; and Ma'ayan, A. 2016. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016.

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

Sondka, Z.; Bamford, S.; Cole, C. G.; Ward, S. A.; Dunham, I.; and Forbes, S. A. 2018. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11): 696–705.

Tokheim, C. J.; Papadopoulos, N.; Kinzler, K. W.; Vogelstein, B.; and Karchin, R. 2016. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, 113(50): 14330–14335.

Vogelstein, B.; Papadopoulos, N.; Velculescu, V. E.; Zhou, S.; Diaz Jr, L. A.; and Kinzler, K. W. 2013. Cancer genome landscapes. *Science*, 339(6127): 1546–1558.

Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; and Stuart, J. M. 2013. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10): 1113–1120.

Xiao, X.-H.; Lv, L.-C.; Duan, J.; Wu, Y.-M.; He, S.-J.; Hu, Z.-Z.; and Xiong, L.-X. 2018. Regulating Cdc42 and its signaling pathways in cancer: small molecules and MicroRNA as new treatment candidates. *Molecules*, 23(4): 787.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, 12310–12320. PMLR.

Zhang, T.; Zhang, S.-W.; Xie, M.-Y.; and Li, Y. 2023. A novel heterophilic graph diffusion convolutional network for identifying cancer driver genes. *Briefings in Bioinformatics*, 24(3): bbad137.

Zhao, W.; Gu, X.; Chen, S.; Wu, J.; and Zhou, Z. 2022. MODIG: integrating multi-omics and multi-dimensional gene network for cancer driver gene identification based on graph attention network model. *Bioinformatics*, 38(21): 4901–4907.

Zheng, Q.-X.; Wang, J.; Gu, X.-y.; Huang, C.-H.; Chen, C.; Hong, M.; and Chen, Z. 2021. TTN-AS1 as a potential diagnostic and prognostic biomarker for multiple cancers. *Biomedicine & Pharmacotherapy*, 135: 111169.