OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf131
https://doi.org/10.1093/bib/bbaf131
Problem Solving Protocol

MethPriorGCN: a deep learning tool for inferring DNA methylation prior knowledge and guiding personalized medicine

Jie Ni[1,2,3], Bin Li[1,3], Shumei Miao[4], Xinting Zhang[1,3], Donghui Yan[1,3], Shengqi Jing[2,4,5], Shan Lu[6], Zhuoying Xie[1,3,*], Xin Zhang[4,*], Yun Liu[4,*]

[1]Institute for Molecular Medical Technology, State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, No. 2, Southeast University Road, Jiangning District, Nanjing, Jiangsu 211102, China
[2]Department of Medical Informatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, No. 101, Longmian Avenue, Jiangning District, Nanjing, Jiangsu 211166, China
[3]Institute of Biomedical Devices (Suzhou), Southeast University, No. 8, Jinfeng Road, Suzhou New District, Suzhou, Jiangsu 215163, China
[4]Department of Information, The First Affiliated Hospital, Nanjing Medical University, No. 300, Guangzhou Road, Gulou District, Nanjing, Jiangsu 210029, China
[5]Center for Data Management, The First Affiliated Hospital, Nanjing Medical University, No. 300, Guangzhou Road, Gulou District, Nanjing, Jiangsu 210029, China
[6]Women and Children Department, The First Affiliated Hospital, Nanjing Medical University, No. 300, Guangzhou Road, Gulou District, Nanjing, Jiangsu 210029, China

*Corresponding authors. Zhuoying Xie, Institute for Molecular Medical Technology, State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing, Jiangsu 211102, China. E-mail: zyxie@seu.edu.cn; Xin Zhang, Department of Information, the First Affiliated Hospital, Nanjing Medical University, Nanjing, Jiangsu 210029, China. E-mail: zx6800@njmu.edu.cn; Yun Liu, Department of Information, the First Affiliated Hospital, Nanjing Medical University, Nanjing, Jiangsu 210029, China. E-mail: liuyun@njmu.edu.cn

Abstract

DNA methylation plays a crucial role in human diseases pathogenesis. Substantial experimental evidence from clinical and biological studies has confirmed numerous methylation-disease associations, which provide valuable prior knowledge for advancing precision medicine through biomarker discovery and disease subtyping. To systematically mine reliable methylation prior knowledge from known DNA methylation-disease associations and develop robust computational methods for precision medicine applications, we propose MethPriorGCN. By integrating layer attention mechanisms and feature weighting mechanisms, MethPriorGCN not only identified reliable methylation digital biomarkers but also achieved superior disease subtype classification accuracy.

Keywords: DNA methylation; graph convolutional network; feature weighting; disease classification; biomarker discovery

Introduction

DNA methylation [1] is a crucial epigenetic modification that regulates gene expression and maintaining genomic stability in disease pathogenesis [2]. In cancer biology, abnormal DNA methylation patterns can lead to the silencing of tumor suppressor genes [3–5]. DNA methylation alterations often occur early in tumorigenesis, even before morphological or clinical manifestations, providing a critical window for early cancer detection. The stability of methylation patterns in cell-free DNA (cfDNA) from liquid biopsies further enables non-invasive screening approaches, reflecting tumor heterogeneity with high clinical feasibility [6]. These findings have catalyzed the development of targeted epigenetic therapies, with DNA methyltransferase inhibitors demonstrating clinical efficacy in hematological malignancies [7–11].

With the advancement of DNA methylation research, various disease-related DNA methylation patterns have been identified [12–14]. Recent computational advances have accelerated methylation analysis: For instance, MethylTree [15] leverages sparse single-cell methylation data for lineage tracing through graph-based algorithms, while the DNA Programmable Gate Array framework [16] demonstrates molecular computation architectures for scalable biological signal integration. Recent studies integrating multisource biological data, such as the MNDCDA model for circRNA-disease association prediction through deep feature projection networks, highlight the potential of multimodal approaches in uncovering complex molecular interactions [17]. Currently, numerous algorithms can infer potential prior knowledge from known associations across various omics [18–20]. For example, at the gene level, Xie *et al.* and Jia *et al.* proposed improved graph convolutional network (GCN) methods to predict potential gene-disease prior knowledge from a large number of known gene-disease associations, providing reliable prior knowledge for identifying potential gene targets of diseases [21]. Similarly, frameworks like scMCGraph, which constructs consensus cell–cell graphs by integrating pathway activity and gene expression for robust cell type annotation, exemplify the power of combining heterogeneous biological features to enhance model generalizability [22]. In the transcriptome, many effective methods for predicting potential prior knowledge for microRNA [23–27], long non-coding RNA [28–32], and circular

Received: January 20, 2025. **Revised:** February 27, 2025. **Accepted:** March 5, 2025

RNA [33–37] have been proposed, offering important references for precision medicine research [38, 39]. Notably, knowledge graph-based models such as BioKG-CMI (circRNA-miRNA interaction), which employs multi-feature fusion and subcellular localization to predict CMIs, further underscore the importance of embedding biological logic into computational frameworks [40].

However, existing models face limitations in methylation-specific contexts. While MethylTree excels in lineage tracing, its focus on clonal memory limits applicability to disease biomarker discovery. Recent studies on immunodeficiency, centromeric instability, and facial anomalies (ICF) syndrome [41] further highlight the complexity of dynamic methylation regulation in non-B DNA structures, a challenge overlooked by current computational tools. Despite significant progress at the gene and transcriptome levels, the more specific field of DNA methylation still faces challenges [42, 43]. The inference of reliable potential prior knowledge from relatively few known meth-disease associations and utilize the prior knowledge to guide clinical precision medicine remains a topic worthy of in-depth research. In other fields, such as physics, the use of prior knowledge is already well-established [44, 45]. For example, Physics-Informed Neural Networks [46–48] and Physics-Informed Convolutional Networks [49, 50] use physical laws and formulas to guide neural network training, achieving the integration of physical prior knowledge [51–53]. In biomedicine, pioneering works like mitoBEs exemplify knowledge-driven design in gene editing [54], yet analogous approaches for methylation-based machine learning remain nascent. While interdisciplinary integration has been successful in fields like physics, the use of prior knowledge is not yet very deep in the biomedical field, with only initial attempts being made, and few machine learning models embedding biological knowledge in some form [55].

Therefore, we developed a deep learning model based on GCN [56, 57], MethPriorGCN, to predict potential prior knowledge and deeply integrate it with DNA methylation data, thereby improving the identification of disease subtypes and important biomarkers [58]. MethPriorGCN includes two GCN-improved models, Prior knowledge Inferring GCN (PIGCN) and GCN for Disease Classification (FWGCN). PIGCN uses a layer attention mechanism to infer potential prior knowledge from meth-disease heterogeneous networks. Meanwhile, FWGCN utilizes potential prior knowledge and feature weighting strategies to drive downstream tasks, achieving precise integration of prior knowledge with clinical precision medicine. Through compared to existing prior prediction methods on the reliability of potential prior knowledge, literature research validation of identified digital methylation biomarkers, and five-fold cross-validation of disease subtype classification, we demonstrated the superior performance of MethPriorGCN in predicting potential prior knowledge and using them to drive precision medicine.

Materials and method

Known meth-disease association

The known meth-disease association serves as a source for inferring potential prior knowledge of meth-disease associations. We selected DiseaseMeth 3.0 as our database, which aggregates DNA methylation-disease data from public sources and literature [59]. This comprehensive database contains pre-processed statistical results across multiple samples. We filtered entries with *P*-values lower than or equal to 0.001, yielding 15 374 validated associations between 60 diseases and 2044 DNA methylations.

Table 1. Overview of datasets. The table encapsulates data on methylation expression. It specifically highlights methylations that are concurrently present in both methylation expression datasets and meth-disease association datasets.

Dataset	Categories	Number of meths and shared meths
ESCA	ESCC: 94, EAC: 89	2000, 308
THCA	Classical: 398, Other: 102	2000, 412
UCEC	EEA: 311, MSEAC: 21, SEA: 98	2000, 273

DNA methylation expression datasets

To validate the clinical application of MethPriorGCN, we selected the large-scale pan-cancer multi-omics dataset from The Cancer Genome Atlas (TCGA) as our research data. Considering the potential impact of noise and redundant features on overall quality, we used the independent preprocessing method by Lu *et al.* and preselected features for each omics category [60]. To ensure the effectiveness of MethPriorGCN, we conducted text comparisons and selected three disease datasets that exactly matched the disease names in the relevant data: ESCA (Esophageal Carcinoma), THCA (Thyroid Carcinoma), and UCEC (Uterine Corpus Endometrial Carcinoma). The ESCA dataset was divided into Esophagus Squamous Cell Carcinoma (ESCC) and Esophagus Adenocarcinoma. The THCA dataset was divided into classical/common (papillary NOS) and other subtypes. The UCEC dataset was classified into Endometrioid Endometrial Adenocarcinoma (EEA), Mixed Serous and Endometrioid Adenocarcinoma (MSEAC), and Serous Endometrial Adenocarcinoma (SEA) types. Table 1 shows the number of patients and label distribution for these datasets.

Gaussian interaction profile kernel similarity for meth and disease

Operating under the premise that methylations with analogous functions tend to correlate with similar diseases, the Gaussian interaction profile kernel function was employed to ascertain the kernel similarity for methylations and diseases [61, 62]. Specifically, in the known meth-disease association matrix A, the i^{th} row $I(d_i)$ and the j^{th} column $I(m_j)$ was defined as interaction profiles for disease d_i and methylation m_j, respectively. Then, the Gaussian interaction kernel similarity $SGD(d_i, d_j)$ for disease between d_i and d_j can be calculated as follows:

$$SGD(d_i, d_j) = \exp\left(-\rho_d \left\| I(d_i) - I(d_j) \right\|^2\right). \quad (1)$$

Also, Gaussian interaction profile kernel similarity $SGM(m_i, m_j)$ for methylation between m_i and m_j can be calculated as follows:

$$SGM(m_i, m_j) = \exp\left(-\rho_m \left\| I(m_i) - I(m_j) \right\|^2\right), \quad (2)$$

where $\exp(\cdot)$ represents the exponential function with natural constant e as the base, and the content in parentheses is the power, $\left\| I(d_i) - I(d_j) \right\|^2$ and $\left\| I(m_i) - I(m_j) \right\|^2$ represent the Euclidean norm of $I(d_i) - I(d_j)$ and $I(m_i) - I(m_j)$, respectively, ρ_d and ρ_m represent the normalized kernel bandwidth of disease and methylation, respectively. Moreover, ρ_d and ρ_m can be calculated by the following method:

$$\rho_d = \rho'_d \Big/ \left((1/nd) \sum_{i=1}^{nd} \left\| I(d_i) \right\|^2 \right), \quad (3)$$

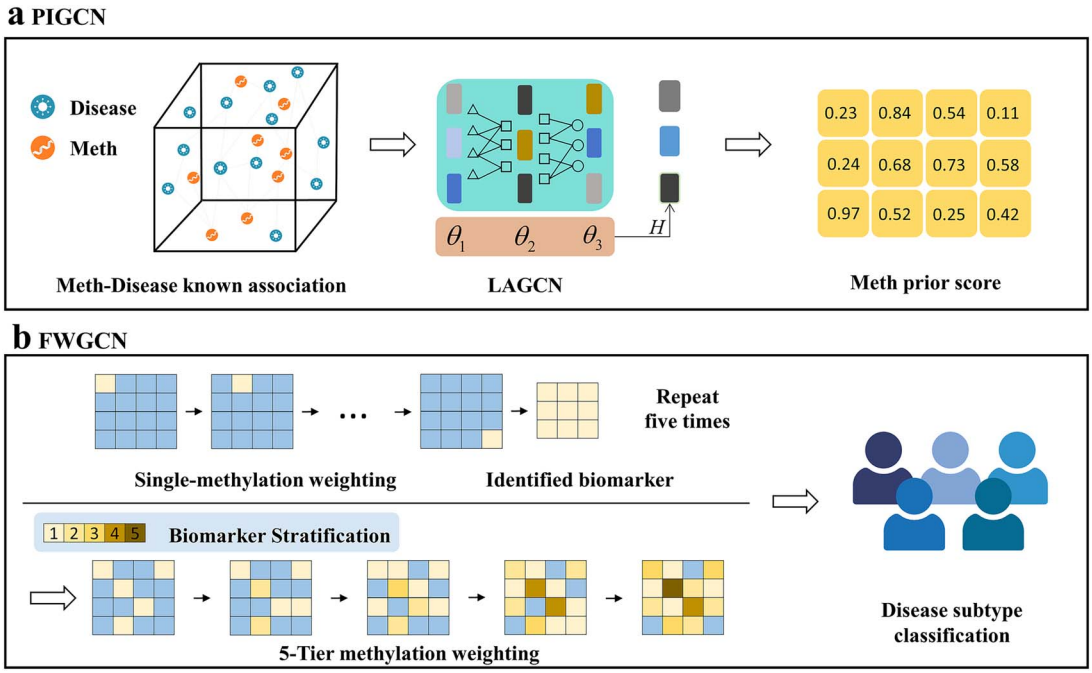

Figure 1. The flowchart of MethPriorGCN.

$$\rho_m = \rho'_m \Big/ \left((1/nm) \sum_{i=1}^{nm} \left\| I\left(d_i\right) \right\|^2 \right), \tag{4}$$

where nd represents the number of diseases, nm represents the number of methylations. Based on previous studies [63], the original bandwidths ρ'_d and ρ'_m was set to 1.

MethPriorGCN

MethPriorGCN primarily consists of two distinct GCNs: Prior knowledge Inferring GCN (PIGCN) integrates known meth-disease associations, methylation Gaussian kernel similarities, and disease Gaussian kernel similarities to construct a heterogeneous network, generating a potential prior score matrix for meth-disease relationships. Feature Weighting GCN (FWGCN) employs a feature weighting layer to identify critical methylations as digital biomarkers. These biomarkers are enhanced using prior scores and fed into FWGCN for disease subtype prediction. The workflow is illustrated in Fig. 1.

PIGCN

GCN models methylation-disease associations by capturing dynamic node relationships in graphs [64]. The proposed PIGCN enhances the Layer Attention GCN (LAGCN) framework by integrating a Meth-Disease Association Masking (MDAM) mechanism to improve prediction reliability. Specifically, LAGCN first infers preliminary potential associations, followed by MDAM-driven local normalization. Due to LAGCN's established utility in prior tasks [65, 66], its implementation details are provided in Supplementary Methods 1. The process of deriving preliminary potential prior knowledge for methylation-disease using LAGCN can be summarized as follows:

$$A^* = \mathrm{LAGCN}\left(SM, A, SD\right). \tag{5}$$

Given the inherent sparsity of the meth-disease association matrix, the initial potential prior A^* maintains scores of 1 (or near 1) at known association positions, leading to disproportionately low normalized scores at unknown positions. To address this, PIGCN introduces the MDAM module to record known association positions. During normalization, only unknown positions are processed and subsequently merged with known associations to form the final potential score matrix. Specifically, we first calculate the matrix A' that retains only the unknown associations:

$$A^{**} = A^* \odot (1 - A). \tag{6}$$

The symbol $\odot$ represents element-wise multiplication, followed by normalization of the non-known associated elements for each column (i.e., each disease). The normalization factor for column j is given by:

$$S_j = \sum_{i=1}^{m} A_{ij}^{**}. \tag{7}$$

If $S_j = 0$, set it to 1 (to avoid division by zero errors). Then, normalize the non-known associated values for each column as follows:

$$\tilde{A}_{ij} = \begin{cases} \frac{A_{ij}^{**}}{S_j}, & if\ S_j > 0 \\ 0, & else \end{cases}. \tag{8}$$

The final normalized matrix is:

$$A' = \tilde{A} + M, \tag{9}$$

where M ensures that the originally known association of 1 is retained, while $\tilde{A}$ contains the normalized values of the unknown associations.

FWGCN

In FWGCN, we evaluated MethPriorGCN's enhancement effect on disease subtype classification. First, we aligned the meth-disease association data with methylation expression profiles to identify shared methylations. The expression data were weighted using corresponding potential prior scores by Eq. (10), optimized with

L2 regularization to reduce feature weights and overfitting. For a given expression datum $p_{i,j}$ between the i_{th} methylation and the j_{th} disease, it was weighted according to $a_{i,j}$, which represents the association score between the i_{th} methylation and the j_{th} disease, to yield the updated expression datum $p'_{i,j}$, delineated as follows:

$$p'_{i,j} = 2p_{i,j}A^{k}_{i,j}\,\mathbf{exp}\,\left(\delta\left|I\left(p_{i,j}\right)\right|\right), k \in \{1,2,3,4,5\}. \tag{10}$$

where k denotes the importance level of digital methylation markers selected through five-fold cross-validation screening. The framework employs a multiplicative factor of 2 for weighting: methylation abundance metrics decrease when the potential association score is below 0.5, increase when exceeding 0.5, and remain unchanged at 0.5. We incorporated L1 regularization with coefficient δ (determined via five-fold cross-validation) to constrain feature weights through the exponential term $\mathbf{exp}\,(\cdot)$, effectively mitigating overfitting.

FWGCN

After obtaining the weighted methylation expression data, the data were put into the GCN for Disease Classification (FWGCN). The performance of FWGCN in supervised learning applications was then explored by combining the labels of each sample. Specifically, the propagation rule of FWGCN is shown as follows:

$$H^{(l+1)}_{mp} = f\left(H^{(l)}_{mp}, G_{mp}\right) = \sigma\left(D^{-\frac{1}{2}}_{mp}G_{mp}D^{-\frac{1}{2}}_{mp}H^{(l)}_{mp}W^{(l)}_{mp}\right), \tag{11}$$

where G_{mp} is constructed by calculating the cosine similarity between pairs of nodes, and edges with cosine similarity larger than a threshold ϵ are retained. Specifically, $G_{mp}[i,j]$ is the adjacency between the i_{th} node and the j_{th} node in the graph, can be calculated as follows:

$$G_{mp}[i,j] = \begin{cases} s\left(\frac{x_i \cdot x_j}{\|x_i\|_2\|x_j\|_2}\right), & \text{if } i \neq j \text{ and } s\left(\frac{x_i \cdot x_j}{\|x_i\|_2\|x_j\|_2}\right) \geq \epsilon \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where x_i and x_j are the feature vectors of the i_{th} node and the j_{th} node in P, The threshold ϵ is determined given a parameter γ, which represents the average number of edges per node that are retained, including self-connections as follows:

$$\gamma = \sum_{i,j} I\left(s\left(\frac{x_i \cdot x_j}{\|x_i\|_2\ \|x_j\|_2}\right) \geq \epsilon\right)/n, \tag{13}$$

where $I(\cdot)$ is an indicator function and n is the number of nodes. In this study, the application of FWGCN is further extended to supervised classification tasks. For the training data X_{tr}, the corresponding adjacency matrix G^{tr}_{mp} can be computed from Eq. (12). Then, a GCN can be used to train on X_{tr} and G^{tr}_{mp}, and the predictions for the training data can be expressed as follows:

$$\hat{Y}_{tr} = \mathrm{GCN}\left(X_{tr}, G^{tr}_{mp}\right), \tag{14}$$

where the i_{th} row of $\hat{Y}_{tr}$ represents the predicted label probability for the i_{th} training sample, and the number of columns of $\hat{Y}_{tr}$ indicates the number of classes in the classification task. Thus, FWGCN utilizes both the features and the geometric structure of the training data to learn the classification task. For a new test sample X_{te}, it is added to the training set, extending X_{tr} to $X_{tr,te}$, and generate the extended adjacency matrix $G^{tr,te}_{mp}$ according to Eq. (12). Specifically, the entries in the last row and column of $G^{tr,te}_{mp}$ are the only ones computed during testing, reflecting the association between the test sample X_{te} and the training samples X_{tr}. Therefore, given $X_{tr,te}$, $G^{tr,te}_{mp}$ and the trained model $GCN(\cdot)$, $\hat{Y}_{tr,te} = GCN\left(X_{tr,te}, G^{tr,te}_{mp}\right)$ can be obtained. The predicted label probability distribution for the test sample is the last row of $\hat{Y}_{tr,te}$. Hence, by exploiting both the features of the test sample and the association between the test sample and the training samples, the label of the new test sample X_{te} can be predicted. Overall, the loss function of FWGCN can be written as follows:

$$L_{mp} = \sum_{j=1}^{n_{tr}} L_{CE}\left(\hat{y}_j, y_j\right) = \sum_{j=1}^{n_{tr}} -\log\left(\frac{e^{\hat{y}_j \cdot y_j}}{\sum_k e^{\hat{y}_{j,q}}}\right), \tag{15}$$

where $L_{CE}(\cdot)$ denotes the cross-entropy loss function. y_j is the true label of the j_{th} training sample, and $\hat{y}_{j,q}$ is the q_{th} element of the vector $\hat{y}_j$.

Parameter settings

The value of δ was determined within a specified range of $\delta \in \{0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3\}$ through the five-fold cross-validation. In the intermediate fusion model FWGCN, the parameter $\gamma = 2$ and $\gamma = 10$ was set for the binary classification task and the multi-classification task by adjusting the parameters empirically, respectively.

Results

Meth-disease potential prior score matrix

Within our research scope, we processed a methylation-disease heterogeneous network using PIGCN, resulting in a potential prior score matrix that depicts the potential prior scores of methylation-disease associations. The structure of this matrix is such that individual rows represent different methylations, while columns correspond to various diseases, with the 5th, 22nd, and 30th columns specifically representing Esophageal carcinoma, Thyroid carcinoma, and Uterine Corpus Endometrial Carcinoma.

Before utilizing the resulting meth-disease potential prior score matrix for subsequent studies, we first validated the reliability of the potential prior scores by calculating the area under the receiver operating characteristic curve (AUC) and employing five-fold cross-validation. Specifically, all known meth-disease associations were randomly divided into five approximately equal parts. Each time, one part of the known associations was used as the test sample, the remaining four parts as the training sample, and all unknown meth-disease pairs as candidate samples. After sample division, PIGCN scored the test and candidate samples. Based on the scores, the ranking of the test samples among all candidate samples was determined. If the ranking of the test samples was above a set threshold, PIGCN was considered to have made an accurate prediction. Subsequently, based on the rankings of all test samples, ROC curves under different thresholds were plotted, and the AUC was calculated by comparing true positive rates and false positive rates across varying thresholds to evaluate the accuracy of the association scores. Previous studies had determined some basic parameters in this model; therefore, we considered only two key parameters: total training steps $\alpha \in \{200, 400, 600, 800\}$ and learning rate $lr \in \{0.0001, 0.001, 0.01, 0.1\}$. By implementing a grid search within this range, we performed five-fold cross-validation for each round of search and recorded the average values. As shown in Fig. 2, higher training steps (>2000) improved AUC by allowing the model to capture deeper network interactions, while excessive steps

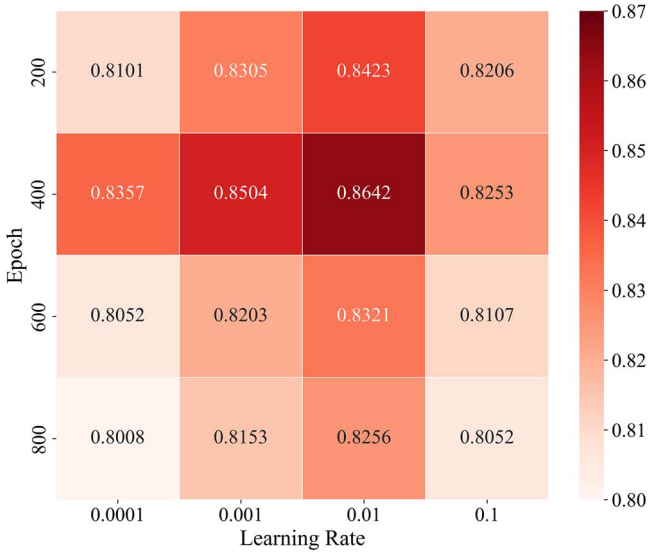

Figure 2. AUC values for different training steps and learning rates.

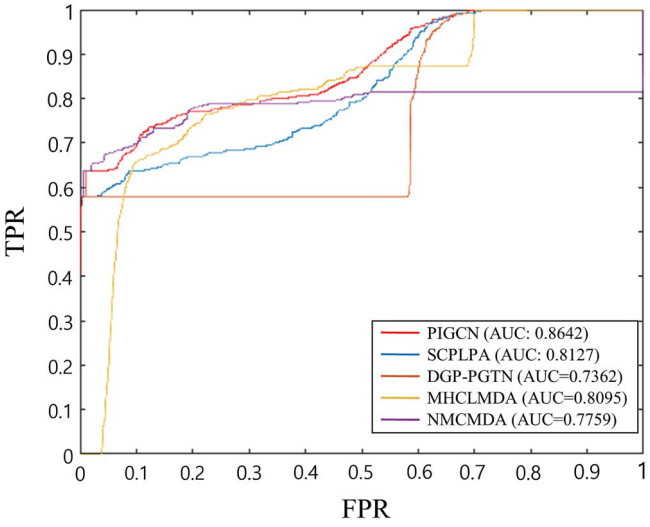

Figure 3. ROC curves of PIGCN and four comparison algorithms.

(>3000) led to marginal gains, indicating convergence. For learning rate, values between 0.001 and 0.005 balanced training stability and efficiency. Finally, we plotted the results as a heatmap in Fig. 2. By comparing the AUC values, we selected the values of $lr = 0.01$ and $\alpha = 400$ through grid search optimization.

Additionally, we plotted the ROC curves of the PIGCN results and the comparison algorithms, specifically NMCMDA (2021) [67], MHCLMDA (2023) [68], DGP-PGTN (2023), [69], and SCPLPA (2024), [70], and compared the AUC of PIGCN with these algorithms, as shown in Fig. 3. The superior AUC of PIGCN arises from its ability to integrate both topological and biological features in the methylation-disease network. For example, the graph attention mechanism dynamically prioritizes interactions involving hub methylation sites linked to multiple cancers, while suppressing noisy connections from unrelated nodes. Additionally, iterative message passing captures multi-hop dependencies, such as the regulatory chain in FHIT, MAPK pathway, and Esophageal carcinoma, which shallow models fail to model [71]. The results indicated that the accuracy of the potential prior scores predicted by PIGCN was significantly higher than that of several selected comparison algorithms, further demonstrating the advancement of PIGCN and the reliability of the predicted potential prior scores.

Performance evaluation of disease classification

The digital meth biomarker screening employed five-fold cross-validation. After splitting data into 80% training and 20% testing, the training set was further divided into five equal folds. In each iteration, four folds trained the model and one validated. Shared methylation markers were iteratively weighted by their prior scores. Markers improving validation ACC versus unweighted baseline were retained as candidate biomarkers. This generated five biomarker lists post cross-validation. Markers appearing in all five lists (consistently enhancing performance) were prioritized and applied with five prior score weightings during testing, while those absent in all lists were excluded.

After screening for digital methylation markers, we applied feature weighting to the methylation expression dataset to improve the accuracy of distinguishing disease subtypes. Notably, the weighting process was based on the level of digital methylation markers. Specifically, level 5 digital methylation markers are the most important, so the weighting process for level 5 digital markers was performed five times, further enhancing the importance of high-level digital methylation markers in the model. We compared the classification performance of MethPriorGCN with the following four existing omics data classification algorithms: (i) Adaboost: It builds a strong model by combining several weak models, focusing on errors each time. (ii) Naive Bayes: It predicts class probabilities based on feature likelihoods using Bayes' theorem. (iii) Highly trustworthy multiomics learning (HTML) framework [60]: Developed by Lu *et al.*, HTML employs multi-omics adaptive dynamic learning to process each sample with a data-dependent architecture and computational workflow. (iv) IBPGNET [72]: Developed by Xu *et al.*, IBPGNET is an interpretable network for disease recurrence prediction and related biomarker discovery. HTML and IBPGNET have been proven to outperform nine benchmark methods across various omics data. Therefore, we selected these methods as comparative algorithms on DNA methylation expression data to demonstrate the superior performance of MethPriorGCN.

The performance of MethPriorGCN was evaluated using three metrics: for tumors classified into two subtypes, we used accuracy (ACC), F1 score (F1), and AUC; for malignant tumors classified into three subtypes, we used ACC, support-weighted average F1 score (F1_weighted), and macro-average F1 score (F1_macro). Specific formula definitions for all evaluation indicators are given in Supplementary Method 1.

We evaluated the effectiveness of the MethPriorGCN algorithm and comparison algorithms through five-fold cross-validation. The dataset was divided into five equal-sized subsets for cross-validation, with one subset used as the validation set and the remaining four as the training set in each iteration. The mean and standard deviation were then reported. Comparative analysis across multiple datasets demonstrated that MethPriorGCN outperformed other benchmark algorithms on all evaluation metrics, validating the effectiveness of digital methylation markers and confirming their advantage in disease subtype classification. The notably higher AUC in ESCA suggests MethPriorGCN's enhanced ability to distinguish subtle epigenetic differences between subtypes, potentially driven by prioritizing methylation markers in critical regulatory regions (e.g., tumor suppressor promoters). In THCA, the improved ACC aligns with better identification of dominant methylation-driven pathways, while the moderate F1 score (0.4787) may reflect challenges in resolving rare subtypes with overlapping methylation profiles. For UCEC, the higher F1_weighted (0.8167) compared to F1_macro (0.5531) indicates

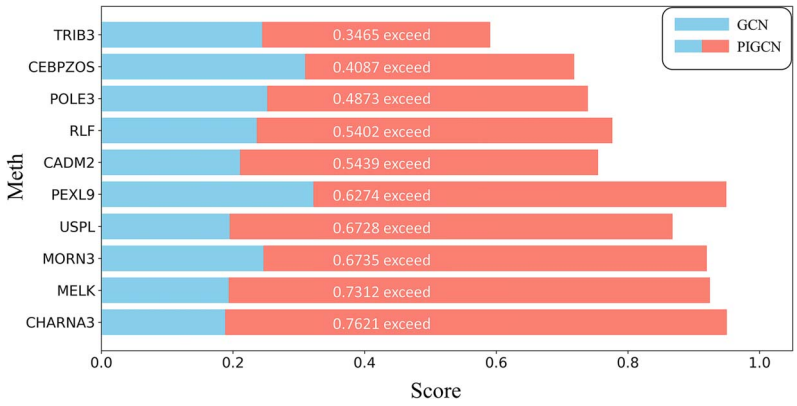

Figure 4. Top 10 potential associated methylations for esophageal carcinoma.

Table 2. Classification results on ESCA dataset.

Method	ACC	F1	AUC
Adaboost	0.8110 ± 0.0382	0.7830 ± 0.0890	0.8177 ± 0.0561
Naive Bayes	0.8465 ± 0.0414	0.8029 ± 0.0719	0.8057 ± 0.0656
HTML	0.8965 ± 0.0454	0.8871 ± 0.0852	0.8726 ± 0.0680
IBPGNET	0.9007 ± 0.0170	0.8768 ± 0.0338	0.9271 ± 0.0124
MethPriorGCN	0.9479 ± 0.0188	0.9460 ± 0.0213	0.9880 ± 0.0157

Table 3. Classification results on THCA dataset.

Method	ACC	F1	AUC
Adaboost	0.7056 ± 0.0331	0.3271 ± 0.0768	0.6501 ± 0.0541
Naive Bayes	0.6965 ± 0.0422	0.3572 ± 0.0350	0.6940 ± 0.0405
HTML	0.7480 ± 0.0310	0.3670 ± 0.0645	0.7107 ± 0.0525
IBPGNET	0.7410 ± 0.0464	0.4260 ± 0.0401	0.7031 ± 0.0514
MethPriorGCN	0.8060 ± 0.0490	0.4787 ± 0.1032	0.7993 ± 0.0573

Table 4. Classification results on UCEC dataset.

Method	ACC	F1_weighted	F1_macro
Adaboost	0.7161 ± 0.0667	0.7851 ± 0.0793	0.7282 ± 0.0745
Naive Bayes	0.7342 ± 0.0692	0.7730 ± 0.0738	0.6297 ± 0.0821
HTML	0.7757 ± 0.0610	0.7148 ± 0.0767	0.7719 ± 0.0813
IBPGNET	0.8302 ± 0.0448	0.8001 ± 0.0525	0.7689 ± 0.0524
MethPriorGCN	0.8272 ± 0.0286	0.8167 ± 0.0330	0.5531 ± 0.0254

robust performance in prevalent subtypes but highlights potential limitations in class-imbalanced scenarios. Detailed comparative results for the ESCA, THCA, and UCEC datasets are provided in Tables 2, 3, and 4, respectively.

Ablation experiments

Layer attention mechanism

In PIGCN, the layer attention mechanism sets the initial weights to 1/2, 1/3, and 1/4, respectively. The motivation for this weight distribution is as follows: the first layer provides the most basic feature information, thus it is given the highest weight. The second layer further processes the features from the first layer, which reflects its importance in further abstracting features. The third layer has a weight of 1/4, which is relatively less significant.

This weight distribution allows the model to focus on learning basic features during the initial training phase and gradually integrate more complex feature relationships. To demonstrate the effectiveness of the layer attention mechanism, we conducted ablation experiments by removing the layer attention module. Specifically, we compared the potential methylation prediction scores of PIGCN and GCN for esophageal carcinoma and selected the top 10 methylations with the most significant differential methylation scores, as shown in Fig. 4. Additionally, we validated these methylations through literature review. The results showed that among these 10 methylations, TRIB3 was validated by existing literature to be significantly associated with the development of esophageal carcinoma [73]. MELK has been predicted by literature to potentially be associated with the development of the esophageal carcinoma [74]. Additionally, CADM2 and RLF have been confirmed to play roles in the progression of other cancers [75, 76]. This indicates that PIGCN's ability to predict potential associations is significantly higher than that of GCN, further proving the effectiveness of the layer attention mechanism.

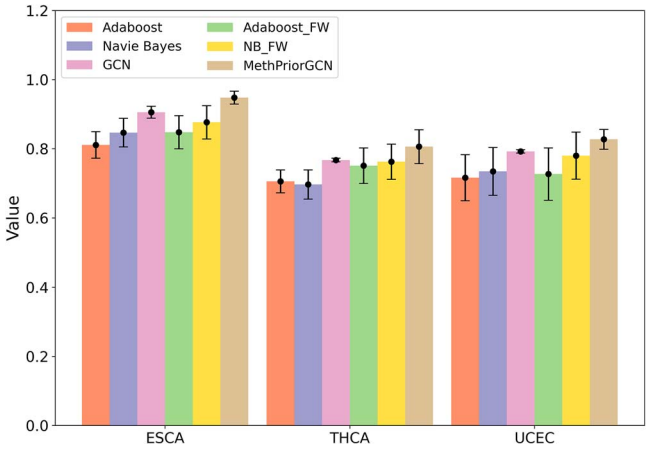

Figure 5. Impact of feature weighting on classification accuracy.

Feature weighting

Feature weighting is the most crucial step in MethPriorGCN, used to connect known association prior big data with methylation expression sequencing data. To validate the effectiveness of feature weighting, we removed the feature weighting module from MethPriorGCN and used only methylation expression data for disease classification tasks, comparing the prediction results. Figure 5 shows the comparative performance of MethPriorGCN and ordinary GCN on three parameters across three datasets.

Feature weighting integrates prior knowledge with expression data, thereby enhancing subtype-specific classification. For

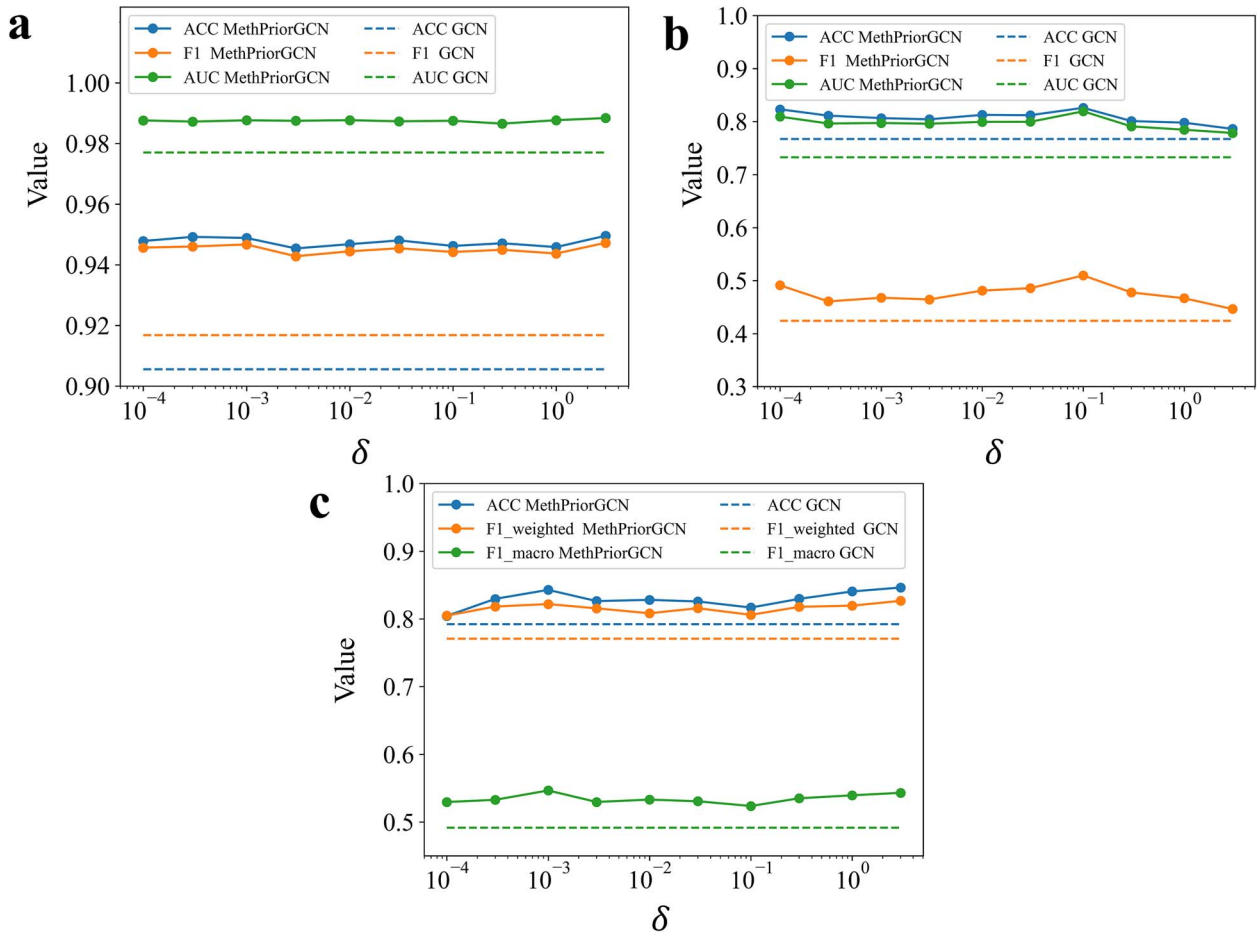

Figure 6. Performance of MethPriorGCN under different values of hyperparameter δ. The dashed lines represent the results of GCN (classification model without feature weighting method).

example, in UCEC, the observed accuracy decline after removing the weighting module underscores its critical role in prioritizing high-impact markers, which are overlooked by standard GCN.

MethPriorGCN with different Hyperparameters δ

In the Feature Weighting stage of FWGCN, the L1 regularization coefficient plays a crucial role. Applying regularization to the weighted methylation expression data enhances the model's effectiveness and clarity. This coefficient controls the sparsity of the expression intensity of digital markers and influences the model's adaptability to new datasets. While a lower L1 regularization coefficient might retain unnecessary markers, increasing the model's complexity, a higher coefficient could overlook important markers with significant association scores, reducing the model's effectiveness. Therefore, precise calibration of the L1 regularization coefficient is vital for optimizing MethPriorGCN's performance.

To evaluate the impact of the L1 regularization coefficient on MethPriorGCN's performance in binary and multiclass classification tasks, we conducted systematic trials across the ESCA, THCA, and UCEC datasets. As demonstrated in Fig. 6, the L1 coefficient critically modulates classification performance through its regulation of feature sparsity. In the ESCA dataset, the optimal coefficient ($\lambda = 0.01$) achieved a balance between feature relevance and noise reduction by retaining high-impact methylation markers while filtering non-informative ones, resulting in the highest AUC of 0.988. In contrast, THCA and UCEC exhibited greater sensitivity to coefficient variations, with performance fluctuations exceeding those observed in ESCA. This disparity reflects the necessity to dynamically prioritize domain-specific features in heterogeneous datasets, where feature weighting compensates for inherent data quality limitations.

Despite variations in marker expression intensity, MethPriorGCN maintained consistent metric outputs across datasets. For instance, UCEC showed robust ACC stability alongside F1_macro fluctuations, underscoring the model's capacity to reliably capture subtype-defining epigenetic alterations. These results collectively demonstrate that feature weighting predominantly enhances classification in data-constrained scenarios like THCA and UCEC, while exerting minimal impact on inherently well-separated datasets such as ESCA.

Discussion

DNA methylation plays a crucial role in the development of human diseases. However, due to individual heterogeneity, the same disease may manifest differently across individuals. This underscores the importance of precision medicine in contemporary clinical practice. Simultaneously, individual heterogeneity implies limited generalizability of universal predictive models. To address this, we developed MethPriorGCN, a GCN-based framework that infers potential prior knowledge in known meth-disease associations and leverages it to screen epigenetic biomarkers for disease subtype classification. This approach aims to

mitigate challenges posed by individual heterogeneity in clinical diagnostics.

MethPriorGCN employs GCN as its foundational architecture. This design enables effective integration of established meth-disease associations while accounting for coordinated regulation patterns among methylation sites in expression profiles. During the PIGCN phase, our methodological rationale stems from two key observations: (i) While known meth-disease association databases contain substantial entries, the average number of methylation sites per disease remains sparse; (ii) Literature-mined associations may introduce confirmation bias during biomarker identification. We therefore systematically inferred latent associations and computed quantitative scores for all meth-disease pairs. Experimental validation demonstrated that incorporating these inferred associations significantly improved disease subtype classification accuracy and identified more clinically actionable biomarkers compared to conventional approaches.

While MethPriorGCN demonstrates significant advancements, three limitations warrant discussion. First, the inherent heterogeneity and sparse sampling of meth-disease associations coupled with expression data variability may compromise association score reliability, impacting both subtype classification validity and biomarker discovery reproducibility. Second, the optimal integration strategy for prior knowledge in GCNs requires further investigation, particularly regarding mathematically rigorous weighting schemes to improve model interpretability. Future work should incorporate multimodal omics integration (e.g., genomics, proteomics, metabolomics) and expand disease cohorts to better capture pathophysiological complexity. These advancements could yield clinically translatable methylation biomarkers through enhanced biological plausibility and technical robustness.

Key Points

- In the PIGCN, a GCN with layer attention mechanism is utilized to learn prior knowledge for meth-mediated disease subtype classification from a heterogeneous network of known meth-disease associations.
- In the FWGCN, a GCN with a feature weighting strategy for disease subtype classification (FWGCN) is employed, effectively inform disease subtype classification using the learned prior knowledge.
- By combining PIGCN and FWGCN, the digital methylation markers are identified, and accuracy of classifying three disease subtypes is successfully improved. The effectiveness of the feature weighting strategy is validated through ablation experiments.

Supplementary data

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

Acknowledgments

Yun Liu, Xin Zhang, and Zhuoying Xie led the project, supervised the writing and did the revision of manuscript. Jie Ni did the experiments and wrote the manuscript. Bin Li, Shumei Miao, Xinting Zhang, Donghui Yan, Shenqi Jing, and Shan Lu did the subsequent revisions. All authors participated in the revisions and approved the final version of the manuscript.

Conflict of interest: The authors declare no conflict(s) of interest.

Funding

This work was supported by Ministry of Science and Technology of the People's Republic of China National Key Research & Development Program (grant number 2023YFC3605800), the Fundamental Research Funds for the Central Universities of China (grant number No. 2242023 K5005), Natural Science Foundation of Jiangsu Province - Major Project (BK20222008), the Nanjing Science and Technology Bureau Project (grant number 202205053), the Social Development Plan of the Provincial Department of Science and Technology in Jiangsu Province (grant number BE2023781), Suzhou Science and Technology Project (SJC2023005), and SEU Innovation Capability Enhancement Plan for Doctoral Students (CXJH_SEU 25145).

Data availability

Omics data were obtained from The Cancer Genome Atlas Program (TCGA). The known meth-disease association, preprocessed data, and all source code are openly available at (https://github.com/Jie-Ni/MethPriorGCN).

References

1. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012;**13**:484–92. https://doi.org/10.1038/nrg3230
2. Mattei AL, Bailly N, Meissner A. DNA methylation: a historical perspective. *Trends Genet* 2022;**38**:676–707. https://doi.org/10.1016/j.tig.2022.03.010
3. Papanicolau-Sengos A, Aldape K. DNA methylation profiling: an emerging paradigm for cancer diagnosis. *Annual Rev Pathol Mech Dis* 2022;**17**:295–321. https://doi.org/10.1146/annurev-pathol-042220-022304
4. Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. *Trends Genet* 2021;**37**:1012–27. https://doi.org/10.1016/j.tig.2021.05.002
5. Frenster John H, Herstein PR. Gene De-repression. *New Engl J Med* 1973;**288**:1224–9. https://doi.org/10.1056/NEJM197306072882310
6. Khawar MB, Afzal A, Ali MM. *et al.* A novel approach for tumor-released methylated DNA detection using molecularly imprinted polymers. *Adv Funct Mater* 2024;**34**:2405786. https://doi.org/10.1002/adfm.202405786
7. Feinberg AP, Levchenko A. Epigenetics as a mediator of plasticity in cancer. *Science* 2023;**379**:eaaw3835. https://doi.org/10.1126/science.aaw3835
8. Yamada Y, Venkadakrishnan VB, Mizuno K. *et al.* Targeting DNA methylation and B7-H3 in RB1-deficient and neuroendocrine prostate cancer. *Sci Transl Med* 2023;**15**:eadf6732. https://doi.org/10.1126/scitranslmed.adf6732
9. Xia B, Lu Y-L, Peng J. *et al.* Galactin-8 DNA methylation mediates macrophage autophagy through the MAPK/mTOR pathway to alleviate atherosclerosis. *Sci Rep* 2025;**15**:603. https://doi.org/10.1038/s41598-024-85036-1
10. Conway ME, McDaniel JM, Graham JM. *et al.* STAT3 and GR cooperate to drive gene expression and growth of basal-like

triple-negative breast cancer. *Cancer Res* 2020;**80**:4355–70. https://doi.org/10.1158/0008-5472.CAN-20-1379
11. Wong KK. DNMT1: a key drug target in triple-negative breast cancer. *Semin Cancer Biol* 2021;**72**:198–213. https://doi.org/10.1016/j.semcancer.2020.05.010
12. Hop PJ, Zwamborn RAJ, Hannon E. *et al.* Genome-wide study of DNA methylation shows alterations in metabolic, inflammatory, and cholesterol pathways in ALS. *Sci Transl Med* 2022;**14**:eabj0264. https://doi.org/10.1126/scitranslmed.abj0264
13. Shi Y, Zhang H, Huang S. *et al.* Epigenetic regulation in cardiovascular disease: mechanisms and advances in clinical trials. *Signal Transduct Target Ther* 2022;**7**:200. https://doi.org/10.1038/s41392-022-01055-2
14. Wu Y-L, Lin Z-J, Li C-C. *et al.* Epigenetic regulation in metabolic diseases: mechanisms and advances in clinical study. *Signal Transduct Target Ther* 2023;**8**:98. https://doi.org/10.1038/s41392-023-01333-7
15. Chen M, Fu R, Chen Y. *et al.* High-resolution, noninvasive single-cell lineage tracing in mice and humans based on DNA methylation epimutations. *Nature Methods*. 2025;**22:**488–98. https://doi.org/10.1038/s41592-024-02568-0
16. Lv H, Xie N, Li M. *et al.* DNA-based programmable gate arrays for general-purpose DNA computing. *Nature* 2023;**622**:292–300. https://doi.org/10.1038/s41586-023-06484-9
17. Liang S-Z, Wang L, You Z-H. *et al.* Predicting circRNA–disease associations through multisource domain-aware Embeddings and feature projection networks. *J Chem Inf Model* 2025;**65**:1666–76. https://doi.org/10.1021/acs.jcim.4c02250
18. Zhou J, Weinberger DR, Han S. Deep learning predicts DNA methylation regulatory variants in specific brain cell types and enhances fine mapping for brain disorders, science. *Advances* 2025;**11**:eadn1870. https://doi.org/10.1126/sciadv.adn1870
19. Li Y, Wu X, Fang D. *et al.* Informing immunotherapy with multi-omics driven machine learning. *NPJ Digital Med* 2024;**7**:67. https://doi.org/10.1038/s41746-024-01043-6
20. Wang G, Yao H, Gong Y. *et al.* Metabolic detection and systems analyses of pancreatic ductal adenocarcinoma through machine learning, lipidomics, and multi-omics, science. *Advances* 2021;**7**:eabh2724. https://doi.org/10.1126/sciadv.abh2724
21. Xie JC, Rao JH, Xie JJ. *et al.* Predicting disease-gene associations through self-supervised mutual infomax graph convolution network. *Comput Biol Med* 2024;**170**:108048. https://doi.org/10.1016/j.compbiomed.2024.108048
22. Huang Y-A, Li Y-C, You Z-H. *et al.* Consensus representation of multiple cell–cell graphs from gene signaling pathways for cell type annotation. *BMC Biol* 2025;**23**:23. https://doi.org/10.1186/s12915-025-02128-8
23. Wang W, Chen H. Predicting miRNA-disease associations based on lncRNA-miRNA interactions and graph convolution networks. *Brief Bioinform* 2023;**24**:bbac495. https://doi.org/10.1093/bib/bbac495
24. Li Z, Zhong T, Huang D. *et al.* Hierarchical graph attention network for miRNA-disease association prediction. *Mol Ther* 2022;**30**:1775–86. https://doi.org/10.1016/j.ymthe.2022.01.041
25. He J, Li M, Qiu J. *et al.* HOPEXGB: a consensual model for predicting miRNA/lncRNA-disease associations using a heterogeneous disease-miRNA-lncRNA information network. *J Chem Inf Model* 2024;**64**:2863–77. https://doi.org/10.1021/acs.jcim.3c00856
26. Bai T, Yan K, Liu B. DAmiRLocGNet: miRNA subcellular localization prediction by combining miRNA-disease associations and graph convolutional networks. *Brief Bioinform* 2023;**24**:bbad212. https://doi.org/10.1093/bib/bbad212
27. Peng W, Che Z, Dai W. *et al.* Predicting miRNA-disease associations from miRNA-gene-disease heterogeneous network with multi-relational graph convolutional network model. *IEEE/ACM Trans Comput Biol Bioinform* 2023;**20**:3363–75. https://doi.org/10.1109/TCBB.2022.3187739
28. Lu Z, Zhong H, Tang L. *et al.* Predicting lncRNA-disease associations based on heterogeneous graph convolutional generative adversarial network. *PLoS Comput Biol* 2023;**19**:e1011634. https://doi.org/10.1371/journal.pcbi.1011634
29. Wang B, Zhang C, Du XX. *et al.* lncRNA-disease association prediction based on the weight matrix and projection score. *PloS One* 2023;**18**:e0278817. https://doi.org/10.1371/journal.pone.0278817
30. Lan W, Lai D, Chen Q. *et al.* LDICDL: LncRNA-disease association identification based on collaborative deep learning. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:1715–23. https://doi.org/10.1109/TCBB.2020.3034910
31. Xuan P, Zhan L, Cui H. *et al.* Graph triple-attention network for disease-related LncRNA prediction. *IEEE J Biomed Health Inform* 2022;**26**:2839–49. https://doi.org/10.1109/JBHI.2021.3130110
32. Zhang GZ, Gao YL. BRWMC: predicting lncRNA-disease associations based on bi-random walk and matrix completion on disease and lncRNA networks. *Comput Biol Chem* 2023;**103**:107833. https://doi.org/10.1016/j.compbiolchem.2023.107833
33. Wang L, Wong L, You ZH. *et al.* NSECDA: natural semantic enhancement for CircRNA-disease association prediction. *IEEE J Biomed Health Inform* 2022;**26**:5075–84. https://doi.org/10.1109/JBHI.2022.3199462
34. Qiao LJ, Gao Z, Ji CM. *et al.* Potential circRNA-disease association prediction using DeepWalk and nonnegative matrix factorization. *IEEE/ACM Trans Comput Biol Bioinform* 2023;**20**:3154–62. https://doi.org/10.1109/TCBB.2023.3264466
35. Lan W, Dong Y, Chen Q. *et al.* KGANCDA: predicting circRNA-disease associations based on knowledge graph attention network. *Brief Bioinform* 2022;**23**:bbab494. https://doi.org/10.1093/bib/bbab494
36. Zhang HY, Wang L, You ZH. *et al.* iGRLCDA: identifying circRNA-disease association based on graph representation learning. *Brief Bioinform* 2022;**23**:bbac083. https://doi.org/10.1093/bib/bbac083
37. Wang L, You ZH, Huang DS. *et al.* MGRCDA: Metagraph recommendation method for predicting CircRNA-disease association. *IEEE Trans Cybern* 2023;**53**:67–75. https://doi.org/10.1109/TCYB.2021.3090756
38. Uhlen M, Zhang C, Lee S. *et al.* A pathology atlas of the human cancer transcriptome. *Science* 2017;**357**:eaan2507. https://doi.org/10.1126/science.aan2507
39. Horak P, Heining C, Kreutzfeldt S. *et al.* Comprehensive genomic and transcriptomic analysis for guiding therapeutic decisions in patients with rare cancers. *Cancer Discov* 2021;**11**:2780–95. https://doi.org/10.1158/2159-8290.CD-21-0126
40. Wei M, Wang L, Li Y. *et al.* BioKG-CMI: a multi-source feature fusion model based on biological knowledge graph for predicting circRNA-miRNA interactions. *Sci China Inform Sci* 2024;**67**:189104. https://doi.org/10.1007/s11432-024-4098-3
41. Hardikar S, Ren R, Ying Z. *et al.* The ICF syndrome protein CDCA7 harbors a unique DNA binding domain that recognizes a CpG dyad in the context of a non-B DNA, science. *Advances* 2024;**10**:eadr0036. https://doi.org/10.1126/sciadv.adr0036
42. Li LY, Sun YL. Circulating tumor DNA methylation detection as biomarker and its application in tumor liquid biopsy: advances and challenges. *Medcomm* 2024;**5**:5. https://doi.org/10.1002/mco2.766
43. Rendek T, Pos O, Duranova T. *et al.* Current challenges of methylation-based liquid biopsies in cancer diagnostics. *Cancers (Basel)* 2024;**16**:2001. https://doi.org/10.3390/cancers16112001

44. Jarolim R, Thalmann JK, Veronig AM. *et al.* Probing the solar coronal magnetic field with physics-informed neural networks. *Nature Astronomy* 2023;**7**:1171–9. https://doi.org/10.1038/s41550-023-02030-9
45. Ji W, Chang J, Xu H-X. *et al.* Recent advances in metasurface design and quantum optics applications with machine learning, physics-informed neural networks, and topology optimization methods. *Light: Sci Appl* 2023;**12**:169. https://doi.org/10.1038/s41377-023-01218-y
46. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* 2019;**378**:686–7. https://doi.org/10.1016/j.jcp.2018.10.045
47. Vemuri SK, Denzler J. Gradient statistics-based multi-objective optimization in physics-informed neural networks. *Sensors* 2023;**23**:23. https://doi.org/10.3390/s23218665
48. Yuan L, Ni YQ, Deng XY. *et al.* A-PINN: auxiliary physics informed neural networks for forward and inverse problems of nonlinear integro-differential equations. *J Comput Phys* 2022;**462**:111260. https://doi.org/10.1016/j.jcp.2022.111260
49. Yuan B, Wang H, Heitor A. *et al.* f-PICNN: a physics-informed convolutional neural network for partial differential equations with space-time domain. *J Comput Phys* 2024;**515**:113284. https://doi.org/10.1016/j.jcp.2024.113284
50. Zhang Z. A physics-informed deep convolutional neural network for simulating and predicting transient Darcy flows in heterogeneous reservoirs without labeled data. *J Petrol Sci Eng* 2022;**211**:110179. https://doi.org/10.1016/j.petrol.2022.110179
51. Jahani-nasab M, Bijarchi MA. Enhancing convergence speed with feature enforcing physics-informed neural networks using boundary conditions as prior knowledge. *Sci Rep* 2024;**14**:23836. https://doi.org/10.1038/s41598-024-74711-y
52. Meinders MBJ, Yang J, Linden. *et al.* Application of physics encoded neural networks to improve predictability of properties of complex multi-scale systems. *Sci Rep* 2024;**14**:15015. https://doi.org/10.1038/s41598-024-65304-w
53. Tang Y, Fan J, Li X. *et al.* Physics-informed recurrent neural network for time dynamics in optical resonances. *Nat Comput Sci* 2022;**2**:169–78. https://doi.org/10.1038/s43588-022-00215-2
54. Zhang X, Zhang X, Ren J. *et al.* Precise modelling of mitochondrial diseases using optimized mitoBEs. *Nature* 2025;**639**:1–11. https://doi.org/10.1038/s41586-024-08469-8
55. Yu MK, Ma JZ, Fisher J. *et al.* Visible machine learning for biomedicine. *Cell* 2018;**173**:1562–5. https://doi.org/10.1016/j.cell.2018.05.056
56. Corso G, Stark H, Jegelka S. *et al.* Graph neural networks. *Nat Rev Methods Primers* 2024;**4**:17. https://doi.org/10.1038/s43586-024-00294-7
57. Li R, Yuan X, Radfar M. *et al.* Graph signal processing, graph neural network and graph learning on biological data: a systematic review, Ieee reviews in biomedical. *Engineering* 2023;**16**:109–35. https://doi.org/10.1109/RBME.2021.3122522
58. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet* 2018;**19**:129–47. https://doi.org/10.1038/nrg.2017.86
59. Xing J, Zhai R, Wang C. *et al.* DiseaseMeth version 3.0: a major expansion and update of the human disease methylation database. *Nucleic Acids Res* 2022;**50**:D1208–d1215. https://doi.org/10.1093/nar/gkab1088
60. Lu Y, Peng R, Dong L. *et al.* Multiomics dynamic learning enables personalized diagnosis and prognosis for pancancer and cancer subtypes. *Brief Bioinform* 2023;**24**:24. https://doi.org/10.1093/bib/bbad378
61. Chen X, Huang YA, You ZH. *et al.* A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 2017;**33**:733–9. https://doi.org/10.1093/bioinformatics/btw715
62. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 2011;**27**:3036–43. https://doi.org/10.1093/bioinformatics/btr500
63. Chen X, Yan GY. Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 2013;**29**:2617–24. https://doi.org/10.1093/bioinformatics/btt426
64. Mnih V, Heess N, Graves A. *et al.* Recurrent models of visual attention. *Adv Neural Inform Processing Syst* 2014;**3**:2204–12. https://doi.org/10.48550/arXiv.1406.6247
65. Yu Z, Huang F, Zhao X. *et al.* Predicting drug-disease associations through layer attention graph convolutional network. *Brief Bioinform* 2021;**22**:22. https://doi.org/10.1093/bib/bbaa243
66. Han H, Zhu R, Liu JX. *et al.* Predicting miRNA-disease associations via layer attention graph convolutional network model. *BMC Med Inform Decis Mak* 2022;**22**:69. https://doi.org/10.1186/s12911-022-01807-8
67. Wang J, Li J, Yue K. *et al.* NMCMDA: neural multicategory MiRNA-disease association prediction. *Brief Bioinform* 2021;**22**:bbab074. https://doi.org/10.1093/bib/bbab074
68. Peng W, He Z, Dai W. *et al.* MHCLMDA: multihypergraph contrastive learning for miRNA-disease association prediction. *Brief Bioinform* 2023;**25**:bbad524. https://doi.org/10.1093/bib/bbad524
69. Li Y, Guo Z, Wang K. *et al.* End-to-end interpretable disease-gene association prediction. *Brief Bioinform* 2023;**24**:24. https://doi.org/10.1093/bib/bbad118
70. Chen M, Deng Y, Li Z. *et al.* SCPLPA: an miRNA-disease association prediction model based on spatial consistency projection and label propagation algorithm. *J Cell Mol Med* 2024;**28**:e18345. https://doi.org/10.1111/jcmm.18345
71. Chen Z, Che D, Gu X. *et al.* Upregulation of PEDF predicts a poor prognosis and promotes esophageal squamous cell carcinoma progression by modulating the MAPK/ERK signaling pathway. *Front Oncol* 2021;**11**:625612. https://doi.org/10.3389/fonc.2021.625612
72. Xu Z, Liao H, Huang L. *et al.* IBPGNET: lung adenocarcinoma recurrence prediction based on neural network interpretability. *Brief Bioinform* 2024;**25**:bbae080. https://doi.org/10.1093/bib/bbae080
73. Liu S, Ni C, Li Y. *et al.* The involvement of TRIB3 and FABP1 and their potential functions in the dynamic process of gastric cancer. *Front Mol Biosci* 2021;**8**:790433. https://doi.org/10.3389/fmolb.2021.790433
74. Xu Y, Wang Z, Pei B. *et al.* DNA methylation markers in esophageal cancer. *Front Genet* 2024;**15**:1354195. https://doi.org/10.3389/fgene.2024.1354195
75. Li D, Zhang Y, Zhang H. *et al.* CADM2, as a new target of miR-10b, promotes tumor metastasis through FAK/AKT pathway in hepatocellular carcinoma. *J Exp Clin Cancer Res* 2018;**37**:46. https://doi.org/10.1186/s13046-018-0699-1
76. Ciampricotti M, Karakousi T, Richards AL. *et al.* Rlf-Mycl gene fusion drives tumorigenesis and metastasis in a mouse model of small cell lung cancer. *Cancer Discov* 2021;**11**:3214–29. https://doi.org/10.1158/2159-8290.CD-21-0441

Briefings in Bioinformatics, 2025, **26(2)**, bbaf131
https://doi.org/10.1093/bib/bbaf131
Problem Solving Protocol