Supplementary manuscript of

A semi-supervised graph neural network method for prioritizing personalized driver genes of individual patients with cancer

Han-Wen Wan1,2†, Meng-Han Wu3†, Wen-Shan Zhao3, Han Cheng3, Xiang-Rui Zhang3, Yan Li4, Wei-Feng Guo1,5\*

1 School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China

2 School of International College, Zhengzhou University, Zhengzhou 450001, China

3 School of Life Sciences, Zhengzhou University, Zhengzhou 450001, China

4 Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xian, 710072, China

5 State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

\* To whom correspondence should be addressed.

[Tel:86-0371-677810](Tel:86-21-54920100)18; Fax: [86-0371-677810](Tel:86-21-54920100)18; Email: [guowf@zzu.edu.cn](mailto:liangjing@zzu.edu.cn,)

† The authors should be regarded as Joint First Authors.

## Supplementary note 1: The statistic information of positive and negative genes in PGIN

Table S1 The statistic information of positive and negative genes in PGIN

|  |  |  |  |
| --- | --- | --- | --- |
|  | BRCA | LUSC | LUAD |
| #posi | 18.4602+-1.5749 | 27.7292+-3.0479 | 29.4286+-3.7005 |
| #neg | 191.9115+-15.8389 | 419.7292+-23.0389 | 412.5000+-24.4861 |

## Supplementary note 2: details of class weighted loss function

To balance the loss of imbalanced classes or labels in PGIN, we adopted a class-balanced softmax cross-entropy loss function by introducing a normalized weighting factor,1/where is the effective number of genesfor class *y*in PGIN. The effective number of genes for class *y* is defined as ),=*N*/(*N*-1) where *ny* is the number of genes in the class or label *y* and *N* is the number of genes in PGIN. The class weighted loss function can be written as



wheredenotes the model's estimated class probabilities. The weighting factor is a normalized vector=*C; C* is the number of classes (here *C* = 2); is softmax cross-entropy loss function. If the gene *x*is within a positive label set, its label , otherwise, .

## Supplementary note 3: Details of some methods

**Construction of PGIN with Paired-SSN**

To obtain the PGIN to characterize the personalized state transition of an individual patient in cancer development, we used the Paired-SSN method on gene expression and SNV mutation datasets of BRCA, LUSC, and LUAD. The Paired-SSN method constructs PGIN by mainly integrating the gene somatic mutation data, personalized gene expression data, and network topology information from the known human genetic interaction network The PGIN of an individual patient *k* composes the adjacency matrix  and node features, in which *N* is the gene number. The feature of gene *i* is calculated with the following formula:











where Norm denotes the min-max normalized function; *G*(*i*) and *G*(*j*) denote the set of individual tumors for mutated gene *i* and gene *j*, respectively, by inspecting somatic mutations in a given cancer dataset; is the *PCC* between gene *i* and gene *j* in the reference network with *n* reference samples, and is the *PCC* in the perturbed network with one additional sample (tumor or normal sample) for individual patient *k*. *N*(*i*) represents the neighboring node set of node *i* in the PGIN.

**Training process of GAT**

Given a graph *G=*(*V*, *E*) of PGIN, we denote *A* and as the adjacency matrix and node features, respectively, where is the set of nodes, is the set of edges, *N* is the number of nodes and *F* is the dimension of the node feature. A one-hot encoded matrix , is denoted as label matrix with *C* being the number of classes (here *C*=2). Each node is associated with a feature vector and label , assuming that only the first *M* nodes can be observed during training. For each dataset associated with a graph *G*, we have the training set and the test set . The goal of training GAT is to predict the labels of non-labeled nodes.

The training process of GAT for predicting PDGs includes two aspects. One is node feature representation with graph attention layers, and the other is gene label prediction to be a cancer driver gene.

**Node feature representation with graph attention layers.** The graph attention layer learns the node feature representation in the PGIN. The graph attention layer is inputted with a set of node features . The output of the graph attention layer is a new set of learned node features, (of potentially different cardinality). The graph attention layer utilizes a multi-head self-attention mechanism to compute the node representation. The attention coefficient between node *i* and node *j* is computed by performing a shared attention mechanism as follows:





where denotes the parameter vector of a single-layer feedforward neural network applying the LeakyReLU nonlinearity (with a slope of 0.2), denotes the learnable linear transformation weight matrix to transform the input features into higher-level features || represents concatenation operation, and denotes the first-order neighbors of node *I* in PGIN*.* The node representation is computed by using *K* independent attention mechanisms as follows:



where denotes normalized attention coefficient computed by the *k-*th attention mechanism, is the corresponding input linear transformation's weight matrix, and

Especially if the final (prediction) layer of the GAT network is performed by multi-head attention, concatenation is no longer sensible, and the node representation is generated by averaging the features with multiple attentions,



**Label prediction to be a cancer driver gene.** After obtaining the node representation vector through the GAT layer, the gene label prediction in PGIN can be obtained in the following form:

where denotes the learned node feature vector from graph attention layers, *N* is the number of genes in the PGIN, and *C* is the number of classes or labels (here *C*=2). All labeled nodes are used to evaluate the cross-entropy error in this semi-supervised node classification task.

**Label reuse strategy**

To fully use the structural information of PGIN and the limited well-established cancer tissue-specific driver gene information, we used label reuse strategy on GAT, which is a data augmentation method to allow propagation of both node features and labels during the semi-learning process. The pseudo-code of the label reuse strategy is shown below.

|  |
| --- |
| The pseudo-code of the label reuse strategy |
| **INPUT**: , *h*, *Y*  1. // The number of epochs is set as 2000 |
| 2. **for** each epoch **do** |
| 3. Obtain by randomly splitting into *n* sub-datasets |
| 4. **for** *k* 1 to *n* **do** |
| 5.  6. |
| 7. **end for** |
|  |
| 9.  10. // denotes the GAT model with parameter |
| 11. Compute and update via back propagation |
| 12. **end for** |

In detail, we first obtained initial predicted labels by considering adjacency matrix *A* of PGIN, initial node features *h*, and original node labels *Y* (i.e., positive labels, negative labels, and other non-labels) in the PGIN as the inputs of an initial GAT model for the label reuse strategy's iteration process. Then the label reuse strategy was applied to GAT for voting genes to be cancer driver genes. The label reuse strategy iteratively recycles the predicted soft labels of the previous iteration (or epoch) as the feature of GAT in the current iteration (the number of epochs or iterations was set as 2000). During each epoch/iteration, the label reuse strategy splitsinto *n* random sub-datasets (here *n* = 3). For one of the sub-datasets, its true label *Y* was concatenated with the previous iteration's predicted label of other sub-datasets as one generated feature. The same operation was applied on all sub-datasets to obtain all generated features, which were updated correspondingly while new predicted labels were utilized for a new iteration. All generated features were concatenated with the initial feature *h* to construct an augmented feature matrix for each iteration. Finally, we outputted the predicted labels of all iterations by considering their corresponding augmented feature matrix as the trained feature of the model.

**DIseAse MOdule Detection (DIAMOnD) method for detecting driver gene module of NPDGs**

DIseAse MOdule Detection (DIAMOnD) method [40] is used to gradually build driver gene module (DGM) in an iterative scheme that exploits the network’s topology. Given a targeting set of **NPDGs**, DIAMOnD calculates the statistical connectivity signiﬁcance of each node to the candidate DGM at each iteration. If DGM at the current iteration is composed of *s* genes, then a candidate node with degree *k* and *ks* edges connected to the *s* genes in the module has a *p-value*.

 (14)

 (15)

where *N* denotes the total number of genes in PGIN. *P*-value threshold is chose as 1e-2 to obtain the significant connected genes at each interaction. This operation is repeated for a ﬁxed number of iterations *Nmax=*10, reaching ﬁnal DGM and corresponding *module scores* (i.e., the ratio of the weight sum of interactions within modules to the weight sum of interactions connecting the genes in module and genes outside module).

## Supplementary note 4: Experimental parameters sensitivity analysis

To further demonstrate the effectiveness of the setting parameters in our PersonalizedGNN, we tuned one of the parameters over suitable ranges of parameter values while we fixed other parameters in BRCA, LUSC, LUAD data sets. For each cancer data set, we used the same parameters among all individual patients andcalculated the average gene ranking score among all individual patients for each cancer data. Based on the average gene ranking score of the cancer data, we obtained the gene ranking result and calculated the *AveragePrecision* of different parameter choices on each cancer data. The parameter setting resultson BRCA, LUAD, and LUSC were shown in **Fig S1 of Additional file 1**. As shown in Figure S1, PersonalizedGNN has the best performance for our setting parameters in PersonalizedGNN.

**para check**

**Fig S1**The error bar in terms of *AveragePrecision*for different parameter choices among top *k*(*k*=1,2,...,30) ranking genes on BRCA, LUSC, LUAD data sets.

## Supplementary note 5: parameter setting of PersonalizedGNN and other methods

* **The parameters in PersonalizedGNN**

Table S2 The parameters of PersonalizedGNN on cancer driver gene prediction

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The number of layers | The number of nodes in each layer | weight decay | learn rate | coefficient of Dropout | coefficient of DropEdge | coefficient of DropAttention | The number of attention heads | epoch |
| 3 | 750 | 1E-7 | 0.02 | 0.25 | 0.75 | 0.75 | 5 | 2000 |

* **Parameters in other methods**

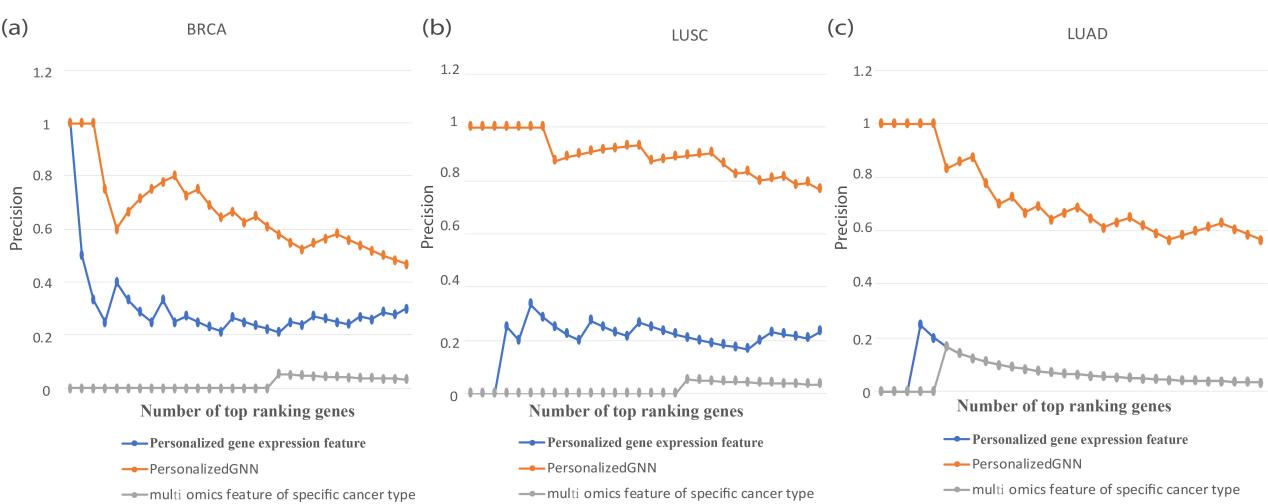
ChebNet: the graph convolution kernel size is 2; the number of layers is 3;the size of each hidden layer is 750; the number of epochs is 2000; bias in graph convolution is enabled, the activation function is ReLU; the dropout rate is 0.25; the learning rate is 0.02, and the optimizer is Adam.

GraphSAGE: the number of layers is 3; the size of each hidden layer is 750; the number of epochs is 2000; the number of sampled neighbors is 10; the batch size is 50; and the learning rate is 0.02.

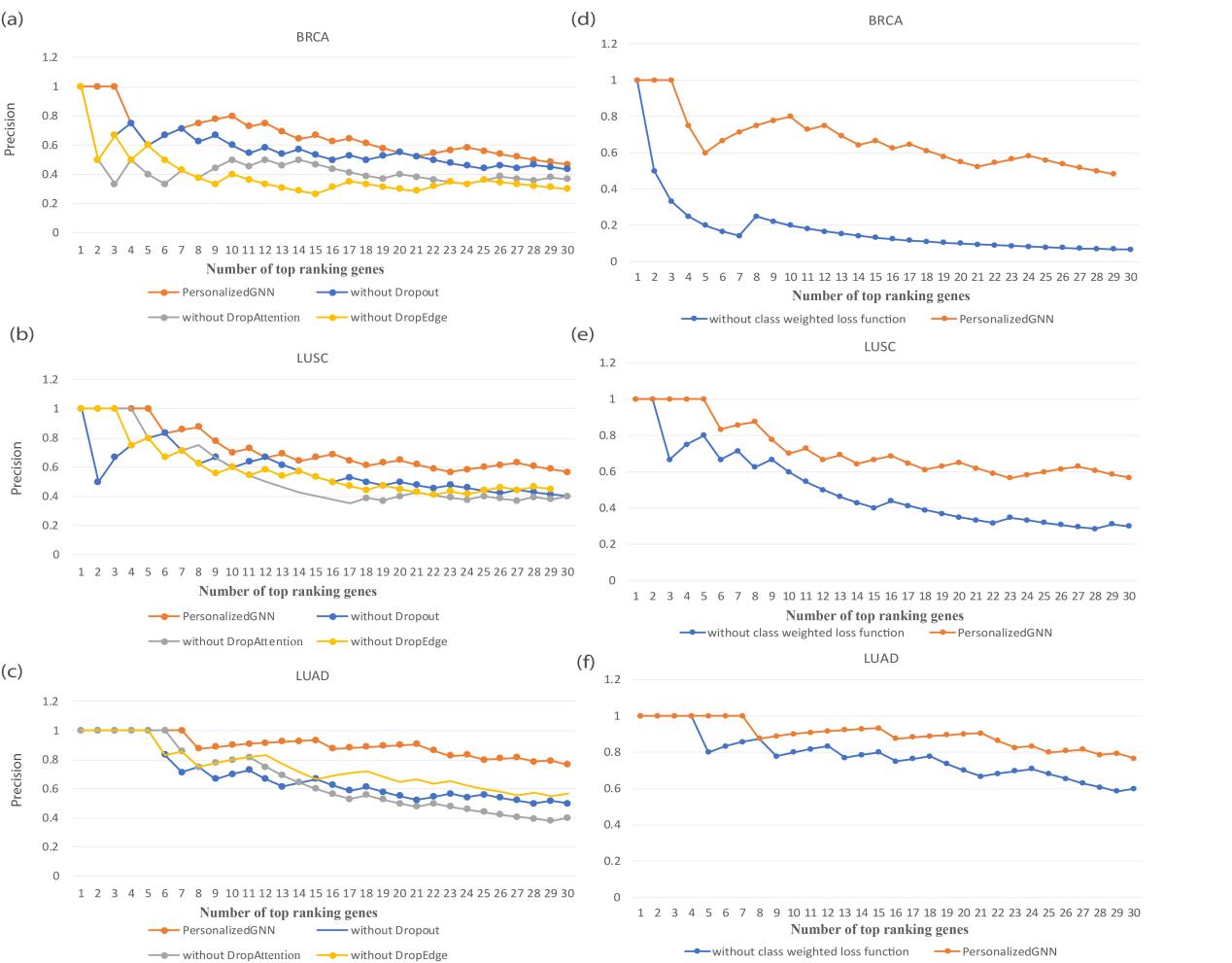
GCN: the number of layers is 3; the size of hidden layers is 750; the num of epochs is 2000; the the dropout rate is 0.25; learning rate is 0.02; and the weight decay is 1e-7.

GAT without label reuse: the heads of attention mechanism is 5; the number of layers is 3; the size of each hidden layer is 500; the number of epochs is 2000; the coefficient of Dropout, DropEdge and DropAttention is 0.25, 0.75 and 0.75 respectively; the learning rate is 0.02; the weight decay is 1e-7; and the activation function is LeakyReLU with the slope alpha equals 0.2.

## Supplementary note 6: supplementary results of PersonalizedGNN



**Fig. S2 The *Precision* resultsof feature selection strategy in our PersonalizedGNNand other feature selection strategies (i.e., personalized gene expression feature and multi-omics feature of specific cancer type) in BRCA, LUSC, and LUAD cancer datasets.**

****

**Fig. S3 The validation results of regularization strategies and class weighted loss function on PersonalizedGNN. (a-c)** The comparison results of PersonalizedGNN with regularization strategies and without regularization strategies in terms of the *Precision* of ranking genes for BRCA, LUSC, and LUAD cancer datasets. **(d-f)** The comparison results of PersonalizedGNN with class weighted loss function and without class weighted loss functionin terms of the *Precision* of ranking genes for BRCA, LUSC, and LUAD cancer datasets.

342b84516eba339c4ea137fec87cc6a4

**Fig.S4** Enrichment pathway results of NPDGs on different subtypes of BRCA cancer patients.

770023a6d07613f2e38a5fa451d25461

**Fig.S5** Enrichment pathway results of NPDGs on different subtypes of LUNG cancer patients.

**Table S3 Results of *Average Precision* of PersonalizedGNN and other methods for BRCA, LUAD, and LUSC cancers.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **BRCA** | **LUAD** | **LUSC** |
| PersonalizedGNN | **0.661139** | **0.897047** | **0.720994** |
| IMCDriver\_CGC | 0.16937103 | 0.330106 | 0.559499 |
| IMCDriver\_TissueSpecificCGC | 0.035534 | 0.143266 | 0.192144 |
| PNC | 0 | 0 | 0 |
| SCS | 0 | 0 | 0 |
| ActiveDriver | 0 | 0.079398 | 0 |
| Prodigy | 0.133166 | 0 | 0 |
| DriverML | 0.287985 | 0.041073 | 0.063648 |

**Table S4** Performance comparisons in terms of the *Average Precision* of PersonalizedGNN and other GNN models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **BRCA** | **LUAD** | **LUSC** |
| PersonalizedGNN | **0.661139** | **0.897047** | **0.720994** |
| GCN | 0.158598 | 0.233371 | 0.155025 |
| ChebNet | 0.132337 | 0 | 0.088503 |
| GraphSAGE | 0.08628 | 0.025352 | 0.108519 |
| GAT | 0.299799 | 0.416707 | 0.648696 |

**Table S5** *Average Precision* of different strategies in PersonalizedGNN for BRCA, LUAD, and LUSC cancer datasets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item** | **Operation** | **BRCA** | **LUAD** | **LUSC** |
| Ours | PersonalizedGNN | **0.661139** | **0.897047** | **0.720994** |
| Network structure | SSN | 0.163885 | 0.293982 | 0.306612 |
| random network | 0.015832 ± 0.02179 | 0.305975 ± 0.200504 | 0.327346 ± 0.202255 |
| Negative label | random negative label | 0.118170 ± 0.064653 | 0.179472 ± 0.065943 | 0.183914 ± 0.048409 |
| Feature | multi-omics feature | 0.016663 | 0.057055 | 0.135739 |
| Personalized gene expression feature | 0.299665 | 0.072055 | 0.203805 |
| Regularization strategy | Dropout strategy | 0.560644 | 0.674893 | 0.569981 |
| DropEdge strategy | 0.391050 | 0.742409 | 0.574552 |
| DropAttention strategy | 0.428262 | 0.657250 | 0.544985 |
| Loss function | without class weighted loss function | 0.179904 | 0.772945 | 0.497805 |

## Supplementary note 7: details of in vitro cell-based assays

**Cell lines and cell culture**

LUSC cell line H1703 was purchased from Procell Life Science & Technology (Wuhan, China) and cultured in RPMI 1640 medium (GIBCO, USA) supplemented with 10% fetal bovine serum (FBS, BI, USA), 100 U/mL penicillin (Solarbio, China), and 100 μg/mL streptomycin (Solarbio, China). All cells were cultured at 37◦C in a humidified atmosphere of 95% air and 5% CO2.

**Plasmid construction and lentiviral transfection**

Short hairpin RNA (shRNA) was cloned into PLKO.1 vector and transfected into H1703(H1703 shFZD7) cells using a lentiviral transfection system. The empty vector was transfected into H1703 cells as a control (H1703 shNT). The target sequence is 5’-GCCGCTTCTACCACAGACTTA-3'.

**qRT‑PCR analysis**

According to the manual, total RNA was extracted from cell lines and tissue samples using a total RNA isolation kit (Tiangen, DP419, China). Tefrst strand cDNA was synthesized from 2 µg of total RNA using the Reverse Assisted First Strand cDNA Synthesis Kit (ThermoFisher, K1622, USA). The qRT-PCR assay was performed using SYBR Green I Master (Roche, 04887352001, Switzerland) in Roche LightCycler 480 II. The primers for FZD7 were 5’-CGTGTCGTTCTCTGTGCGAG-3'and 5'-GTAGGGCGCGGTAGGGTAG-3'. *GAPDH* was used as a control to normalize the expression of each gene.

**Cell viability assay**

H1703 cells (3000 cells/well) were seeded in 96-well plates and cultured overnight. MTT (Sigma) was added to the culture wells at 24 h, 48 h, 72 h, 96 h, and 120 h and incubated at 37 °C for 4 h. After incubation, the medium was discarded, and 150 μL of dimethyl sulfoxide was added to each well to dissolve the formazan crystals. The absorbance at 490 nm was measured using a microplate reader.

**Colony formation assay**

The siRNA-transfected H1703 cells were seeded in six-well plates with 3000 cells/well and cultured for 7 days. Then the medium was aspirated, and the cells were washed with phosphate-buffered saline (PBS) and fixed with 4% paraformaldehyde solution at room temperature for 30 min. Thereafter, the fixative was removed, and 0.2% crystal violet staining solution was added to stain the cells at room temperature for 30 min. After staining, the crystal violet was rinsed with PBS, and the cells were dried at room temperature. Then, the colonies of cells were counted and photographed.

**Invasion assay**

In vitro invasion assays were performed using Transwell chambers (Corning, MA, USA). H1703 cells (1×105) transfected with shRNA and negative shRNA were plated on the 24-well plate and cultured for 24 h. After the incubation, the medium was discarded and fixed at room temperature for 30 min. Then, 0.2% crystal violet staining solution was added to stain the cells at room temperature for 30 min. After staining, the chamber was washed with PBS, and the cells were carefully wiped off with a cotton swab. The stained cells were counted under a microscope and photographed.

**Western blot assay**

Anti-FZD7 antibody (A4213), anti-CyclinA2antibody (A7632), and anti-CyclinE2 antibody (A9305) were purchased from ABclonal (Cambridge, MA, USA). Anti-P27 antibody (25614-1-AP) was purchased from Abcam. Anti-β-actin (GB11001) antibody was purchased from Servicebio. Total cellular protein was extracted by adding a protein lysis buffer. Proteins were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and transferred to a polyvinylidene difluoride (PVDF) membrane (Merck Millipore, IPVH00010, USA). The membrane was blocked with 5% milk dissolved in Tris-buffered saline (TBST) for 1 h at room temperature. Then, the PVDF membrane was incubated with the antibody (1:1000)dissolved in 5% milk overnight at 4 °C. After that, the membrane was washed six times for 5 min each time using TBST. Thereafter, the membrane was incubated with horseradish peroxidase-conjugated mouse anti-rabbit antibody (1:8,000) for 2 h at room temperature. Finally, the membrane was washed six times, and the enhanced chemiluminescence detection kit (Applygen Technologies, Beijing, China) was used to expose the protein bands. The results were analyzed using ImageJ software (NIH, Bethesda, MD, USA).

**Cell cycle assays**

The cell cycle was analyzed using a Cell Cycle Staining Kit (eBioscience). Each sample with 5 × 106 cells was washed with PBS and incubated with 500 μL DNA staining solution and 5 μL permeabilization solution for 30 min at room temperature. The samples were analyzed by flow cytometry.