BIBS seminar

GraphPath: a graph attention model for molecular stratification with interpretability based on the pathway-pathway interaction network

2024. 5. 9

Sangyeon Shin

Bioinformatics, 2024, **40(4)**, btae165
https://doi.org/10.1093/bioinformatics/btae165
Advance Access Publication Date: 26 March 2024
Original Paper

OXFORD

Systems biology

GraphPath: a graph attention model for molecular stratification with interpretability based on the pathway–pathway interaction network

Teng Ma[1] and Jianxin Wang [1,*]

[1]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 41083, Hunan, China

*Corresponding author. Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Computing Building 303, South LuShan Road 932, Changsha 41083, Hunan, China. E-mail: jxwang@mail.csu.edu.cn (J.W.)

Associate Editor: Pier Luigi Martelli

Contents

Introduction

Characteristic of Cancer

- Molecular heterogeneity is one of the most important concerns in cancer.
- During the growth and proliferation of the tumor cell, the daughter cells exhibit differences at the molecular level, resulting in variations in tumor growth rate, invasion and metastasis patterns, drug sensitivity, and other aspects.
- Because of heterogeneity, same type of cancer can get varying treatment responses and overall prognosis among patients.

Cancer subtype classification with omics

- Recent advances in genomics, proteomics, and molecular pathology have led to the discovery of candidate biomarkers.
  - Cancer subtype classification, staging, and prognosis.
- The application of molecular profiling technologies can be used for personalized medicine.
- For more accurate prediction of tumor subtypes, projects such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) was progressed.

Interpretability of prediction models

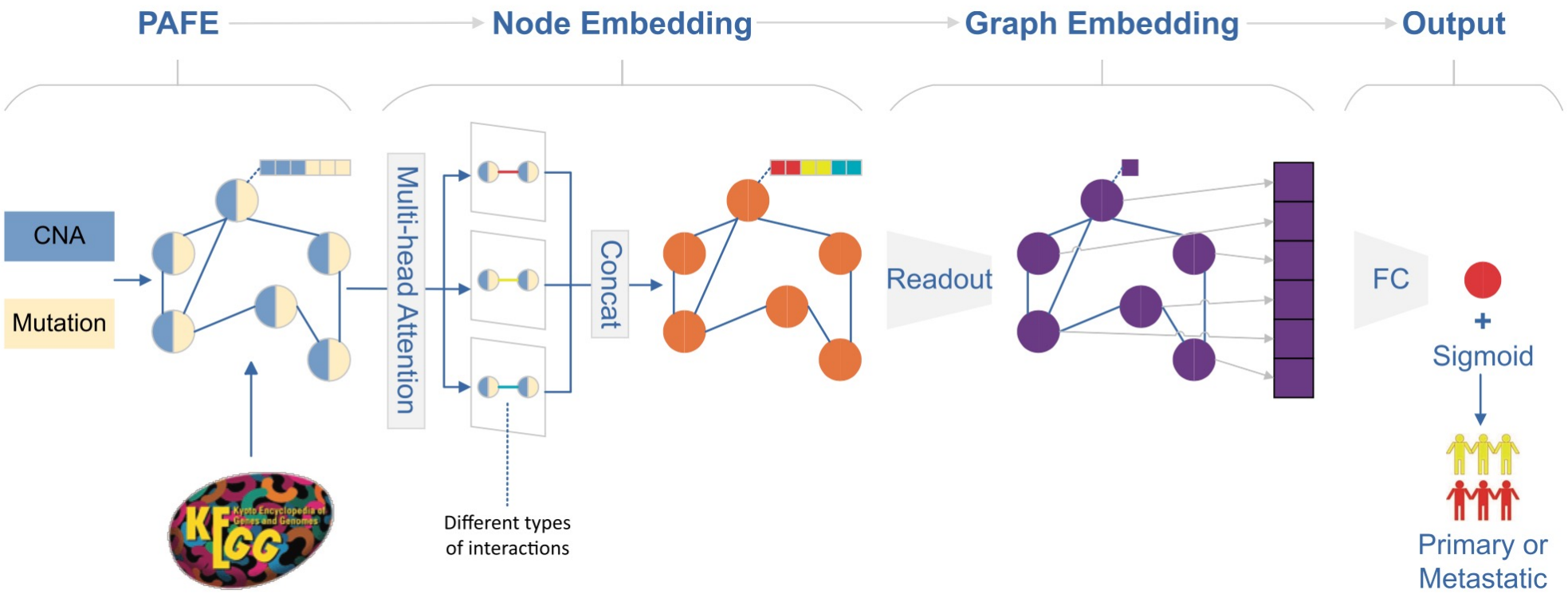
- High accuracy and interpretability are crucial for clinical prediction models that allow healthcare professionals to make informed decisions.
- Recently, with the development of cancer research, it has become a consensus that multiple driver pathways are cooperatively involved in the transformation process from a normal cell to a tumor during cancer development.
- Therefore, more attention has been paid to identifying driver pathways and functional modules rather than individual genes, leading to the emergence of pathway-based deep neural networks(e.g. P-NET).

Materials and Methods

Materials

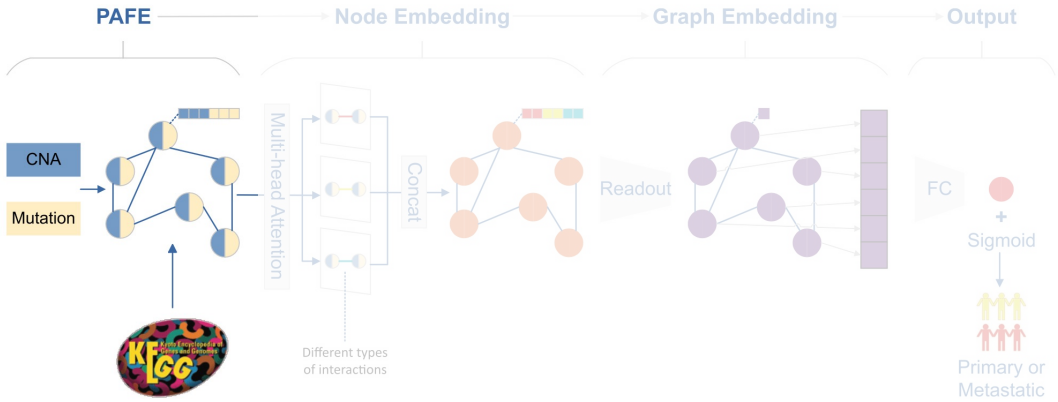
- KEGG pathway database
  - Pathway-pathway interaction annotation
  - Gene-pathway membership annotation
  - 511 pathways
    - Each pathway has a gene set ranges of its size from 2 to 1926.
- Cancer datasets
  - 1,013 prostate cancers
  - Armenia, et al.(2018, *Nat Genet*)
  - Same as P-NET dataset
    - Copy Number Alteration(CNA)
    - Somatic Mutation

Architecture of GraphPath



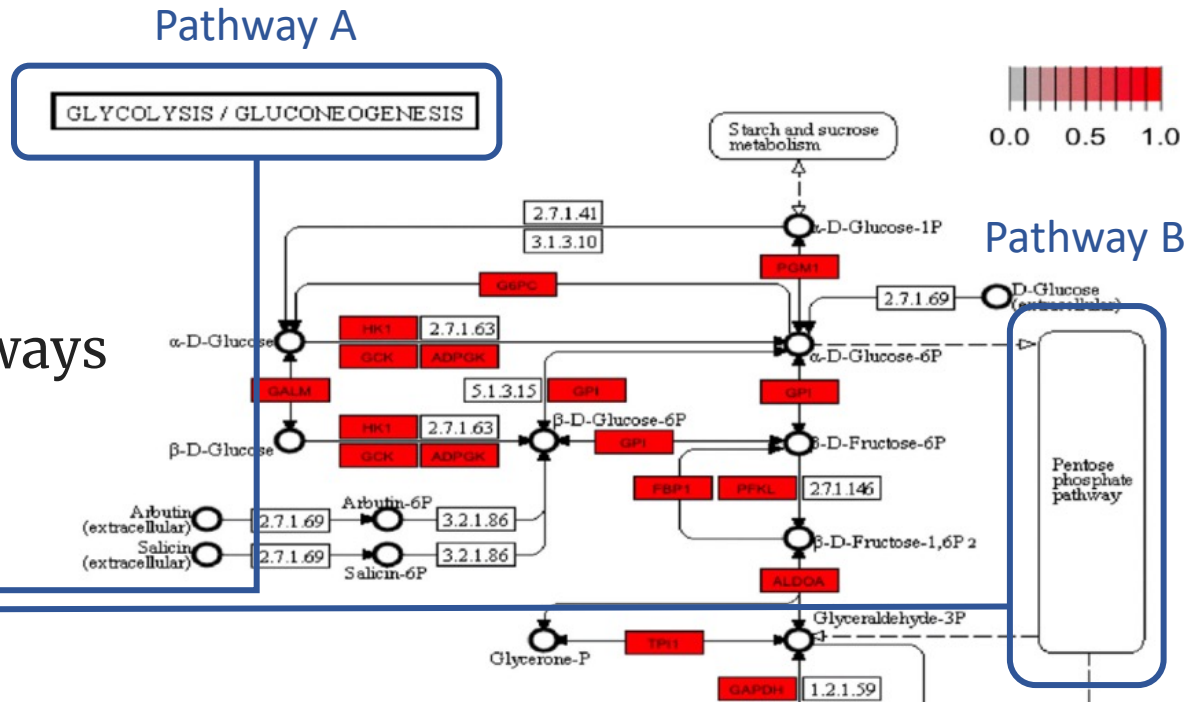
*PAFE: Pathway annotation-based feature extraction

PAFE

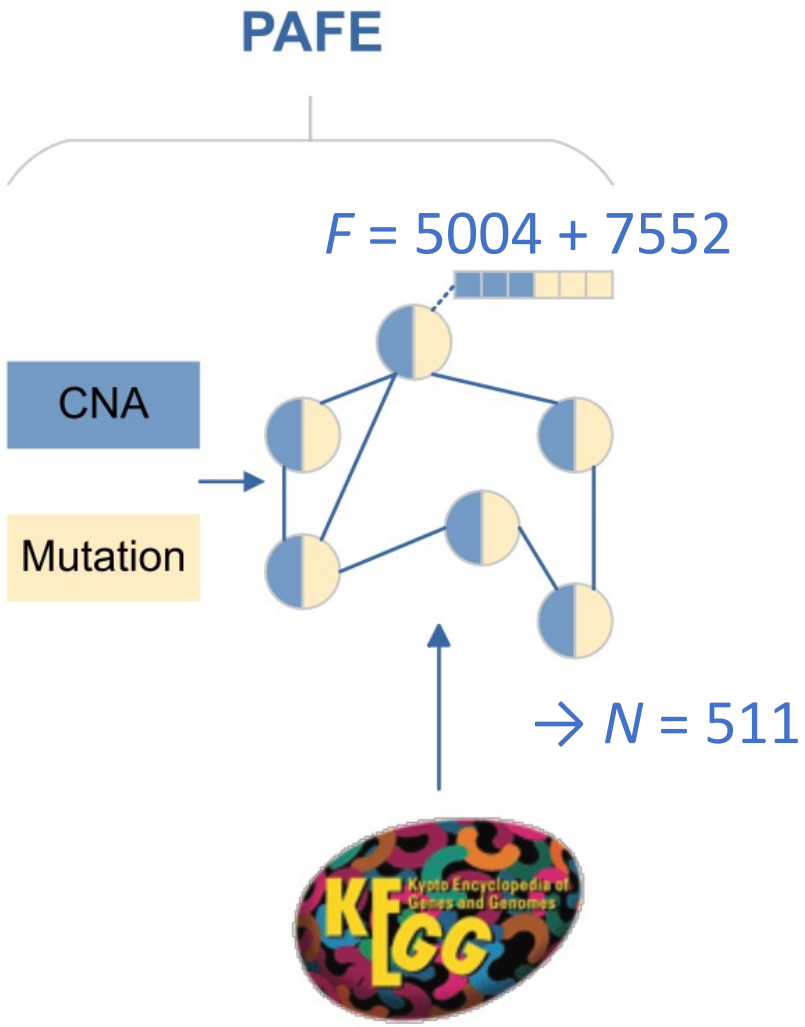


- Pathway annotation-based feature extraction
- Identify intersect gene set between omics data and KEGG pathways
  - CNA datasets: 5004 genes
  - Mutation datasets: 7552 genes
- Compose the graph $Graph = (V, E)$
  - V: Node set, pathway
  - E: Edge set, interaction between pathways
  - A: 511 * 511, adjacency matrix

$$A_{ij} = 1$$



PAFE



Initial Node

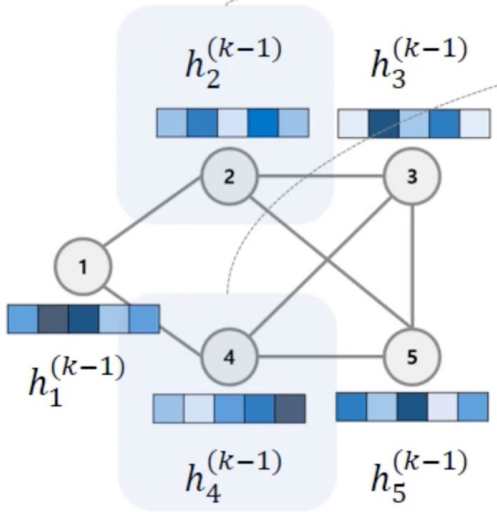
$$h = h_1, h_2, \ldots, h_N, h_i \in \mathbb{R}^F$$

N: Number of pathway

F: Number of genes(from CNA & Mutation)

Graph Neural Network

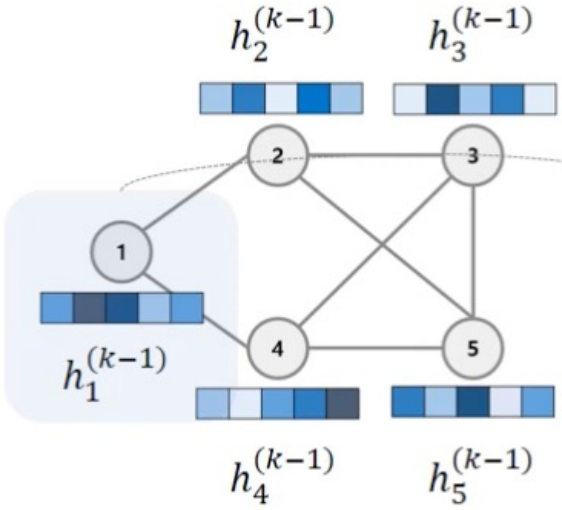
1. Aggregate(Message passing)



$$a_1^{(k-1)} = aggregate^k(\{ \quad , \quad \})$$

$$a_v^{(k-1)} = aggregate^k(\{h_u^{(k-1)}, \forall u \in N(v)\})$$

2. Combine(Update)



$$a_1^{(k-1)} = aggregate^k(\{ \quad , \quad \})$$

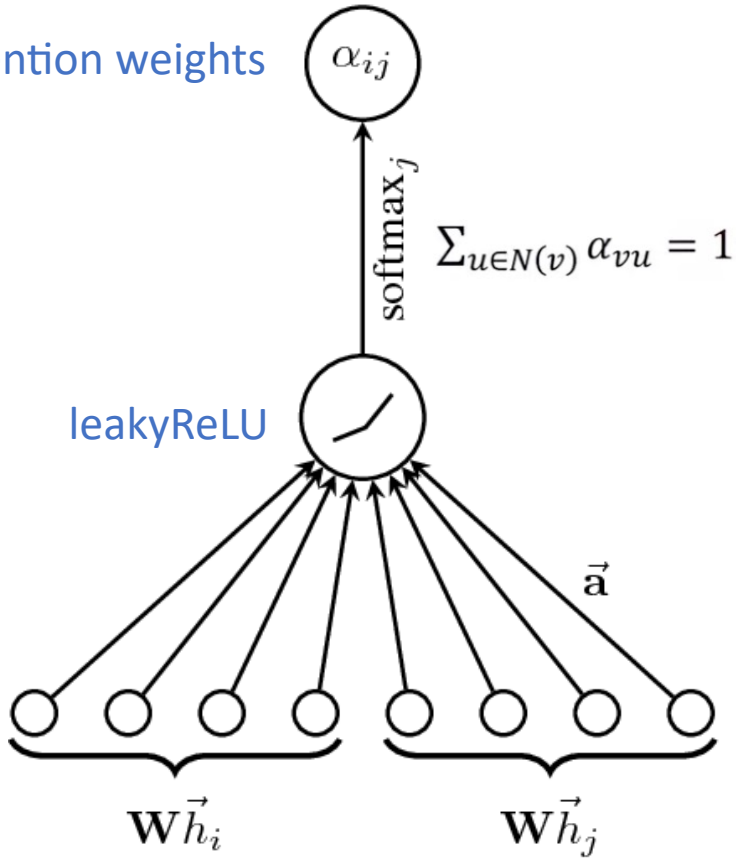
$$h_1^{(k)} = combine^k(a_1^{(k-1)}, \quad)$$

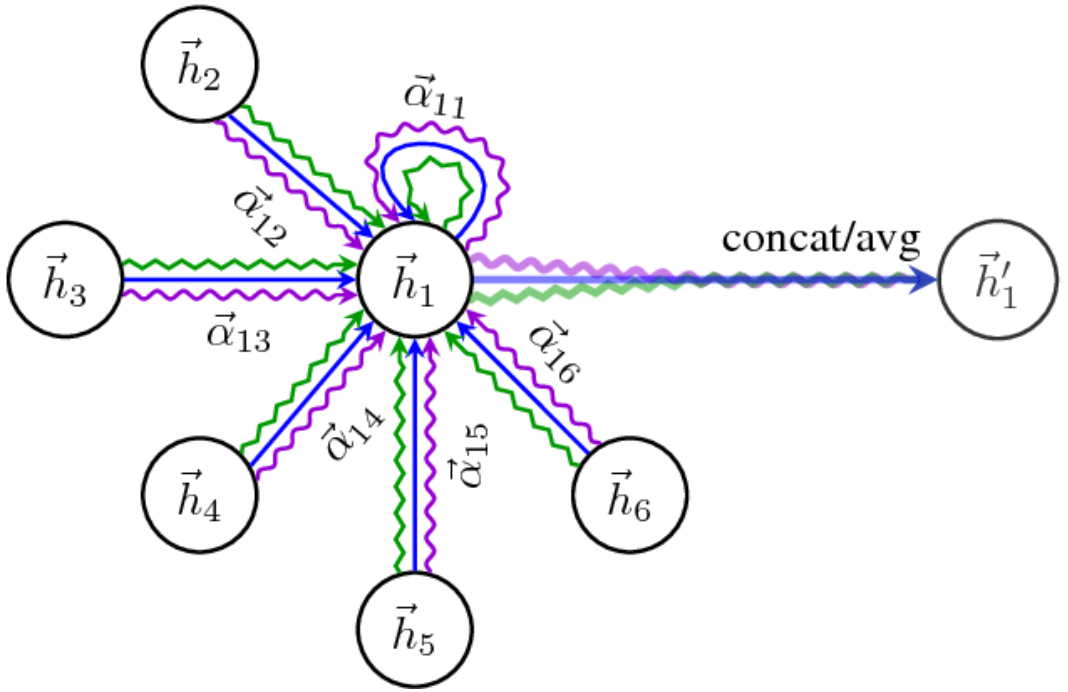
$$=$$

$$h_v^{(k)} = combine^k(a_u^{(k-1)}, h_v^{(k-1)})$$

Graph Attention Networks(GAT)

Attention weights





Different weights for each neighbor nodes

GAT on GraphPath

- Attention score

$$e_{ij} = softmax(LeakyReLU(a^T [Wh_i \| Wh_j])) \quad (1)$$

Concatenate

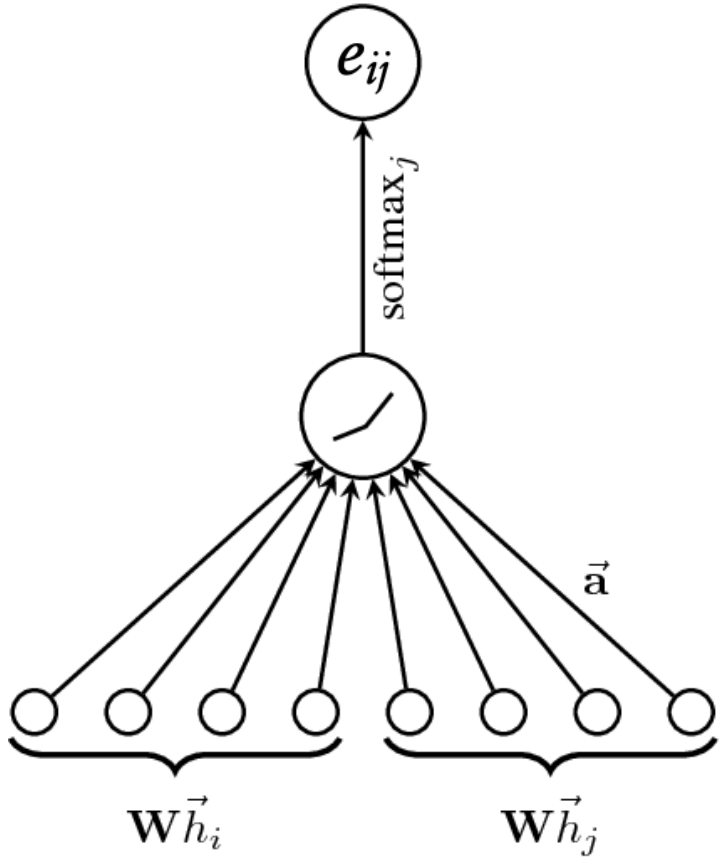
Trainable parameters

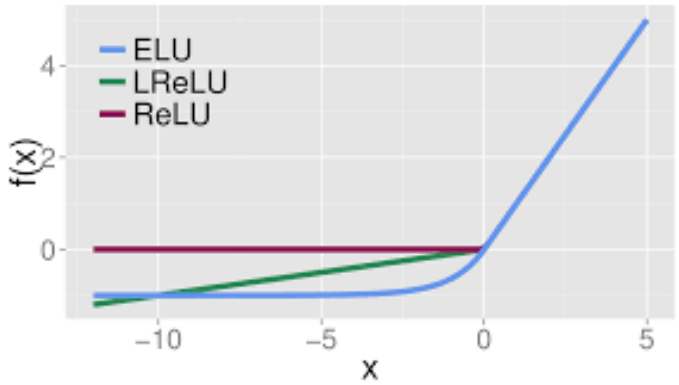
- Multi-head attention

$$h'_i = \|_{k=1}^{K} \sigma\left(\sum_{j \in N_i} e_{ij}^k W^k h_j\right) \quad (2)$$

Self-attention is repeated K times(this paper used 3)

$$\sigma = ELU(x) = \begin{cases} x, & x > 0 \\ \alpha(\exp(x) - 1), & x \leq 0 \end{cases}$$





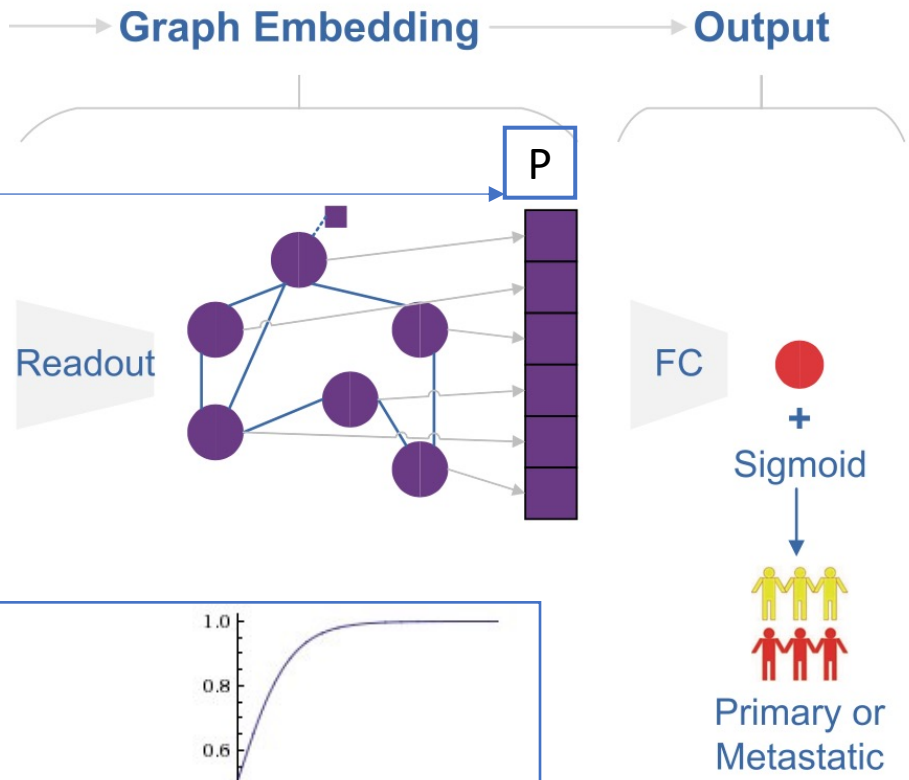
Graph embedding and output

- Summarizing graph representation

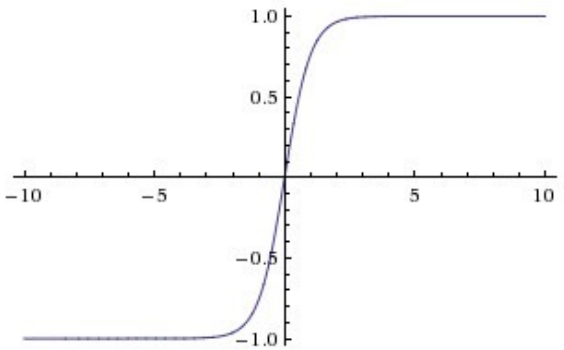
$$P_i = \mathrm{Tanh}(W^p h_i') \qquad (4)$$

- Generate the final prediction output

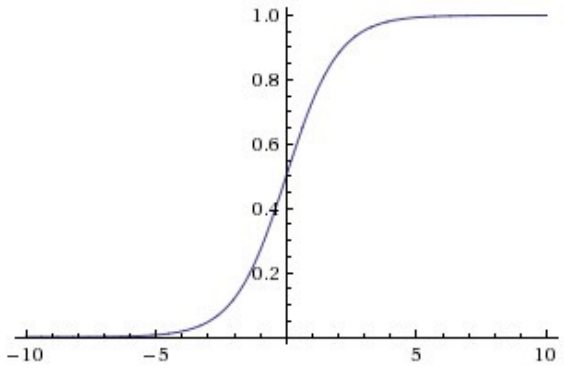
$$y = Sigmoid(W^y P) \qquad (5)$$



$$\mathrm{Tanh} = \frac{e^{2x}-1}{e^{2x}+1}$$



$$Sigmoid = \frac{1}{1+e^{-x}}$$



Training and testing

- ☑ Change random seeds in 30 times
- ☑ Split the dataset

- ☑ Binary cross-entropy loss function

$$BCELoss = \frac{1}{N}\sum_{i=1}^{N} y_j * \log(p(\hat{y}_i)) + (1-y_i) * \log(1-p(\hat{y}_i))$$

Results

Comparison with baseline methods

Table 2. Performance comparison of GraphPath with four methods based on the benchmark dataset.

Method	AUPR	AUC	F1	Recall	Precision	Accuracy
MLP	0.798 ± 0.066	0.872 ± 0.044	0.722 ± 0.067	0.704 ± 0.094	0.746 ± 0.058	0.821 ± 0.040
PPI-based	0.825 ± 0.058	0.899 ± 0.036	0.733 ± 0.078	0.701 ± 0.124	0.784 ± 0.072	0.833 ± 0.041
PAFE+MLP	0.863 ± 0.052	0.922 ± 0.031	0.777 ± 0.067	0.742 ± 0.091	0.823 ± 0.072	0.859 ± 0.040
P-NET	0.862 ± 0.052	0.905 ± 0.038	0.749 ± 0.054	0.671 ± 0.075	**0.857 ± 0.059**	0.851 ± 0.029
GraphPath	**0.887 ± 0.045**	**0.933 ± 0.028**	**0.808 ± 0.059**	**0.779 ± 0.074**	0.843 ± 0.070	**0.877 ± 0.038**

The best results are highlighted in bold.

- MLP: Fully connected neural network
- PPI-based: protein-protein interaction + pathway network
- PAFE + MLP: Sparse neural network based on the gene-pathway membership annotation
- The GraphPath outperformed → pathway interaction network has better performance

Table 3. *P*-values of comparing GraphPath to P-NET for six metrics.

	AUPR	AUC	F1	Recall	Precision	Accuracy
P-NET versus GraphPath	0.0500	0.0036	0.0007	3.75E-06	0.4352	0.0104

Simulation experiment

- Make a graph with 800 pathways by R igraph package
- Considering the graph as known prior knowledge, get a variance-covariance matrix M
- Generate a multivariate normal distribution with mean 0 and M
- Among them, 20 pathways have nonlinear effect

$$Y_i = \sum_{j=1}^{20} \{0.05X_{ij}^2 + 0.01X_{ij}^3 + \exp(-X_{ij}^2/10)\}$$

- Finally, obtained 566 positive samples and 377 negative samples with a threshold

Table 4. Performance comparison between GraphPath and MLP on simulation dataset.

Method	AUPR	AUC	F1	Recall	Precision	Accuracy
MLP	0.866 ± 0.037	0.819 ± 0.042	0.786 ± 0.039	0.785 ± 0.048	**0.789 ± 0.046**	0.744 ± 0.048
GraphPath	**0.890 ± 0.029**	**0.840 ± 0.036**	**0.811 ± 0.028**	**0.843 ± 0.038**	0.783 ± 0.036	**0.764 ± 0.037**

The best results are highlighted in bold.

30 times
Monte-Carlo Cross Validation

Table 5. *P*-values of comparing GraphPath to MLP for six metrics.

	AUPR	AUC	F1	Recall	Precision	Accuracy
MLP versus GraphPath	0.0070	0.0016	6.16E-05	1.83E-09	0.2507	0.0031

Jialin Qu, Yuehua Cui, Gene set analysis with graph-embedded kernel association test, *Bioinformatics*, Volume 38, Issue 6, March 2022, Pages 1560–1567, https://doi.org/10.1093/bioinformatics/btab851

External validation

- Two independent prostate cancer cohorts as inputs
  - Fraser et al.(n = 130): primary prostate cancer
  - Robinson et al.(n = 40): metastatic prostate cancer

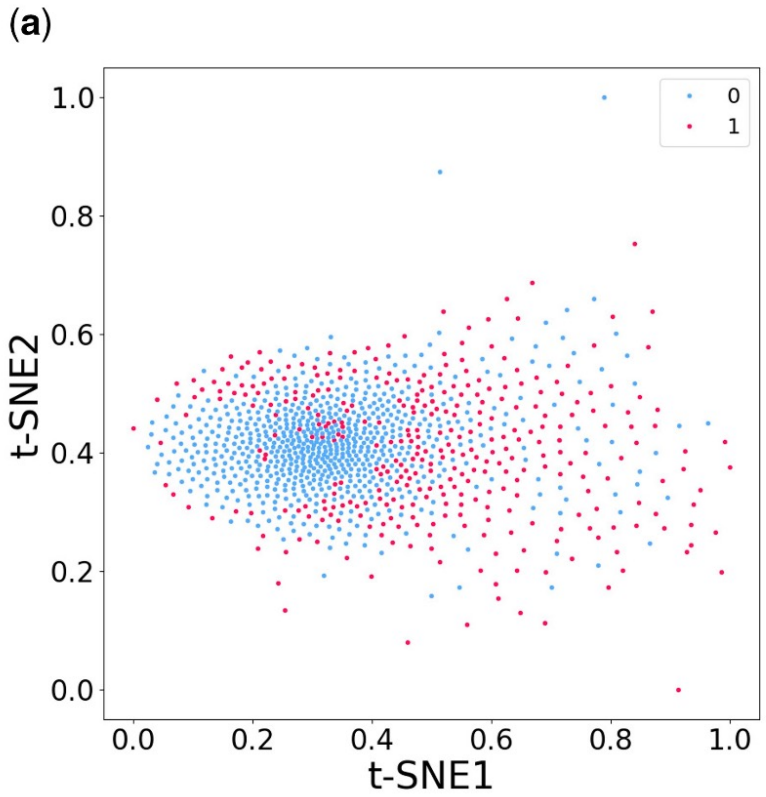
  → merged into a single test set
- For train set, used previous dataset with 333 negative and 333 positive samples

Table 6. Performance comparison of GraphPath with four methods based on the external validation datasets.

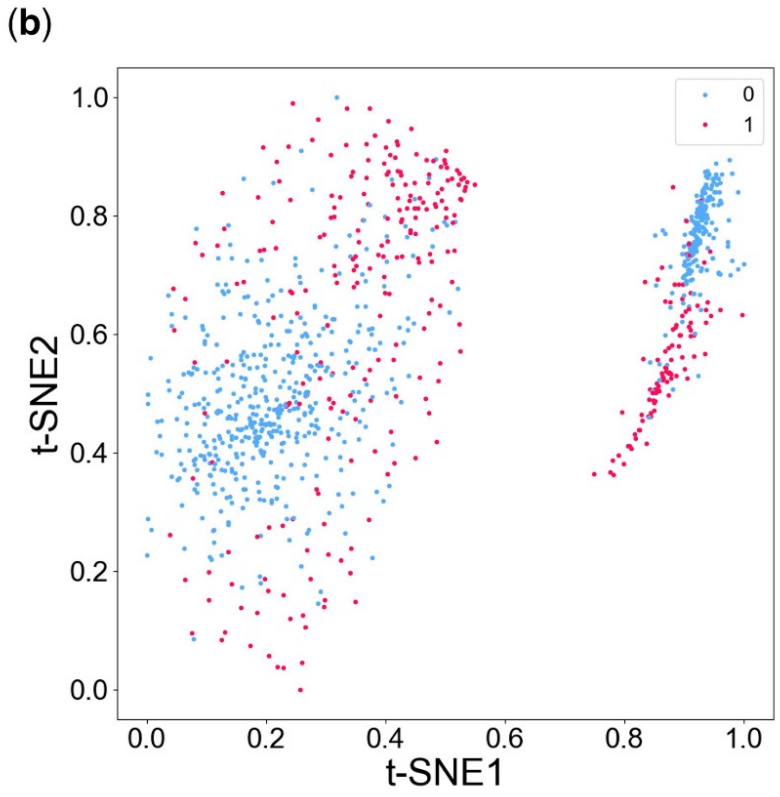
Method	TPR	TNR
MLP	0.725	0.731
PAFE+MLP	0.725	0.746
PPI-based	0.800	0.739
P-NET	0.800	0.754
GraphPath	**0.800**	**0.796**

The best results are highlighted in bold.

Cluster structure in t-SNE



Pathway initial feature matrix



Graph embedding

☑ Model successfully learns a meaningful latent representation

Interpretation of GraphPath

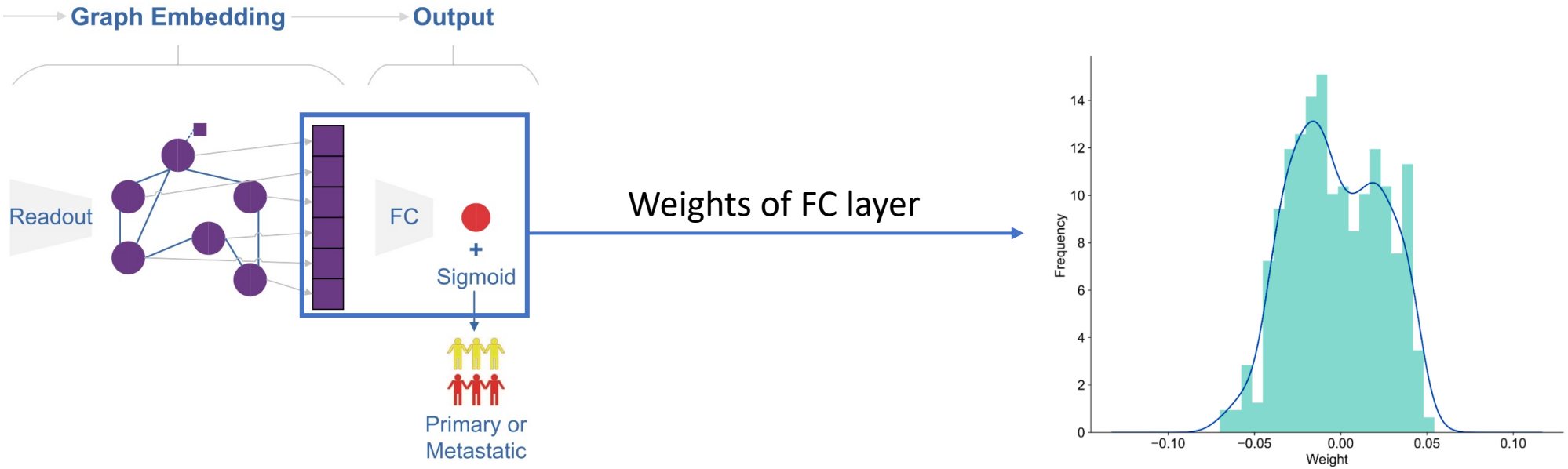
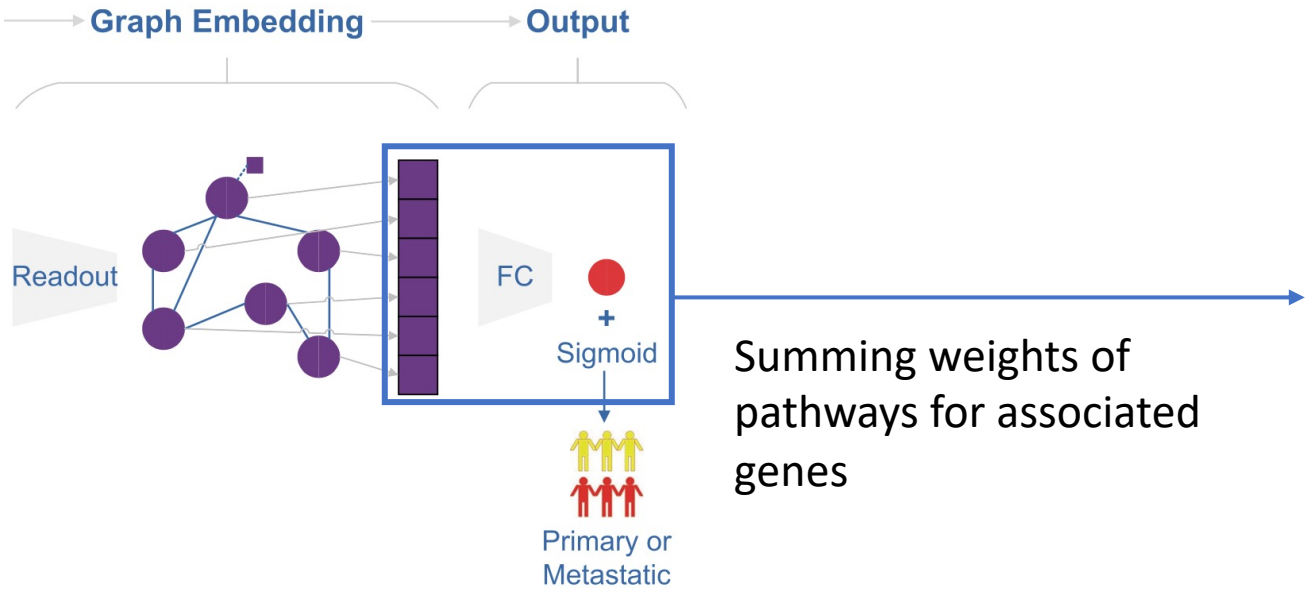


Table 8. The top 10 prostate cancer-related pathways inferred by GraphPath.

Rank	Pathway	Name	Evidence
1	ko05200	Pathway in cancer	
2	ko03000	Transcription factors	Pisano *et al.* (2021)
3	ko04131	Membrane trafficking	
4	ko05215	Prostate cancer	KEGG
5	ko05207	Chemical carcinogenesis-receptor activation	
6	ko03310	Nuclear receptors	Wang *et al.* (2021)
7	ko04147	Exosome	Akoto and Saini (2021)
8	ko03036	Chromosome and associated proteins	
9	ko04121	Ubiquitin System	Ummanni *et al.* (2011)
10	ko01001	Protein kinases	Shorning *et al.* (2020)

Interpretation of GraphPath



Summing weights of pathways for associated genes

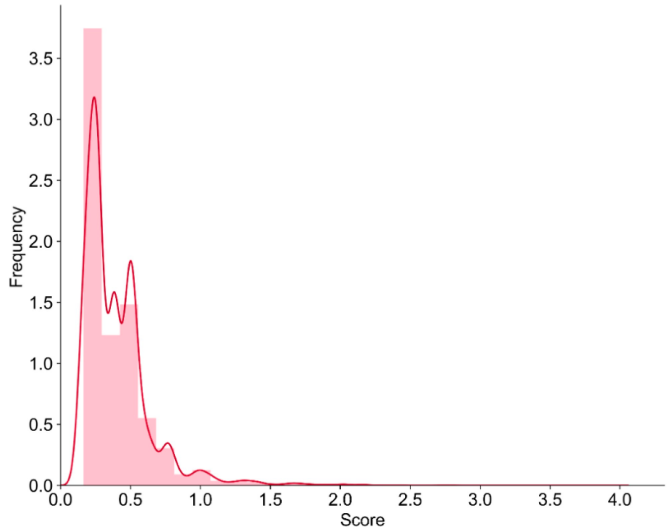


Figure 4. Contribution score distribution of the 6418 genes.

Table 9. The top 10 prostate cancer-related genes inferred by GraphPath.

Rank	Gene name	Aliases	Evidence
1	ERK	EPH Receptor B2	Wang *et al.* (2018)
2	MTOR	Mechanistic Target of Rapamycin Kinase	Wang *et al.* (2018)
3	HRAS	HRas Proto-Oncogene, GTPase	
4	NRAS	NRAS Proto-Oncogene, GTPase	
5	KRAS	KRAS Proto-Oncogene, GTPase	Shtivelman *et al.* (2014)
6	NFKB1	Nuclear Factor Kappa B Subunit 1	
7	RELA	RELA Proto-Oncogene, NF-KB Subunit	
8	TP53	Tumor Protein P53	Wang *et al.* (2018)
9	EP300	E1A Binding Protein P300	
10	RAF1	Raf-1 Proto-Oncogene, Serine/Threonine Kinase	Shtivelman *et al.* (2014)

Discussion

Discussion

- Cancers that share similar morphological characteristics often exhibit distinct treatment responses and prognoses.
- The research focuses on exploring how pathway knowledge can be used to design deep neural networks for cancer prediction and interpretation.
- The GraphPath achieves superior predictive performance in the task of cancer molecular stratification and provides interpretable predictions.
- Overall, GraphPath is a biological knowledge-driven graph neural network that encodes the pathway–pathway interaction network and achieves optimal performance in stratifying cancer status

Thanks!