frontiers | Research Topics

# Eighth CCF Bioinformatics Conference: CBC 2023

**Edited by**
Fa Zhang, Wang Guohua, Xuefeng Cui and Min Li

**Coordinated by**
Cui Jinghua

**Published in**
Frontiers in Genetics

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-5675-7
DOI 10.3389/978-2-8325-5675-7

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

# Eighth CCF Bioinformatics Conference: CBC 2023

**Topic editors**
Fa Zhang — Beijing Institute of Technology, China
Wang Guohua — Harbin Institute of Technology, China
Xuefeng Cui — Shandong University, China
Min Li — Central South University, China

**Topic coordinator**
Cui Jinghua — Shandong University, China

**Citation**
Zhang, F., Guohua, W., Cui, X., Li, M., Jinghua, C., eds. (2024). *Eighth CCF Bioinformatics Conference: CBC 2023*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-5675-7

# Table of contents

04 Classifying breast cancer using multi-view graph neural network based on multi-omics data
Yanjiao Ren, Yimeng Gao, Wei Du, Weibo Qiao, Wei Li, Qianqian Yang, Yanchun Liang and Gaoyang Li

16 DeepCBS: shedding light on the impact of mutations occurring at CTCF binding sites
Yiheng Wang, Xingli Guo, Zhixin Niu, Xiaotai Huang, Bingbo Wang and Lin Gao

24 PCGIMA: developing the web server for human position-defined CpG islands methylation analysis
Ming Xiao, Yi Xiao, Jun Yu and Le Zhang

33 Acupuncture and tuina knowledge graph with prompt learning
Xiaoran Li, Xiaosong Han, Siqing Wei, Yanchun Liang and Renchu Guan

46 CD-Loop: a chromatin loop detection method based on the diffusion model
Jiquan Shen, Yang Wang and Junwei Luo

59 LcDel: deletion variation detection based on clustering and long reads
Yanan Yu, Runtian Gao and Junwei Luo

73 Specific feature recognition on group specific networks (SFR-GSN): a biomarker identification model for cancer stages
Bolin Chen, Yuxin Wang, Jinlei Zhang, Yourui Han, Hamza Benhammouda, Jun Bian, Ruiming Kang and Xuequn Shang

85 A joint model for lesion segmentation and classification of MS and NMOSD
Lan Huang, Yangguang Shao, Hui Yang, Chunjie Guo, Yan Wang, Ziqi Zhao and Yingchun Gong

102 Identification of DNA motif pairs on paired sequences based on composite heterogeneous graph
Qiuqin Wu, Yang Li, Qi Wang, Xiaoyu Zhao, Duanchen Sun and Bingqiang Liu

115 GIHP: Graph convolutional neural network based interpretable pan-specific HLA-peptide binding affinity prediction
Lingtao Su, Yan Yan, Bo Ma, Shiwei Zhao and Zhenyu Cui

128 MGPPI: multiscale graph neural networks for explainable protein–protein interaction prediction
Shiwei Zhao, Zhenyu Cui, Gonglei Zhang, Yanlong Gong and Lingtao Su

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 20 February 2024
DOI 10.3389/fgene.2024.1363896

Check for updates

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Advait Balaji,
Occidental Petroleum Corporation,
United States
Juexin Wang,
Purdue University Indianapolis, United States

*CORRESPONDENCE
Yanchun Liang,
✉ ycliang@jlu.edu.cn
Gaoyang Li,
✉ lgyzngc@tongji.edu.cn

RECEIVED 31 December 2023
ACCEPTED 02 February 2024
PUBLISHED 20 February 2024

CITATION
Ren Y, Gao Y, Du W, Qiao W, Li W, Yang Q, Liang Y and Li G (2024), Classifying breast cancer using multi-view graph neural network based on multi-omics data.
*Front. Genet.* 15:1363896.
doi: 10.3389/fgene.2024.1363896

COPYRIGHT
© 2024 Ren, Gao, Du, Qiao, Li, Yang, Liang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Classifying breast cancer using multi-view graph neural network based on multi-omics data

Yanjiao Ren[1], Yimeng Gao[1], Wei Du[2], Weibo Qiao[2], Wei Li[1], Qianqian Yang[1], Yanchun Liang[2,3]* and Gaoyang Li[4]*

[1]College of Information Technology, Smart Agriculture Research Institute, Jilin Agricultural University, Changchun, Jilin, China, [2]College of Computer Science and Technology, Jilin University, Changchun, China, [3]School of Computer Science, Zhuhai College of Science and Technology, Zhuhai, China, [4]Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai, China

**Introduction:** As the evaluation indices, cancer grading and subtyping have diverse clinical, pathological, and molecular characteristics with prognostic and therapeutic implications. Although researchers have begun to study cancer differentiation and subtype prediction, most of relevant methods are based on traditional machine learning and rely on single omics data. It is necessary to explore a deep learning algorithm that integrates multi-omics data to achieve classification prediction of cancer differentiation and subtypes.

**Methods:** This paper proposes a multi-omics data fusion algorithm based on a multi-view graph neural network (MVGNN) for predicting cancer differentiation and subtype classification. The model framework consists of a graph convolutional network (GCN) module for learning features from different omics data and an attention module for integrating multi-omics data. Three different types of omics data are used. For each type of omics data, feature selection is performed using methods such as the chi-square test and minimum redundancy maximum relevance (mRMR). Weighted patient similarity networks are constructed based on the selected omics features, and GCN is trained using omics features and corresponding similarity networks. Finally, an attention module integrates different types of omics features and performs the final cancer classification prediction.

**Results:** To validate the cancer classification predictive performance of the MVGNN model, we conducted experimental comparisons with traditional machine learning models and currently popular methods based on integrating multi-omics data using 5-fold cross-validation. Additionally, we performed comparative experiments on cancer differentiation and its subtypes based on single omics data, two omics data, and three omics data.

**Discussion:** This paper proposed the MVGNN model and it performed well in cancer classification prediction based on multiple omics data.

KEYWORDS

**multi-view graph neural network, multi-omics data, attention mechanism, feature selection, cancer differentiation, cancer subtypes**

## 1 Introduction

Cancer is one of the leading causes of death in the world today. According to the global cancer statistics report in 2020, there were nearly 19.3 million new cases of cancer and 10 million cancer-related deaths worldwide (Bray et al., 2018). Due to factors such as globalization and economic growth, the number of new cancer cases is expected to continue to rise. Cancer is a disease characterized by the uncontrolled growth and spreading of specific cells in the body to other parts of the body. These cells can also transfer to distant body parts, forming new tumors through metastasis (Hanahan and Weinberg, 2011). Tumors can be classified into different grades, known as tumor grading, by examining tumor cells under a microscope. Tumor grading compares the degree of cellular and tissue morphological changes between cancer cells and normal cells, indicating the tumor's differentiation. Generally, based on the abnormality of tumor cells observed under a microscope, tumors are classified into grades 1, 2, or 3 (sometimes also 4), called G1, G2, G3, and G4, respectively (Sobin and Fleming, 1997). These represent well-differentiated, moderately differentiated, poorly differentiated, and undifferentiated tumors. Cancer is also a heterogeneous disease that encompasses various subtypes. The same type of cancer can be divided into subtypes based on different mechanisms of occurrence. Different subtypes of the same cancer reflect distinct molecular carcinogenesis processes and clinical outcomes. With the advent of precision medicine, cancer classification has gradually become one of the fundamental goals of cancer informatics. Heterogeneous cancer populations are grouped into clinically meaningful subtypes based on the similarity of molecular spectra.

Breast cancer is a most common cancer worldwide (Loibl et al., 2021). The number of breast cancer patients is increasing year by year, and the proportion of women under the age of 40 with breast cancer has reached 6.6% (Assi et al., 2013). Breast cancer incidence rates have risen in most of the past four decades; during the most recent data years (2010–2019), the rate increased by 0.5% annually (Giaquinto et al., 2022). Breast cancer, as a highly heterogeneous disease, is composed of different biological subtypes, which possess distinct clinical, pathological, and molecular characteristics, as well as prognostic and therapeutic significance (Reis-Filho and Pusztai, 2011).Therefore, studying breast cancer subtypes is of great significance for precision medicine and prognosis prediction (Waks and Winer, 2019).In the year 2000, Perou et al. first proposed the molecular subtyping of breast cancer. They concluded that breast cancer can be divided into four subtypes: Luminal A subtype, Basal-like subtype, HER2-enriched subtype, and Normal-like subtype (Perou et al., 2000). Sorlie et al. subdivided the luminal subtype into luminal A and B subtypes (Sorlie et al., 2003). Waks et al. categorized breast cancer into three major subtypes based on the presence or absence of molecular markers, including estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). These subtypes are ER+/PR+/HER2- (luminal A), HER2-positive, and triple-negative breast cancer (TNBC), where all three of these molecular markers are negative (Yersal and Barutca, 2014). The HER2-positive subtype can be further divided into ER+/PR+/HER2+ (luminal B) and ER-/PR-/HER2+. Tao et al. categorized breast cancer into five subtypes based on immunohistochemistry (IHC) markers, including ER, PR, and HER2 (Tao et al., 2019). These subtypes include luminal A, B, HER2-positive, TNBC, and unclassified.

With the advancement of sequencing technologies, various types of omics data in the biosphere, including transcriptomics data [RNA expression data (Wang et al., 2009; Ozsolak and Milos, 2011)], metabolomics (Shulaev, 2006) data, proteomics (Altelaar et al., 2013) data, methylation patterns (Laird, 2010) data, as well as genomics data [DNA sequence data (Metzker, 2010)], have experienced rapid growth and accumulation. Many researchers have developed corresponding tools to handle this large-scale omics data. Another issue gradually gaining attention from researchers is whether there is interaction between complex traits and omics data. Previous studies mainly focused on the relationship between individual omics data and biological processes. Due to the reliance on a single type of omics data in analyzing the causes of complex traits, there have been few research results in this area until now. Through many existing experimental studies, it is known that there is a specific connection between different omics data, and they can complement each other's missing information. This is crucial for researchers to discover the relationship between complex traits and different omics data (Reif et al., 2004; Sieberts and Schadt, 2007; Hamid et al., 2009; Hawkins et al., 2010; Holzinger and Ritchie, 2012). Integrating different types of omics data and designing reasonable and adequate multi-omics data integration methods to accurately predict cancer differentiation and subtype classification have become hot topics in cancer research.

Deep learning, as an emerging and efficient method in the field of machine learning, is more capable of capturing non-linear complex relationships in complex models. It has been widely used in the research of multi-omics data fusion methods (Cai et al., 2022). Mohammed et al. proposed a LASSO based 1D-CNN method and compared it with SVM, ANN, KNN, and bagging tree methods, the results indicating that the classification performance of the deep stacking method was superior to the traditional machine learning method (Mohammed et al., 2021). Li et al. proposed the MoGCN method by integrating multi-omics data based on a graph Convolutional network (GCN). Autoencoders and similarity network fusion methods are used to reduce and construct a patient similarity network (PSN) respectively to capture complex nonlinear relationships among multi-omics data (Li et al., 2022). Xing et al. Proposed the MLE-GAT method, namely multi-layer embedded graph attention method, uses WGCNA method to format each patient's omics data into a co-expression network and uses the full gradient map significance mechanism to identify disease-related genes (Xing et al., 2021). Blanco et al. points out the need to maintain a certain balance between biology and computer technology, and to integrate biological knowledge into modeling methods (Linares-Blanco et al., 2021). Leng et al. suggests that the best foundational model for predicting the fusion of multiple omics data is the GNN model (Leng et al., 2022).

This paper considers the relations between feature nodes in the aggregation of GCN model, which are constructed based on multiple sets of omics data to form a similarity network. The correlation between samples can be captured through this similarity network, effectively preserving the biological semantic and geometric structures of the data. While for the GAT model, the relations between nodes are learned through network training. However, especially when the sample size is small, the training effect may

TABLE 1 Statistics of breast cancer data.

| Data type | Number of samples | Number of features |
|---|---|---|
| mRNA | 606 | 13195 |
| DNA methylation | 606 | 14285 |
| CNV | 606 | 15186 |

TABLE 2 Statistical information of breast cancer data differentiation.

| Breast cancer differentiation | Number of samples |
|---|---|
| G1 | 75 |
| G2 | 286 |
| G3 | 245 |

not be satisfactory. Therefore, this paper adopts the GCN model instead of the GAT model in the design, and subsequent experiments have also validated this design.

# 2 Materials and methods

## 2.1 Data collection

The breast cancer data used in this study were obtained from The Cancer Genome Atlas (TCGA) database (Weinstein et al., 2013), which contains various cancer types and their corresponding omics data. A total of 606 breast cancer cases were carefully selected, which included gene expression data, DNA methylation data, copy number variation (CNV) data, differentiation annotation, and subtype annotation. The specific statistical information of the mRNA, DNA methylation, and CNV data for the collected breast cancer cases is shown in Table 1. Among the breast cancer cases with differentiation annotation, there were 245 samples labeled as low differentiation (G3), 286 samples labeled as medium differentiation (G2), and 75 samples labeled as high differentiation (G1). The detailed information is presented in Table 2.

In this article, Tao et al. classified breast cancer into four subtypes using immunohistochemistry (IHC) labeling: luminal A, luminal B, HER2-positive, and triple-negative breast cancer (TNBC). The luminal A subtype is the most common, accounting for 60% of all breast cancer subtypes (Malhotra et al., 2010). The majority of patients with the luminal B subtype are elderly. Approximately 25% of breast cancer patients are HER2-positive, which is associated with a poorer prognosis. Most patients with HER2-positive advanced breast cancer are likely to have lymph node metastasis in the axillary region. The TNBC subtype is characterized by the absence of estrogen receptor (ER), progesterone receptor (PR), and HER2 (Tao et al., 2019). Compared to other subtypes of breast cancer, TNBC tends to rapidly deteriorate and metastasize.

In the breast cancer cases with subtype annotation, there were a total of 398 cases. Out of these, 277 cases were annotated as Luminal A, 40 were annotated as Luminal B, 11 were annotated as HER2(+), and 70 were annotated as TNBC. Table 3 provides detailed information on these cases. The above three omics data and two annotation files are provided in the Supplementary Material.

## 2.2 Data preprocessing

Generally, deep learning models do not require separate feature selection, as they can achieve this through the neural network's weights. However, due to the "large p small n" dimensionality catastrophe problem in omics data, training the network weights of omics data using the deep learning model is not adequate. In deep neural networks, fewer features often mean better interpretability and higher training speed. In this study, the collected breast cancer case sample data underwent preprocessing operations using three feature selection algorithms: chi-square test, linear normalization, and minimum redundancy maximum relevance (mRMR) (Yiming, 1997; Peng et al., 2005; Forman, 2008). The specific data preprocessing workflow is shown in Figure 1.

This paper uses the chi-square test to select features for each omics type. The features are sorted based on their number in the hypothesis test using the samples corresponding to each classification task. Then, the top-k features are selected for each omics data. In this study, k is set to 5000. Normalization is performed using linear scaling, transforming the data values to fit within the range of [0,1]. The paper also employs the minimum Redundancy Maximum Relevance (mRMR) feature selection algorithm. The difference between each feature's maximum relevance value and the minimum redundancy value is used as the feature score. The features are then sorted in descending order based on their scores, and the top 500 features are selected for further filtering. These selected features are favorable for cancer differentiation and subtype prediction.

## 2.3 Graph construction

A graph is a complex data structure consisting of nodes and edges. Many scenes in real life shown in the form of graphs or networks. For example, resources and users in recommendation systems can be considered as nodes in a graph, and the relationships between users and items can be considered as edges. Complex terms like chemical molecules can also be abstracted as graphs (Zhou et al., 2020). Most deep learning algorithms use data such as speech, images, and text with tidy and regular data structures. However, conventional deep learning algorithms are difficult to handle for those irregular and complex network structures. The Graph Convolutional Network (GCN) (Kipf and Welling, 2016) model can process such graph structures.

In this paper, patient similarity networks are constructed by using cosine similarity for three kinds of omics data, namely mRNA, DNA methylation, and CNV data, respectively (Pai and Bader, 2018). The calculation formula for cosine similarity is as Eq. 1:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{1}$$

where, $A$ and $B$ are two known attribute vectors, $A_i$ and $B_i$ respectively represent the components of the vector sum.

Each patient sample is a node in the patient similarity network, and the goal of each GCN in the model is to learn features aggregation from the graph-structured data by leveraging the

TABLE 3 Classification of cancer subtypes.

| Breast cancer subtype | Number of samples | IHC markers |
|---|---|---|
| luminal A | 277 | ER/PR+, Her2− |
| luminal B | 40 | ER/PR+, Her2+ |
| HER2(+) | 11 | ER/PR−, Her2+ |
| TNBC | 70 | ER/PR−, Her2− |

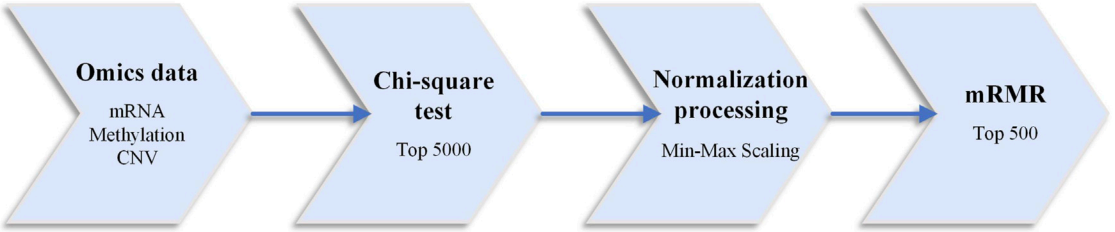


FIGURE 1
Data preprocessing flowchart.

features of each node and the relationships between nodes. Therefore, the input of the GCN module consists of two parts: the feature matrix and the graph structure description. The feature matrix is represented as $\mathbf{X} \in \mathbb{R}^{n\times d}$, where n is the number of nodes and d is the number of input features. The graph structure description is an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, constructed by computing the cosine similarity between node pairs. The computation equation is as Eq. 2:

$$A_{ij} = \begin{cases} s(\mathbf{x}_i, \mathbf{x}_j), & \text{if } i \neq j \text{ and } s(\mathbf{x}_i, \mathbf{x}_j) \geq \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

In the equation, $\mathrm{A}_{\mathrm{ij}}$ represents the adjacency relationship between node i and node j, $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors of node i and node j, and $s(\mathbf{x}_i, \mathbf{x}_j)$ is the cosine similarity between node i and node j. $\epsilon$ is a threshold determined by k, where k represents the average number of edges preserved for each node. The computation equation for k is as Eq. 3:

$$k = \sum\nolimits_{i,j} I\left(s(\mathbf{x}_i, \mathbf{x}_j) \geq \epsilon\right)\Big/ n \tag{3}$$

where $I(\cdot)$ represents an indicator function, and n is the number of nodes. With the similarity network, GCN can be trained using omics features and the corresponding similarity network to learn specific omics data.

## 2.4 Model design

The proposed model in this paper consists mainly of the Graph Convolutional Neural Network (GCN) module and an attention (Velikovi et al., 2017) module. The GCN module is designed for learning the feature aggregation of specific omics data, while the attention module is designed for the fusion of multi-omics features corresponding to different omics data obtained from the output of the GCN module. The attention module can assign different attention weight to each neighbor of a node, thus identifying more important neighbors for better classification of breast cancer differentiation and its subtypes.

This paper presents a detailed architecture of the model for predicting the differentiation degree and subtypes of breast cancer, as shown in Figure 2.

In this paper, the GCN is constructed by stacking multiple convolutional layers. Specifically, each layer is defined as Eq. 4:

$$\mathbf{H}^{(l+1)} = f\left(\mathbf{H}^{(l)}, \mathbf{A}\right) = \sigma\left(\mathrm{AH}^{(l)}\mathbf{W}^{(l)}\right) \tag{4}$$

where $l$ is the number of graph convolutional layers, $\mathbf{H}^{(l)}$ is the input of the $l$ th layer, $\mathbf{W}^{(l)}$ is the weight matrix of the $l$ th layer. $\sigma(\cdot)$ represents a non-linear activation function. $\mathbf{H}^{(l+1)}$ is the output of the $l$ th layer. When the number of graph convolutional layers is too large, the resulting node feature vectors will become overly smooth, meaning that the features of each node become very similar. This is mainly because each layer of the GCN integrates information from the node and its neighbors. As the layers deepen, each node incorporates information from more neighbors, including some unrelated nodes. This ultimately leads to similar feature vectors for different types of nodes.

This paper's model observed that when the number of graph convolutional layers exceeded three, there was no significant improvement in the experimental results. Instead, it increased the computational time and led to overfitting on some datasets. Therefore, the GCN module in this paper's model consists of three graph convolutional layers.

To effectively train GCN, this paper extends the approach of Kipf et al. (Kipf and Welling, 2016) by further modifying the adjacency matrix A as Eq. 5:

$$\tilde{\mathbf{A}} = \hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}} = \hat{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\hat{\mathbf{D}}^{-\frac{1}{2}} \tag{5}$$

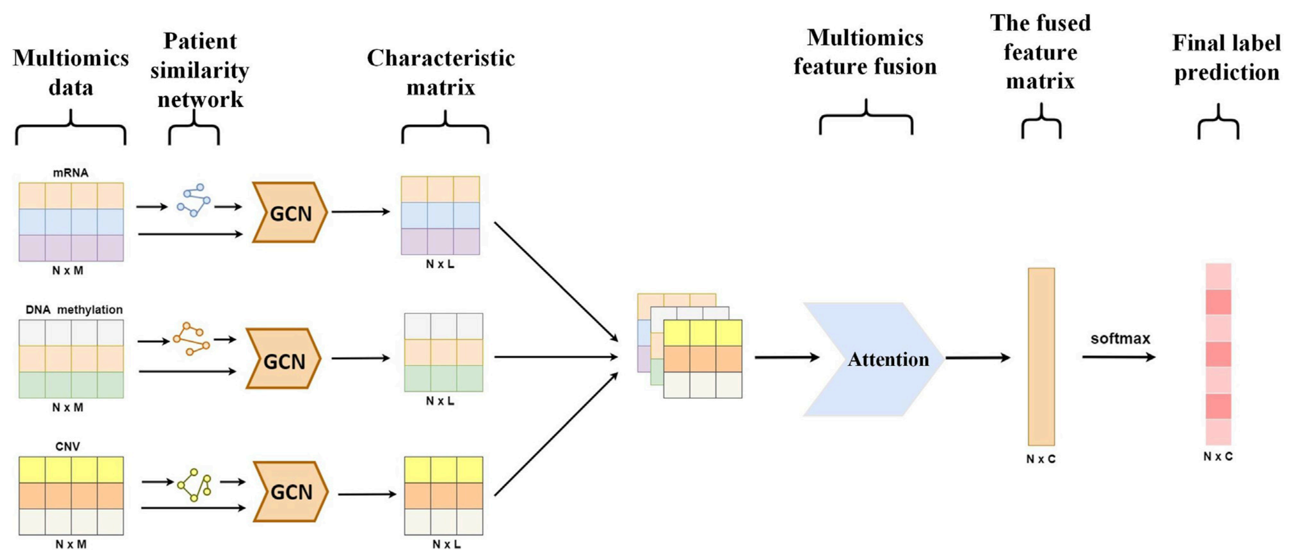


FIGURE 2
Prediction model of MVGNN.

where $\hat{\boldsymbol{D}}$ is the diagonal degree matrix of $\tilde{\mathbf{A}}$, and $\mathbf{I}$ is the identity matrix.

The attention model was introduced by Velikovi et al. (2017). The attention model incorporates a self-attention mechanism during the propagation process in the network. Unlike GCN, which treats all neighbors of a node equally, this attention model assigns different attention scores to all neighbors. A higher score for a neighbor indicates a higher importance level for that node. The attention network is implemented by stacking multiple graph attention layers. The input to a single graph attention layer is a set of node feature vectors as Eq. 6:

$$\mathbf{h} = \left\{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\right\}, \vec{h}_i \in \mathbb{R}^F \tag{6}$$

where N represents the number of nodes in the node-set, and F represents the corresponding eigenvector dimension.

The output of each layer is a new set of node feature vectors as Eq. 7:

$$\mathbf{h}' = \left\{\vec{h}'_1, \vec{h}'_2, \ldots, \vec{h}'_N\right\}, \vec{h}'_i \in \mathbb{R}^{F'} \tag{7}$$

where F′ represents the new node eigenvector dimension.

In order to obtain sufficient expressive power to transform input features into higher-level features, the graph attention layer first performs self-attention processing according to the set of node feature vectors of input as Eq. 8:

$$e_{ij} = a\left(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j\right) \tag{8}$$

The shared attention mechanism $a$ is a mapping of $\mathbb{R}^{F'} \times \mathbb{R}^{F'} \vec{x} \mathbb{R}$, and $W \in \mathbb{R}^{F' \times F}$ is a weight matrix that is shared by all $\vec{h}'_i$. $e_{ij}$ represents the importance of the features of node $j$ to node $i$.

In this study, the attention module is used to compute the attention coefficients for each omics feature matrix. The attention mechanism is then applied to aggregate different types of omics features, resulting in the final omics feature matrix. The fused feature matrix obtained from the attention module is further processed using SoftMax function for final label prediction.

# 3 Results

## 3.1 Performance metrics

Samples are generally divided into positive and negative classes for binary classification tasks. Therefore, the classifier has four classification results: TP, TN, FP, and FN. TP refers to correctly classifying positive samples as positive. TN refers to correctly classifying negative samples as negative. FP refers to incorrectly classifying negative samples as positive. FN refers to incorrectly classifying positive samples as negative. To evaluate the model's predictive performance, we mainly used three evaluation metrics: accuracy, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). The specific calculation formulas are as as Eqs 9–14:

$$\text{accuracy} = \frac{TP + TN}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

$$\text{sensitivity} = \text{recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{specificity} = \frac{TN}{TN + FP} \tag{11}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{F1\_score} = 2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

$$\text{AUC} = \frac{1}{m^+ m^-} \sum\nolimits_{x^+ \in D^+} \sum\nolimits_{x^- \in D^-} \left(W\left(f\left(x^+\right) > f\left(x^-\right)\right) + \frac{1}{2} W\left(f\left(x^+\right) = f\left(x^-\right)\right)\right) \tag{14}$$

In the paper, "accuracy" refers to the proportion of correctly predicted results among all samples. "F1" is the arithmetic average of precision and recall divided by the geometric mean. F1 has the worst effect when the value is 0 and the best effect when the value is 1. The receiver operating characteristic curve is known as ROC, and the area under the curve (AUC) represents the area under the ROC curve. AUC is calculated through the integral of the ROC curve, and a higher AUC indicates better classification results.

We adopt two evaluation indexes for multi-classification tasks, $F1\_macro$ and $F1\_weighted$ (Leng et al., 2022). Its calculation formula are as Eqs 15–17:

$$Precision\_macro = \frac{1}{n}\sum_{i=1}^{n} Precision_i \tag{15}$$

$$Recall\_macro = \frac{1}{n}\sum_{i=1}^{\pi} Recall_i \tag{16}$$

$$F1\_macro = 2 \cdot \frac{Precision\_macro \times Recall\_macro}{Precision\_macro + Recall\_macro} \tag{17}$$

$F1\_macro$ takes values between 0 and 1 and is unaffected by data imbalance. On the other hand, $F1\_weighted$ is the weighted average of F1_score for each category, where the weight is the proportion of each category in the accurate predictions. The difference between $F1\_weighted$ and $F1\_macro$ is that $F1\_macro$ assigns the same weight to each category, while $F1\_weighted$ assigns different weights based on the proportion of each category.

The model proposed in this paper and the comparison model are specifically executed on the workstation based on Ubuntu 18.04.5 LTS system and Pytorch v1.7.0. The working environment of the workstation is as follows: CPU is AMD Ryzen 7 3700X 8-Core, 16-Thread,Memory is 64G, GPU is GeForce GTX 1080 Ti (11G).

## 3.2 Implementation details

In deep learning, networks with many parameters are very powerful (Srivastava et al., 2014). However, dealing with the overfitting problem is a key issue. This paper adopts two approaches to address the overfitting issue. The first approach is to add dropout layers to the model. It randomly drops elements in the neural network during training, preventing overfitting caused by excessive training. Each sub-network channel consists of three sequential graph convolution layers and two dropout layers are used in our model and then weighted each channel using the attention mechanism. The second approach is to employ early stopping during the training process of the network model. Specifically, if the loss function of the validation data does not show a significant decrease in the first 100 epochs of training, the model's training is paused (Prechelt et al., 2012).

This paper computed the cross-entropy between the actual distribution and the predicted distribution of breast cancer differentiation and its subtypes (Tabor and Spurek, 2014). The loss is calculated by minimizing the cross-entropy. The loss function used in this paper's model is shown in Eq. 18:

$$L = -\sum_{l \in Y_L} Y^l \ln\left(C \cdot Z^l\right) \tag{18}$$

where L is the loss function, $Y_L$ is the set of node indexes with labels, $Y^l$ is the label of the label node, that is, the type of breast cancer differentiation and its subtypes, C is the parameter of the classifier, and $Z^l$ is the final node embedding of the label node. This paper optimizes the entire model through end-to-end backpropagation.

TABLE 4 The prediction results of classification in any two degrees of differentiation across different models.

| Method | ACC | AUC | F1 |
|---|---|---|---|
| SVM | 0.658 | 0.645 | 0.623 |
| RF | 0.669 | 0.649 | 0.624 |
| NN | 0.701 | 0.674 | 0.672 |
| NN_NN | 0.725 | 0.708 | 0.760 |
| NN_VCDN | 0.720 | 0.703 | 0.752 |
| GCN | 0.741 | 0.704 | 0.758 |
| GAT | 0.749 | 0.723 | 0.743 |
| MOGONET | 0.744 | 0.731 | 0.772 |
| MVGNN | 0.778 | 0.745 | 0.809 |

## 3.3 The performance of binary classification

### 3.3.1 Analysis of experimental results of binary classification in differentiation degree

In order to comprehensively evaluate the performance of our MVGNN model compared to traditional machine learning methods and recent supervised multi-omics data integration methods, this paper employs 5-fold cross-validation for different models. The average accuracy, average AUC value, and average F1 value obtained on the test dataset are used as evaluation metrics. These models include Support Vector Machine (SVM), Random Forest (RF), Neural Network (NN), GCN, GAT, and Multi-Omics Graph Convolutional Networks (MOGONET). MOGONET is the latest method for multi-omics data integration published by Wang et al. (2021). The View Correlation Discovery Network (VCDN) are used to explores cross-omics correlations in the feature space, enabling effective multi-omics integration. Three pairs of breast cancer differentiation classifications are considered: well-differentiated vs. moderately-differentiated (G1 vs. G2), well-differentiated vs. poorly-differentiated (G1 vs. G3), and moderately-differentiated vs. poorly-differentiated (G2 vs. G3). The same dataset split is used, and the average accuracy, average AUC value, and average F1 value based on 5-fold cross-validation are used as evaluation metrics. The experimental results of all models in predicting any two types of breast cancer differentiation are shown in Table 4.

In the experimental process, SVM, RF, NN, GCN, and GAT were trained using preprocessed multi-omics data directly concatenated as input. All methods were trained using the same preprocessed data. According to Table 4, the proposed MVGNN model for integrating multi-omics data achieved the highest accuracy, AUC value, and F1 value compared to traditional machine learning methods, graph convolutional network models, and the latest methods for integrating multi-omics data in classifying any two types of breast cancer differentiation. The values are: accuracy—0.778, AUC—0.745, F1—0.809. It can be

TABLE 5 Prediction results of each model for any two subtypes of breast cancer.

| Method | ACC | AUC | F1 |
|---|---|---|---|
| SVM | 0.7853 | 0.7725 | 0.5005 |
| RF | 0.8085 | 0.7917 | 0.5092 |
| NN | 0.8310 | 0.8103 | 0.5355 |
| NN_NN | 0.8505 | 0.8433 | 0.5927 |
| NN_VCDN | 0.8417 | 0.8473 | 0.6002 |
| GCN | 0.8627 | 0.8457 | 0.6310 |
| GAT | 0.8812 | 0.8702 | 0.6405 |
| MOGONET | 0.8915 | 0.9160 | 0.6632 |
| MVGNN | 0.9180 | 0.9530 | 0.7155 |

concluded that the proposed model in this study outperforms traditional machine learning models and the latest methods for integrating multi-omics data in classifying any two types of breast cancer differentiation.

### 3.3.2 Analysis of experimental results of binary classification on subtypes

This article adopts a five-fold cross-validation method to train all models, and all methods use the same training set, validation set, and test set. The evaluation metrics are average accuracy (ACC), average area under the curve (AUC), and average F1 score. The classification results of any two subtypes of breast cancer include (1) luminal A vs. luminal B, (2) luminal A vs. HER2(+), (3) luminal A vs. TNBC, (4) luminal B vs. HER2(+), (5) luminal B vs. TNBC, and (6) HER2(+) vs. TNBC. The experimental results of predicting any two subtypes of breast cancer by each model are shown in Table 5.

Based on the data in Table 5, this paper's model achieved the highest accuracy, AUC value, and F1 score compared to traditional machine learning methods, graph convolutional network models, and the latest integrated multi-omics data methods for any two classification results of breast cancer subtypes. The values are as follows: accuracy - 0.9180, AUC - 0.9530, and F1 score - 0.7155. It can be concluded that this paper's model outperforms traditional machine learning methods and the latest multi-omics data integration methods in the overall classification results of any two subtypes of breast cancer.

## 3.4 The performance of multi-classification

### 3.4.1 Analysis of the results of multi-classification experiments on differentiation degree

To better evaluate the performance of the MVGNN model, this paper uses the model to predict the differentiation degree and subtypes of breast cancer based on multi-classification. Specifically, based on the same data set partitioning, this paper uses the average accuracy, average F1_weighted value, and average F1_macro value calculated through 5-fold cross-validation as evaluation metrics. The multi-classification results of breast cancer differentiation degree are G1 vs. G2 vs. G3. The specific experimental results of the MVGNN model and other methods in the multi-classification of breast cancer differentiation degree are shown in Table 6.

TABLE 6 Experimental results of multiple classifications of different models in different degrees of differentiation.

| Method | ACC | F1_weighted | F1_macro |
|---|---|---|---|
| SVM | 0.529 | 0.5 | 0.429 |
| RF | 0.54 | 0.532 | 0.441 |
| NN | 0.56 | 0.547 | 0.464 |
| NN_NN | 0.574 | 0.549 | 0.518 |
| NN_VCDN | 0.572 | 0.547 | 0.506 |
| GCN | 0.59 | 0.575 | 0.473 |
| GAT | 0.608 | 0.587 | 0.476 |
| MOGONET | 0.6 | 0.593 | 0.537 |
| MVGNN | 0.621 | 0.597 | 0.541 |

TABLE 7 Experimental results of multiple classifications of different models in different subtypes.

| Method | ACC | F1_weighted | F1_macro |
|---|---|---|---|
| SVM | 0.617 | 0.627 | 0.535 |
| RF | 0.621 | 0.635 | 0.543 |
| NN | 0.649 | 0.633 | 0.584 |
| NN_NN | 0.699 | 0.679 | 0.611 |
| NN_VCDN | 0.687 | 0.671 | 0.609 |
| GCN | 0.73 | 0.721 | 0.525 |
| GAT | 0.733 | 0.725 | 0.552 |
| MOGONET | 0.712 | 0.717 | 0.614 |
| MVGNN | 0.735 | 0.725 | 0.636 |

According to Table 6, it can be observed that the MVGNN model proposed in this paper achieves the highest ACC value (0.621), the highest F1_weighted value (0.597), and the highest F1_macro value (0.541) compared to traditional machine learning methods, graph convolutional network models, and the latest integrated multi-omics data methods in the multi-classification results of breast cancer differentiation degree. It can be concluded that the model proposed in this paper outperforms traditional machine learning methods and the latest multi-omics data integration methods in the multi-classification problem of breast cancer differentiation degree.

### 3.4.2 Analysis of experimental results of multiple classifications on subtypes

In the same way, the experimental details in Section 3.4.1 are utilized in this study. The multi-classification results of breast cancer subtypes are luminal A vs. luminal B vs. HER2(+) vs. TNBC. The specific experimental results of the MVGNN model compared with other methods on multi-classification of breast cancer subtypes are presented in Table 7.

TABLE 8 Results of any two classifications of different modules in breast cancer differentiation.

| Method | GCN + VCDN | GAT + VCDN | GAT + Attention | GCN + Attention |
|---|---|---|---|---|
| ACC | 0.744 | 0.696 | 0.726 | 0.778 |
| AUC | 0.731 | 0.563 | 0.741 | 0.745 |
| F1_weight | 0.772 | 0.653 | 0.682 | 0.807 |

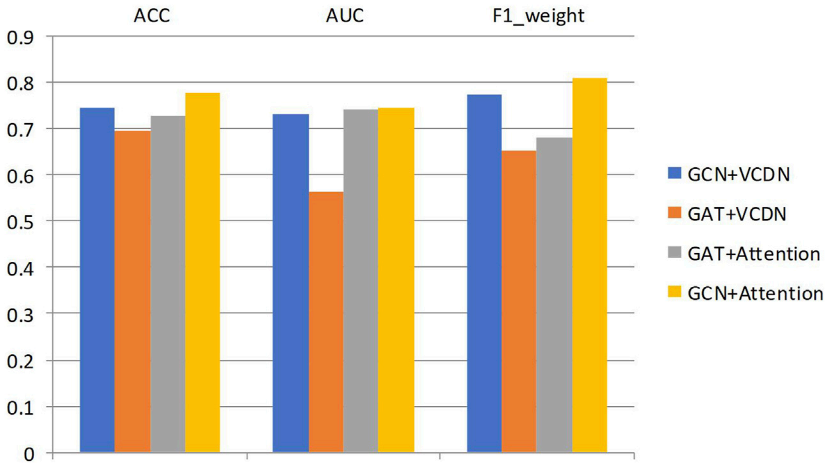


FIGURE 3
Results of any two classifications of different modules in breast cancer differentiation.

According to Table 7, it can be observed that the MVGNN model proposed in this paper, as compared to traditional machine learning methods, graph convolutional network models, and the latest integrated multi-omics data approaches, achieves the best performance in the multi-classification of breast cancer subtypes. The corresponding performance measures are the accuracy (ACC) value of 0.735, the weighted F1 score (F1_weighted) value of 0.725, and the macro F1 score (F1_macro) value of 0.636. Hence, these results are sufficient to demonstrate the effectiveness of the proposed model in this study.

## 3.5 Ablation experiments

### 3.5.1 The performance of different network module

- Analysis of experimental results on differentiation classification

To select the module most beneficial for breast cancer differentiation and subtype classification in the model, this study employed a five-fold cross-validation approach to assess the performance of different modules on the same test dataset. For all models, the same training and validation sets were utilized.

Specifically, this study performed 5-fold cross-validation on the training dataset, with all modules utilizing the same training, validation, and test sets. Mean accuracy, AUC value and mean F1 value were used as measurement metrics. The detailed experimental results of different modules on two types of breast cancer differentiations are presented in Table 8; Figure 3.

By comparing the experimental results of GCN + VCDN and GAT + VCDN, as well as GAT + Attention and GCN + Attention, in predicting any two types of breast cancer differentiations, it can be observed that there exists a specific correlation between biological genomic data. The GAT module did not utilize this correlated information, while the GCN module was able to fully exploit the correlations between biological data, resulting in better differentiation prediction outcomes. Similarly, by comparing the experimental results of GCN + VCDN and GCN + Attention, as well as GAT + VCDN and GAT + Attention, it was found that introducing the attention module improved the performance of predicting breast cancer differentiation. This is because the attention mechanism in the attention module can identify more important neighbors, enabling better classification of breast cancer differentiation. Therefore, this study chose the GCN + Attention model, the MVGNN model, as the final model for predicting breast cancer differentiation.

- Analysis of experimental results on subtype classification

Similarly, the experimental setup for predicting breast cancer differentiation was used. The specific experimental results of different modules on any two breast cancer subtypes are shown in Table 9; Figure 4.

By comparing the experimental results of GCN + VCDN and GAT + VCDN, as well as GAT + Attention and GCN + Attention in predicting two different subtypes of breast cancer, it can be observed that the introduction of the GCN module can improve the accuracy of breast cancer subtype prediction to a certain extent. This is because GCN can effectively utilize the correlation in the biological data. Similarly, by comparing the experimental results

TABLE 9 Results of any two classifications of different modules in breast cancer subtypes.

| Method | GCN + VCDN | GAT + VCDN | GAT + Attention | GCN + Attention |
|---|---|---|---|---|
| ACC | 0.892 | 0.818 | 0.888 | 0.918 |
| AUC | 0.916 | 0.51 | 0.854 | 0.953 |
| F1_weight | 0.663 | 0.283 | 0.438 | 0.716 |

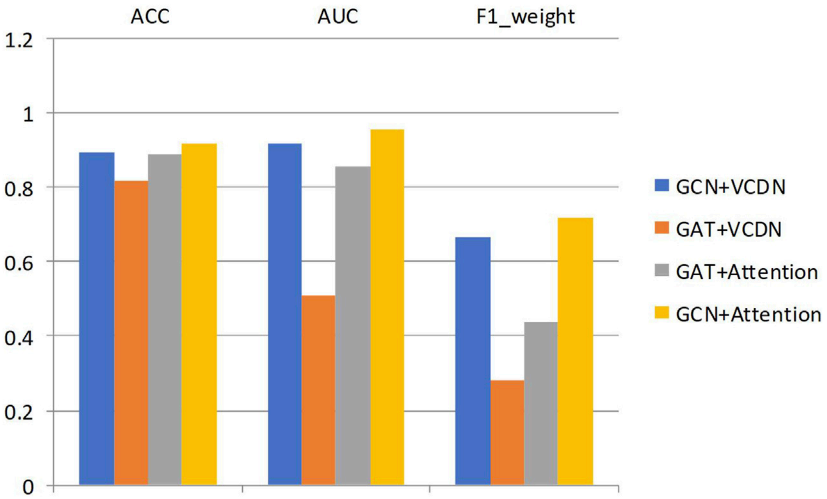


FIGURE 4
Results of any two classifications of different modules in breast cancer subtypes.

of GCN + VCDN and GCN + Attention, as well as GAT + VCDN and GAT + Attention, it can be concluded that the introduction of the attention module increases the precision of predicting breast cancer differentiation. This also indicates that introducing an attention mechanism can improve the model's performance.

### 3.5.2 The performance of multi-omics data fusion

- Analysis of experimental results on differentiation classification

Specifically, for different types of omics data combinations, the same data set partitioning was adopted in this study, and the average accuracy, average AUC value, and average F1 value of 5-fold cross-validation were used as metrics. Figure 5 shows the average accuracy, AUC value, and F1 value of the classification results for different degrees of breast cancer differentiation using different types of omics data. DNA_methylation, mRNA, and CNV in the figure represent the single omics data classification experiments using the MvGNN model with mRNA expression, DNA methylation, and CNV data, respectively. mRNA + DNA_methylation, mRNA + CNV, and DNA_methylation + CNV refer to the classification experiments using two types of omics data simultaneously. mRNA + DNA_methylation + CNV refers to the classification experiments simultaneously using all three types of omics data. The specific experimental results are shown in Table 10; Figure 5.

From Table 10; Figure 5, it can be observed that compared to using a single type of omics data or combining two types of omics data, the model integrating three types of omics data achieved the highest accuracy AUC, and F1 scores in predicting any two subtypes of breast cancer differentiation. The scores were 0.778, 0.803, and 0.809, respectively. This indicates that the model in this study successfully extracted useful information for classification from different omics data.

- Analysis of experimental results on subtype classification

Similarly, this paper uses the dataset partitioning described in Section 3.5.1 and utilizes the average accuracy, average AUC, and average F1 values from 5-fold cross-validation as performance metrics. Experiments were conducted on the classification of any two subtypes of breast cancer using different types of omics data. The integrated model of three omics data achieved the highest accuracy in classifying any two subtypes of breast cancer, with values of 0.921 (luminal A vs. luminal B), 0.968 (luminal A vs. HER2+), 0.91 (luminal A vs. TNBC), 0.82 (luminal B vs. HER2+), 0.964 (luminal B vs. TNBC), and 0.925 (HER2+ vs. TNBC). This indicates that the model proposed in this paper can extract useful information for classification from different omics data. Furthermore, regarding AUC, the integrated model based on three omics data achieved the highest values in classifying any two subtypes of breast cancer, except for the luminal A vs. HER2+ and luminal A vs. TNBC classifications. The respective AUC values were 0.881 (luminal A vs. luminal B), 0.925 (luminal B vs. HER2+), 0.997 (luminal B vs. TNBC), and 0.979 (HER2+ vs. TNBC). Although the model based on three omics data for the luminal A vs. HER2+ classification was 0.6% lower and for the luminal A vs. TNBC classification was 1.2% lower compared to the models integrating mRNA expression data and CNV data or DNA methylation data, respectively, this still demonstrates the robustness of the proposed model in handling

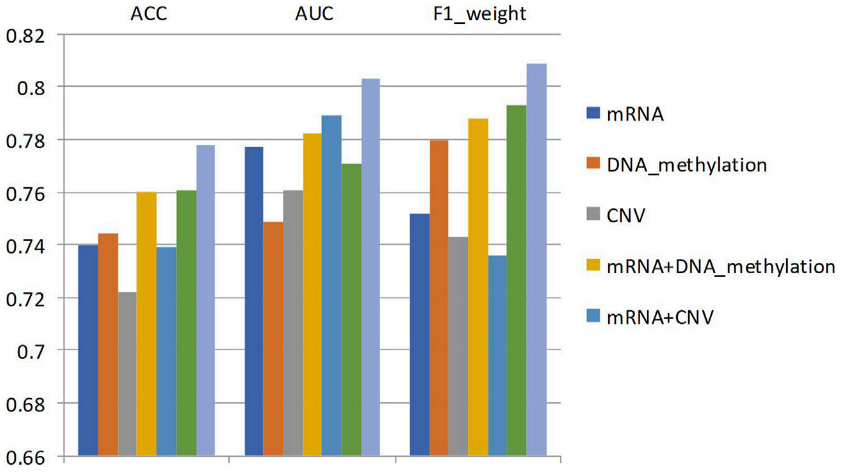


FIGURE 5
The classification results of any two types of breast cancer differentiation in MVGNN model with different combination of omics data.

**TABLE 10 The classification results of any two types of breast cancer differentiation in MVGNN model with different combination of omics data.**

| Omics data | ACC | AUC | F1_weight |
|---|---|---|---|
| mRNA | 0.74 | 0.777 | 0.752 |
| DNA_methylation | 0.744 | 0.749 | 0.78 |
| CNV | 0.722 | 0.761 | 0.743 |
| mRNA + DNA_methylation | 0.76 | 0.782 | 0.788 |
| mRNA + CNV | 0.739 | 0.789 | 0.736 |
| DNA_methylation + CNV | 0.761 | 0.771 | 0.793 |
| mRNA + DNA_methylation + CNV | 0.778 | 0.803 | 0.809 |

imbalanced samples. Similarly, the model based on three omics data achieved the highest F1 values in classifying two breast cancer subtypes, except for the luminal A vs. luminal B classification. The respective F1 values were 0.36 (luminal A vs. HER2+), 0.799 (luminal A vs. TNBC), 0.58 (luminal B vs. HER2+), 0.973 (luminal B vs. TNBC), and 0.959 (HER2+ vs. TNBC).

# 4 Conclusion and discussion

## 4.1 Conclusion

The grading and subtyping of cancer, as a complex trait with distinct molecular features, has significant prognostic and therapeutic implications. Therefore, cancer grading and subtyping research is essential for precision medicine and prognostic cancer prediction. In recent years, numerous supervised multi-omics data integration methods have emerged domestically and internationally. However, these methods do not consider the interrelationships between different types of omics data, which may lead to a bias towards a specific type of omics data in the final prediction results. It is crucial to explore how to improve the predictive performance of models by utilizing the interrelationships between different types of omics data.

This study proposes a multi-omics data fusion algorithm based on a heterogeneous graph neural network. The algorithm combines graph convolutional networks and graph attention networks to predict the differentiation and subtypes of cancer. The breast cancer data from TCGA is used in this study, which includes gene expression data, DNA methylation data, copy number variation (CNV) data, differentiation level annotations, and subtype annotations for each breast cancer sample.

First, preprocessing operations, including chi-square test, normalization, and minimum Redundancy Maximum Relevance (mRMR), are performed on the three types of omics data for breast cancer. Then, we conduct experiments using the MVGNN model, traditional machine learning algorithms, and popular multi-omics data integration methods separately for binary and multi-class classification of breast cancer differentiation and subtypes using 5-fold cross-validation. According to the experimental results, our model achieves the best performance in both binary classification of breast cancer differentiation and subtypes, and multi-class classification.

Furthermore, to select the modules in the model that are more conducive to predicting breast cancer differentiation and subtypes, we also perform 5-fold cross-validation to test the performance of different modules on the test set. Finally, to further test the classification prediction performance of the model, we compare the differentiation and subtype experiments using only one type of omics data, two types of omics data, and all three types of omics data. Based on the experimental results, the breast cancer classification predictions using the MVGNN model with all three types of omics data perform better than those using two or just one type of omics data.

## 4.2 Discussion

The MVGNN model proposed in this paper has achieved good results predicting breast cancer differentiation and subtypes, but some work will be carried out in future. For example:

The overall classification performance of the proposed MVGNN model is satisfactory. However, from the experimental results in Section 3.5.2, it can be observed that our model needs improvement in differentiating between luminal A and HER2(+) subtypes, as well as between luminal A and TNBC subtypes in breast cancer. This also indicates that our gene expression, DNA methylation, and CNV data are insufficient to distinguish the boundaries between luminal A and HER2(+) subtypes and luminal A and TNBC subtypes. Therefore, there may be differences in these subtypes of breast cancer in other types of omics data. In future work, we aim to integrate additional omics data, such as metabolomics data and mutation data, to enhance our breast cancer subtype classification model.

This paper primarily trains the MVGNN model on the breast cancer dataset from TCGA. In order to further demonstrate the performance of the MVGNN model in cancer classification and diagnosis, future studies can include additional datasets of different cancers, such as lung cancer, liver cancer, gastric cancer, and colon cancer, which have high mortality rates.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

## Ethics statement

The studies involving humans were approved by the Cancer Genome Atlas (TCGA) Ethics, Law and Policy Group. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

YR: Conceptualization, Writing–review and editing. YG: Data curation, Writing–original draft. WD: Methodology, Project administration, Writing–original draft. WQ: Formal Analysis, Visualization, Writing–original draft. WL: Writing–original draft, Validation, Visualization. QY: Writing–original draft, Validation, Visualization. YL: Writing–original draft, Project administration, Supervision, Writing–review and editing. GL: Writing–original draft, Conceptualization, Supervision, Writing–review and editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors are grateful to the support of the NSFC (61972174, 62372494, 62202334), Guangdong Universities' Innovation Team Project (2021KCXTD015), Guangdong Key Disciplines Project (2021ZDJS138), the Natural Science Foundation of Jilin Provincial Science and Technology Department (20230101234JC), the Science and Technology Project of Jilin Provincial Education Department (JJKH20200328KJ).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1363896/full#supplementary-material

## References

Altelaar, A. F., Munoz, J., and Heck, A. J. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* 14 (1), 35–48. doi:10.1038/nrg3356

Assi, H. A., Khoury, K. E., Dbouk, H., Khalil, L. E., Mouhieddine, T. H., and El Saghir, N. S. (2013). Epidemiology and prognosis of breast cancer in young women. *J. Thorac. Dis.* 5, S2–S8. Suppl 1. doi:10.3978/j.issn.2072-1439.2013.05.24

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA a cancer J. Clin.* 68 (6), 394–424. doi:10.3322/caac.21492

Cai, Z., Poulos, R. C., Liu, J., and Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *iScience* 25 (2), 103798. doi:10.1016/j.isci.2022.103798

Forman, G. (2008). "An extensive empirical study of feature selection metrics for text classification," in Cikm 08: Proceeding of the Acm Conference on Information and Knowledge Mining, Bethesda, Maryland, USA, November 3-7, 1998.

Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., et al. (2022). Breast cancer statistics, 2022. *CA a cancer J. Clin.* 72 (6), 524–541. doi:10.3322/caac.21754

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics* 2009, 869093. doi:10.4061/2009/869093

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* 11 (7), 476–486. doi:10.1038/nrg2795

Holzinger, E. R., and Ritchie, M. D. (2012). Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics* 13 (2), 213–222. doi:10.2217/pgs.11.145

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. Available at: https://arxiv.org/abs/1609.02907.

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* 11 (3), 191–203. doi:10.1038/nrg2732

Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., et al. (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol.* 23 (1), 171. doi:10.1186/s13059-022-02739-2

Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., et al. (2022). MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Front. Genet.* 13, 806842. doi:10.3389/fgene.2022.806842

Linares-Blanco, J., Pazos, A., and Fernandez-Lozano, C. (2021). Machine learning analysis of TCGA cancer data. *PeerJ Comput. Sci.* 7, e584. doi:10.7717/peerj-cs.584

Loibl, S., Poortmans, P., Morrow, M., Denkert, C., and Curigliano, G. (2021). Breast cancer. *Lancet* 397 (10286), 1750–1769. doi:10.1016/S0140-6736(20)32381-3

Malhotra, G. K., Zhao, X., Band, H., and Band, V. (2010). Histological, molecular and functional subtypes of breast cancers. *Cancer Biol. Ther.* 10 (10), 955–960. doi:10.4161/cbt.10.10.13879

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11 (1), 31–46. doi:10.1038/nrg2626

Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K., and Omolo, B. (2021). A stacking ensemble deep learning approach to cancer type classification based on TCGA data. *Sci. Rep.* 11 (1), 15626. doi:10.1038/s41598-021-95128-x

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12 (2), 87–98. doi:10.1038/nrg2934

Pai, S., and Bader, G. D. (2018). Patient similarity networks for precision medicine. *J. Mol. Biol.* 430 (18), 2924–2938. doi:10.1016/j.jmb.2018.05.037

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Analysis Mach. Intell.* 27 (8), 1226–1238. doi:10.1109/TPAMI.2005.159

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406 (6797), 747–752. doi:10.1038/35021093

Prechelt, L. (2012). “Early stopping — but when?,” in *Neural networks: tricks of the trade* Editors G. Montavon, G. B. Orr, and K.-R. Müller 2 (Berlin, Heidelberg: Springer Berlin Heidelberg), 53–67.

Reif, D. M., White, B. C., and Moore, J. H. (2004). Integrated analysis of genetic, genomic and proteomic data. *Expert Rev. proteomics* 1 (1), 67–75. doi:10.1586/14789450.1.1.67

Reis-Filho, J. S., and Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* 378 (9805), 1812–1823. doi:10.1016/S0140-6736(11)61539-0

Shulaev, V. (2006). Metabolomics technology and bioinformatics. *Briefings Bioinforma.* 7 (2), 128–139. doi:10.1093/bib/bbl012

Sieberts, S. K., and Schadt, E. E. (2007). Moving toward a system genetics view of disease. *Mamm. Genome* 18 (6-7), 389–401. doi:10.1007/s00335-007-9040-6

Sobin, L. H., and Fleming, I. D. (1997). “TNM classification of malignant tumors,” in *Union internationale contre le Cancer and the American joint committee on cancer, cancer. 80(9)*. 5, 1803–1804.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A.* 100 (14), 8418–8423. doi:10.1073/pnas.0932692100

Srivastava, N., Hinton, G., and Krizhevsky, A. (2014). Ilya sutskever and ruslan %J journal of machine learning research salakhutdinov. *Dropout A Simple Way Prev. Neural Netw. Overfitting* 15 (1), 1929–1958. doi:10.5555/2627435.2670313

Tabor, J., and Spurek, P. (2014). Cross-entropy clustering, pattern recognition. Available at: https://arxiv.org/abs/1210.5594.

Tao, M., Song, T., Du, W., Han, S., Zuo, C., Li, Y., et al. (2019). Classifying breast cancer subtypes using multiple kernel learning based on omics data. *Genes* 10 (3), 200. doi:10.3390/genes10030200

Velikovi, P., Cucurull, G., Casanova, A., Romero, A., Pietro, L., and Bengio, Y. (2017). “Graph attention networks,” in ICLR, Toulon, France, April 24 - 26, 2017.

Waks, A. G., and Winer, E. P. (2019). Breast cancer treatment: a review. *JAMA* 321 (3), 288–300. doi:10.1001/jama.2018.19323

Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., et al. (2021). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* 12 (1), 3445. doi:10.1038/s41467-021-23774-w

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10 (1), 57–63. doi:10.1038/nrg2484

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer Genome Atlas pan-cancer analysis project. *Nat. Genet.* 45 (10), 1113–1120. doi:10.1038/ng.2764

Xing, X., Yang, F., Li, H., Zhang, J., Zhao, Y., Gao, M., et al. (2021). “An interpretable multi-level enhanced graph attention network for disease diagnosis with gene expression data,” in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, December 9-12, 2021.

Yersal, O., and Barutca, S. (2014). Biological subtypes of breast cancer: prognostic and therapeutic implications. *World J. Clin. Oncol.* 5 (3), 412–424. doi:10.5306/wjco.v5.i3.412

Yiming, Y. (1997). “A comparative study on feature selection in text categorization,” in ICML, Nashville, Tennessee, USA, July 8-12, 1997.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Cheng, Y., Liu, Z., et al. (2020). “Graph neural networks: a review of methods and applications,” in *AI open*.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 23 February 2024
DOI 10.3389/fgene.2024.1354208

Check for updates

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Advait Balaji,
Occidental Petroleum Corporation,
United States
Zhi-Ping Liu,
Shandong University, China
Shihua Zhang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Xingli Guo,
✉ xlguo@mail.xidian.edu.cn

RECEIVED 12 December 2023
ACCEPTED 30 January 2024
PUBLISHED 23 February 2024

CITATION
Wang Y, Guo X, Niu Z, Huang X, Wang B and Gao L (2024), DeepCBS: shedding light on the impact of mutations occurring at CTCF binding sites.
*Front. Genet.* 15:1354208.
doi: 10.3389/fgene.2024.1354208

COPYRIGHT
© 2024 Wang, Guo, Niu, Huang, Wang and Gao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# DeepCBS: shedding light on the impact of mutations occurring at CTCF binding sites

Yiheng Wang, Xingli Guo*, Zhixin Niu, Xiaotai Huang,
Bingbo Wang and Lin Gao

School of Computer Science and Technology, Xidian University, Xi'an, China

CTCF-mediated chromatin loops create insulated neighborhoods that constrain promoter-enhancer interactions, serving as a unit of gene regulation. Disruption of the CTCF binding sites (CBS) will lead to the destruction of insulated neighborhoods, which in turn can cause dysregulation of the contained genes. In a recent study, it is found that CTCF/cohesin binding sites are a major mutational hotspot in the cancer genome. Mutations can affect CTCF binding, causing the disruption of insulated neighborhoods. And our analysis reveals a significant enrichment of well-known proto-oncogenes in insulated neighborhoods with mutations specifically occurring in anchor regions. It can be assumed that some mutations disrupt CTCF binding, leading to the disruption of insulated neighborhoods and subsequent activation of proto-oncogenes within these insulated neighborhoods. To explore the consequences of such mutations, we develop DeepCBS, a computational tool capable of analyzing mutations at CTCF binding sites, predicting their influence on insulated neighborhoods, and investigating the potential activation of proto-oncogenes. Futhermore, DeepCBS is applied to somatic mutation data of liver cancer. As a result, 87 mutations that disrupt CTCF binding sites are identified, which leads to the identification of 237 disrupted insulated neighborhoods containing a total of 135 genes. Integrative analysis of gene expression differences in liver cancer further highlights three genes: ARHGEF39, UBE2C and DQX1. Among them, ARHGEF39 and UBE2C have been reported in the literature as potential oncogenes involved in the development of liver cancer. The results indicate that DQX1 may be a potential oncogene in liver cancer and may contribute to tumor immune escape. In conclusion, DeepCBS is a promising method to analyze impacts of mutations occurring at CTCF binding sites on the insulator function of CTCF, with potential extensions to shed light on the effects of mutations on other functions of CTCF.

KEYWORDS

**CTCF, insulated neighborhoods, liver cancer, proto-oncogene, deep learning**

## 1 Introduction

3D genomics is a rapidly growing field that investigates the complex folding and organization of chromosomes in eukaryotic cells. Various techniques have been developed to study the 3D structure of chromosomes, such as 3C (Dekker et al., 2002), Hi-C (Lieberman-Aiden et al., 2009), ChIA-PET (Fullwood et al., 2009), 4C (Simonis et al., 2009), and 5C (Dostie et al., 2006). The three-dimensional structure of chromosomes includes chromosomal domains, chromosome compartments, topologically associated domains, insulator regions, and promoter-enhancer loops. In 3D genomics, insulated neighborhood is defined as a CTCF-CTCF homodimer that binds with cohesions and contains at least one gene's chromatin loop (Dowen et al., 2014). Miao have observed that this chromatin loop serves as a unit of gene regulation (Yu and Ren, 2017). When the CTCF binding site is disrupted, improper enhancer-promoter interactions can lead to the dysregulation of local genes (Hnisz et al., 2016b).

Recently, a study has revealed that mutations in CTCF binding sites occur frequently in cancer (Katainen et al., 2015). And CTCF/cohesin binding sites are a major mutational hotspot in the cancer genome. Some of these mutations can cause a decrease in CTCF binding, leading to the disappearance of insulated neighborhoods (Hnisz et al., 2016b; Umer et al., 2016). Non-coding mutations at CTCF binding sites have the potential to disrupt insulated neighborhoods, leading to altered gene expression within these regions. This, in turn, could potentially contribute to the development of diseases. Previous research has indicated that there are typically silent proto-oncogenes within insulated neighborhoods. Additionally, the anchoring regions of insulated neighborhoods containing proto-oncogenes undergo frequent somatic mutations in various types of cancer (Hnisz et al., 2016a). In summary, variations in CTCF binding sites in cancer may lead to the disappearance of insulated neighborhoods and the activation of oncogenes, ultimately promoting the development of cancer. So the identification of variants that have the potential to disrupt insulated neighborhood is a critical task.A few studies focous on this task. Zhang et al. has proposed a method, named CTCF-MP, to predict chromatin loops. This method utilizes a machine learning model based on word2vec and boosted trees (Zhang et al., 2018). CTCF-MP algorithm incorporates sequence variations caused by mutations and enables prediction of the influence of such mutations on the formation of chromatin loops. Sequence-based deep learning methods have shown great potential in predicting the impact of genetic variants on insulated neighborhoods. When provided with a pair of DNA sequences of anchors, this model generates a value ranging from 0 to 1, which can be used to determine the probability or strength of the chromatin loop (Zhang et al., 2018). DeepCTCFLoop takes a pair of DNA sequence containing CTCF motifs with flanking regions and encodes it into one-hot encoding as input, uses a neural network to predict whether this pair of sequences can form a DNA loop (Kuang and Wang, 2021). DeepMILO, a deep learning framework, utilizes one-hot encoding to represent DNA sequences, comprises of an anchor model and an anchor orientation model. It accurately predicts the effects of variants on CTCF/cohesion mediated insulator loops and reveals a novel mechanism for oncogene dysregulation in malignant lymphoma (Trieu et al., 2020).

However, CTCF is a multifunctional protein, associated with a number of vital cellular processes such as transcriptional activation, repression, insulation, imprinting and genome organization (Oh et al., 2017). CTCF not only regulates gene expression by forming loops but also can independently regulate gene expression.

While the discussed methods, such as DeepMILO, have certain limitations as they require paired data, making them effective in predicting the impact of mutations occurring at CTCF binding sites on their insulator function. These methods are not capable of predicting the effects of mutations at CTCF binding sites on other function.

Both DeepMILO and DeepCTCFLoop utilize the one-hot encoding method. One-hot encoding treats each position in the sequence as an independent feature, disregarding the sequential relationships between adjacent nucleotides. However, biological sequences often contain important sequence patterns or motifs that play a critical role in the functionality or structure of the sequence. In comparison, using only one-hot encoding may not fully capture the information conveyed by these patterns. In a recent study, a novel method named dna2vec, has been proposed. This method leverages the human genome sequences as the learning corpus and embeds k-mers into a 100-dimensional continuous vector space (Ng, 2017). By employing this encoding approach, the model can capture a more comprehensive set of information, enhancing its ability to capture relevant patterns and features in the sequences.

Considering the limitations of the discussed methods, we have developed a method named DeepCBS, which employs a DNA sequence as input instead of a paired sequence and utilizes the dna2vec encoding method for representation. Applying DeepCBS to somatic mutation data of liver cancer patients, we predicted the impact of these mutations on CTCF binding sites. Then, through analysis of differential gene expression, we identify three potential liver cancer oncogenes, providing potential therapeutic targets for the treatment of liver cancer. In our study, DeepCBS successfully predicts the impact of mutations occurring at CTCF binding sites on insulated neighborhoods. In the future, it can also be utilized to predict the effects of mutations on other functions of CTCF.

# 2 Materials and methods

## 2.1 Data collection and processing

CTCF ChIP-seq data for GM12878, HepG2, K562, MCF-7, and HMEC cell lines is downloaded from the Encode portal (accession: ENCFF710VEH, ENCFF237OKO, ENCFF738TKN, ENCFF738TKN, ENCFF288RFS). We also download RAD21 CHIP-seq raw data of GM12878 (accession: ENCFF002CPK) and CTCF CHIA-PET raw data of GM12878 (accession: ENCFF780PGS). We download comprehensive gene annotation data from GENCODE.

Positive samples are generated by selecting 100 base pairs from the summit of each ChIP-seq peak. Negative samples are generated using the R package gkmSVM by matching the repeat fraction, length, and GC content of the repetitive sequences in positive samples (Ghandi et al., 2016). Then we get 43,631 positive and 48,753 negative samples for GM12878 cell line, 60,229 positive and 56,099 negative samples for HepG2 cell line, 56,889 positive samples and 53,875 negative samples for K562 cell line.

We collect simple somatic mutations data of 1706 liver cancer patients from ICGC database, we also collected RNA-seq data of

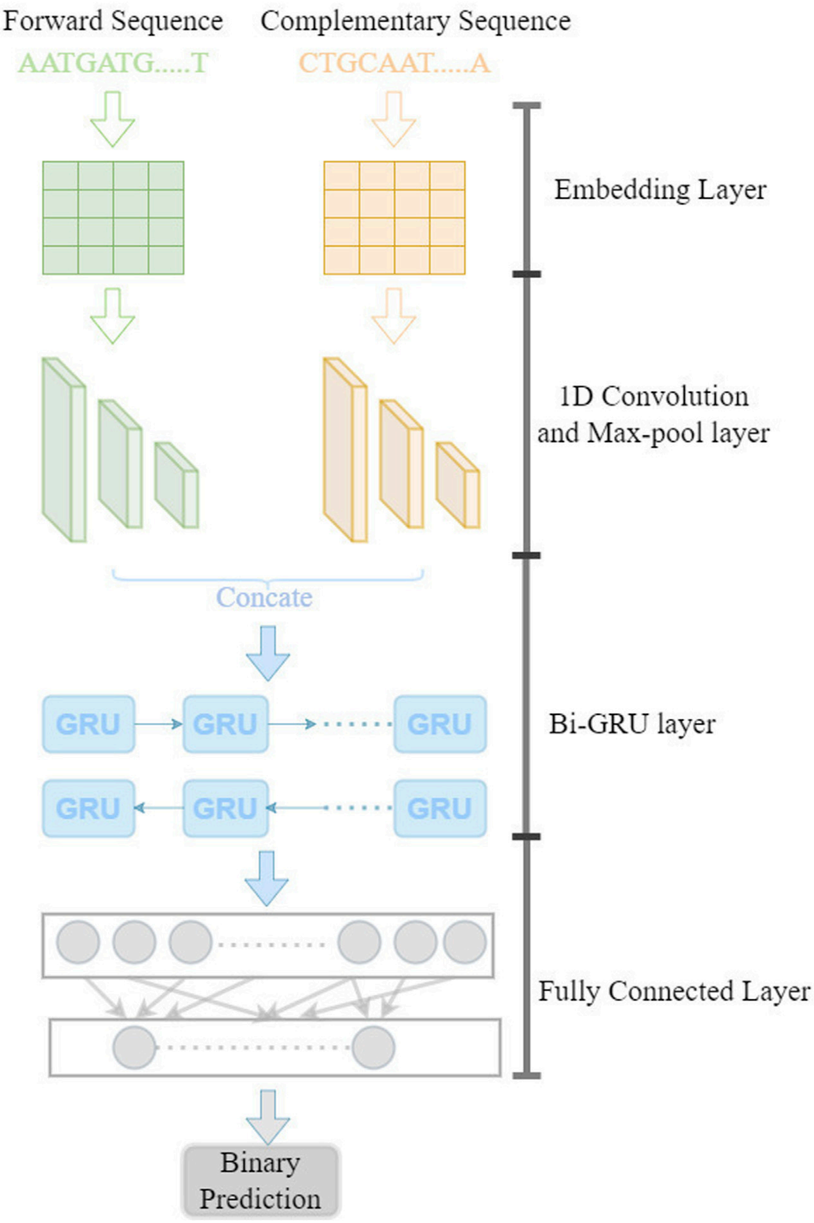


FIGURE 1
Diagram of DeepCBS.

liver cancer from this database, at the same time (see the Supplementary Material).

## 2.2 Construction of DeepCBS

The model is illustrated in Figure 1. In this model, the forward and reverse DNA sequence with CTCF binding are taken as input by encoding into a matrix using the dna2vec (Ng, 2017) approach. Then, a three-layer convolutional neural network is used to learn the sequence motifs and high level features. The Bi-GRU(Bidirectional Gate Recurrent Unit) layer is used to learn the long-range dependencies between the high-level features. Next, two fully connected layer is used to combine the output from the Bi-GRU layer and make the binary prediction.

## 2.3 Identification of disrupted insulated neighborhoods

We obtain insulated neighborhoods by integrating CTCF ChIA-PET data, CTCF ChIP-seq data, RAD21 CHIP-seq data and comprehensive gene annotation data. Specifically, we define an insulated neighborhood as a CTCF loop whose loop anchors overlap with a CTCF CHIP-seq peak and a RAD21CHIP-seq peak, and which contains at least one gene.

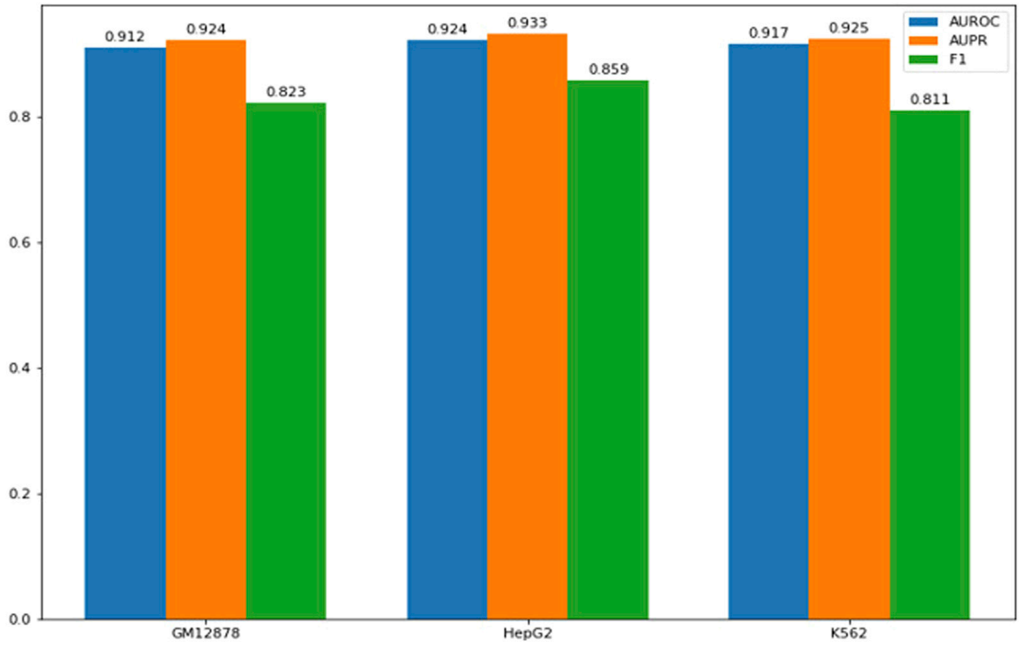


FIGURE 2
Performance of models on cell type specific CTCF binding sites. The *x*-axis represents the cell lines comprising the training set, while the *y*-axis represents the performance scores.

We collect somatic mutation data from liver cancer patients, identify mutated insulated neighborhoods, and utilize the deep learning model proposed in the previous step to predict whether these insulated neighborhoods would be disrupted.

## 2.4 Differential gene expression analysis

We obtain gene expression data from liver cancer patients in ICGC databases. To analyze gene differential expression, we utilized 3 R packages, namely, limma, edgeR, and DESeq2, independently. In order to enhance the robustness of our findings, we obtain differentially expressed genes by taking the intersection of the results from the three packages.

# 3 Results

## 3.1 Workflow of DeepCBS

To elucidate the impact of non-coding mutations occurring at CTCF binding sites, we develop a method named DeepCBS, comprising the following main steps. Initially, we generate positive and negative samples from CTCF ChIP-seq data for 3 cell lines (GM12878, HepG2, K562). Using this data, we train a deep learning model to predict whether mutations on CTCF binding sites lead to the loss of CTCF binding at those sites. Subsequently, we obtain RAD21-mediated loops from RAD21 CHIA-PET data, defining a loop as an insulated neighborhood if both anchors of the loop overlap with CTCF CHIP-seq peaks. And if there are mutations within the CHIP-seq peak region that overlaps with loop anchors, then the insulated neighborhood is considered as a mutated insulated neighborhood. Leveraging the well-trained deep learning model, we predict whether mutations within the mutated insulated neighborhoods disrupt the binding of CTCF, resulting in the disruption of the insulated neighborhooods. In the next step, we observe a significant enrichment of proto-oncogenes in mutated insulated neighborhoods, suggesting that the disruption of these neighborhoods may play a crucial role in cancer development. Consequently, we identify the genes within the disrupted insulated neighborhoods and intersect these genes with the differentially expressed genes in liver cancer. This process yield three genes that may undergo upregulation due to the disruption of insulated neighborhoods. Notably, two out of the three genes have been previously reported as potential oncogenes in liver cancer. The remaining gene, DQX1, is identified as a potential liver cancer oncogene through bioinformatics analysis Figure 2.

**TABLE 1** Cross cell performance evaluation.

| Train set | Test set | Auroc |
|---|---|---|
| GM12878 | HepG2 | 0.97945 |
| GM12788 | K562 | 0.97364 |
| K562 | HepG2 | 0.97965 |
| K562 | GM12878 | 0.98921 |
| HepG2 | K562 | 0.97507 |
| HepG2 | GM12878 | 0.99106 |

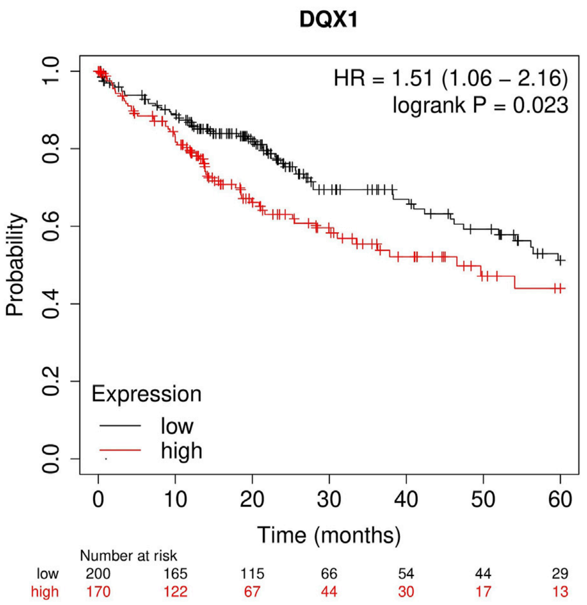


FIGURE 3
Correlation between DQX1 gene expression and survival prognosis of liver cancer.

## 3.2 Performance of DeepCBS

To assess the model's performance across different cell types, we employe 1 cell type's samples as the training data and used samples from other cell types as the testing data, as shown in Table 1. As it can be seen, the model has achieved AUC values over 0.97, demonstrating that our method is a powerful tool for identifying CTCF binding sites.

Since some CTCF binding sites are cell-type specific, we collect CTCF CHIP-seq data from the MCF-7 cell line (breast cancer cell line) and HMEC cell line (normal breast epithelial cell line), and get cell-type specific CTCF binding sites in the MCF-7 cell line. As shown in Figure 2, our models have demonstrated excellent performance on cell-type-specific binding sites.

## 3.3 Identification of potential oncogenes in disrupted insulated neighborhoods in liver cancer

We apply a hypergeometric distribution test to our data and find a significant enrichment of proto-oncogenes in mutated insulated neighborhoods ($p < 0.05$). We have also observed this phenomenon in the data provided by Ji (Ji et al., 2016). This suggests that the disruption of insulated neighborhoods may be a key driver of cancer development, as it can lead to the abnormal activation of proto-oncogene into oncogene. We identify 237 disrupted insulated neighborhoods, comprising a total of 135 genes. We perform differential gene expression analysis, then identify 1,218 differentially expressed genes using 3 R packages. To explore which genes among the 135 affected genes in the disrupted insulated neighborhoods are key genes related to cancer. Then, we take the intersection of the differentially expressed genes with the genes located within the disrupted insulated neighborhoods, which resulted in the identification of three key genes: ARHGEF39, UBE2C, and DQX1. And all of them are upregulated genes, potentially activated due to the disruption of insulated neighborhoods.

ARHGEF39 is a novel member of the Dbl-family of guanine nucleotide exchange factors (Wang et al., 2012). Guanine nucleotide exchange factors are recognized as crucial activators of Rho GTPases, which play a significant role in cell migration (Cook et al., 2014; Goicoechea et al., 2014). Overexpress of ARHGEF39 promotes gastric cancer cell proliferation and migration through the Akt signaling pathway (Wang et al., 2018; Zhou et al., 2018). Previous literature has proposed that ARHGEF39 may act as an oncogene in the progression of liver cancer, and thus represents a potential prognostic indicator and therapeutic target for this disease (Gao and Jia, 2019). Ubiquitin-conjugating enzyme E2C(UBE2C), a member of the E2family, is encoded by the UbcH10gene situated on human chromosome20q13.12. Its function involves the degradation of various target proteins through catalysis. UBE2C has been found to be upregulated in various types of cancer, including breast cancer, and is considered a potent proto-oncogene associated with tumor malignancy (Chou et al., 2014; Han et al., 2015). In liver cancer, UBE2C has been identified as a potential oncogene that can promote cell proliferation, migration, invasion, and drug resistance (Xiong et al., 2019).

Based on the above, we speculate that in liver cancer, the overexpression of ARHGEF39 and UBE2C serves as activated oncogenes and is involved in liver cancer development due to the disruption of the insulated neighborhoods containing them. However, there is currently no literature exploring the role of the DQX1 in liver cancer.

## 3.4 Overexpression of DQX1 is oncogenic in liver cancer

The Kaplan-Meier plotter (https://kmplot.com/analysis/) is a powerful tool that enables the assessment of the impact of 54k genes (including mRNA, miRNA, and protein) on survival across 21 types of cancer (Győrffy, 2023). In this study, we focus on the analysis of the relationship between DQX1 expression and survival in liver cancer. The result of survival analysis, as shown in Figure 3, show that highly expressed DQX1 is linked to poor prognosis of overall survival (OS) for cancers of liver cancer.

Gene set enrichment analysis (GSEA) is further performed to explore the signaling pathways and molecular mechanisms that were differentially affected by DQX1 in liver cancer. In this study, the tumor samples are grouped based on the mean expression level of DQX1. Samples with expression levels higher than the mean are assigned to the high-expression group (DQX1. Hi), while those with expression levels lower than the mean are assigned to the low-expression group (DQX1. Low). In our study, the Hallmark database is utilized for performing the gene set enrichment analysis. As depicted in Figure 4, the analysis reveal that high expression of DQX1 is significantly associated with the activation of cell proliferation-related pathways. This finding suggests that DQX1may play a crucial role in promoting cell proliferation in liver cancer.

DQX1 is one of the RNA-binding protein genes and RNA-binding protein can regulate the infiltration degrees of immune cells (Sun et al., 2021). Therefore, we implement an immunological

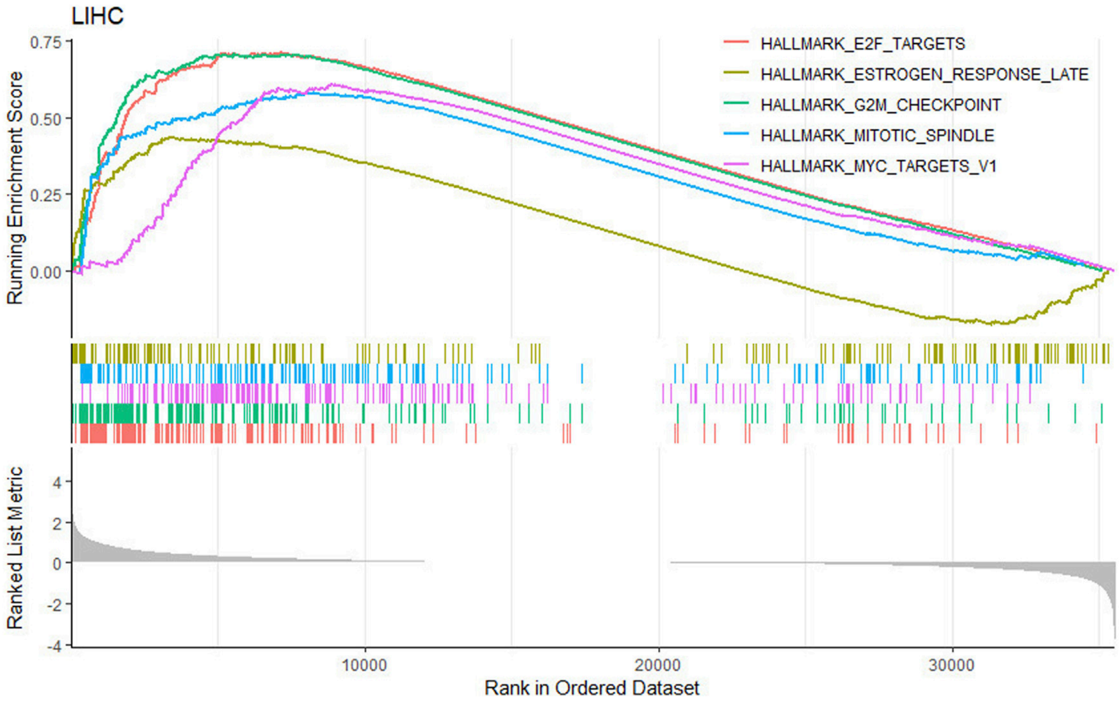


FIGURE 4
Signaling pathways associate with DQX1 in liver cancer. The depicted signaling pathways in the figure are all cell proliferation-related pathways.

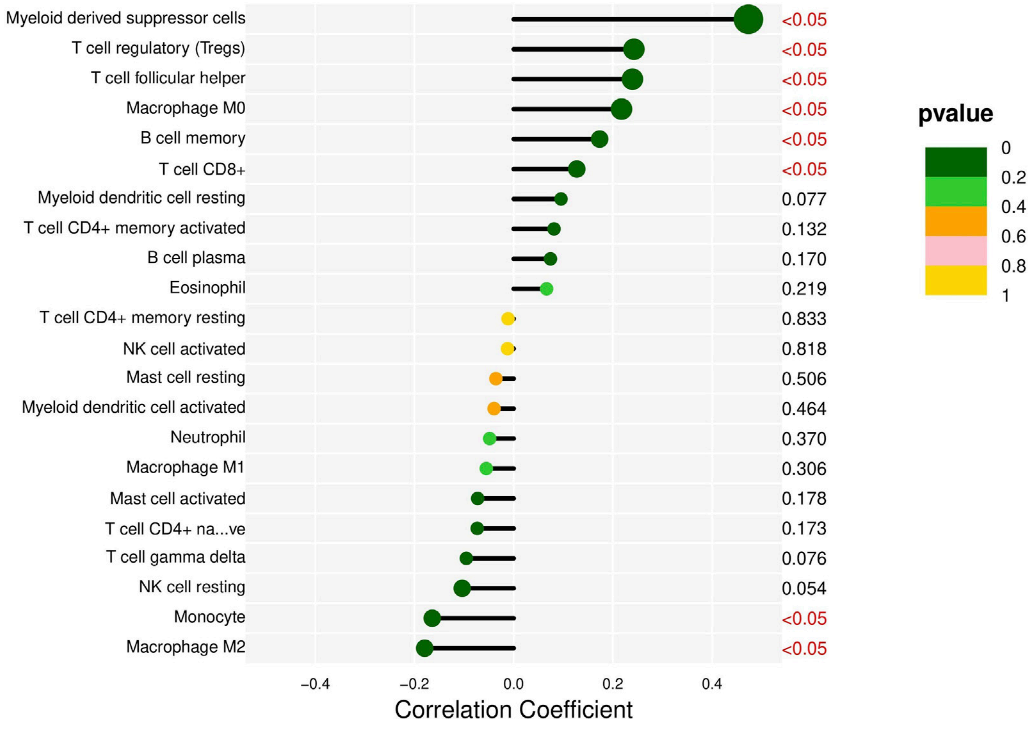


FIGURE 5
The correlation between DQX1 and immune cell infiltration in liver cancer.

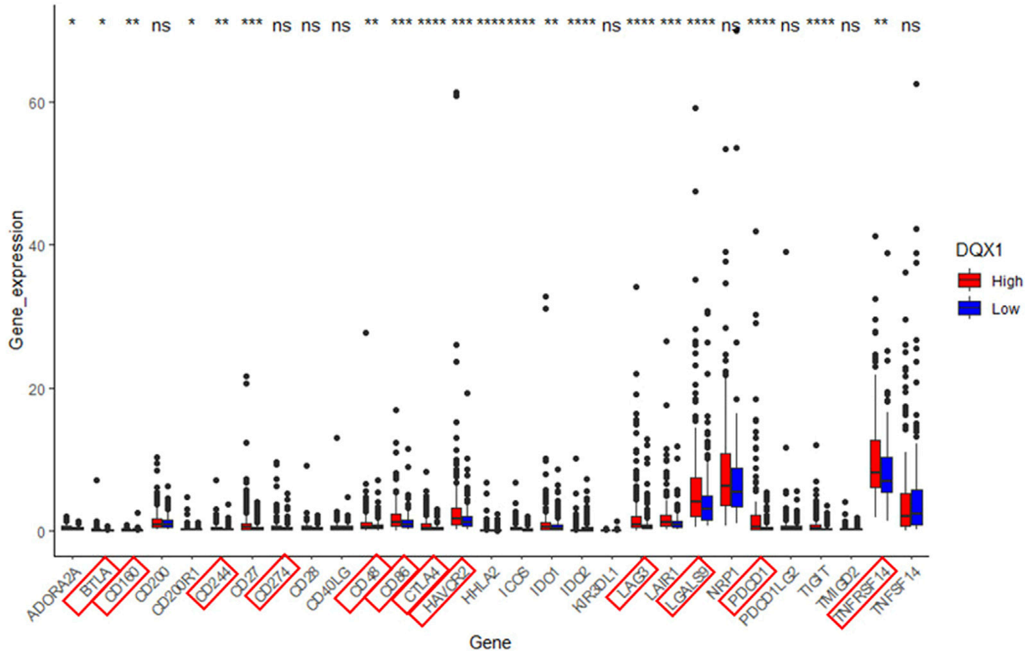


FIGURE 6
Significant analysis of differential expression of immune checkpoint genes between the high-expression group (DQX1. High) and the low-expression group (DQX1. Low). The markers enclosed within the red box represent co-suppressive immune checkpoints. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. ns: non-significant differences ($p > 0.05$).

Analysis of DQX1 in liver cancer. TIMER, a comprehensive online resource for the systematic analysis of immune infiltrates in various cancer types, is employed in this study to explore the correlation between DQX1 expression in liver cancer and different immune infiltrates (Li et al., 2020). The results are presented in Figure 5.

Also, we perform differential expression analysis of immune checkpoint genes in relation to DQX1 using the same grouping approach as in GSEA. The result is depicted in Figure 6.

Overall, the expression of DQX1 shows significant positive correlations with immune infiltration levels of regulatory T cells (Tregs), myeloid-derived suppressor cells (MDSCs), and expressions of co-suppressive immune checkpoints, contributing to immune escape. This suggests that we can develop immunotherapies targeting DQX1 for the treatment of liver cancer, in the future.

Based on the bioinformatics analysis, we have been inferred that DQX1 may potentially act as an oncogene and be involved in the development of liver cancer.

## 4 Conclusion

In summary, the CTCF play an crucialrole in maintaining these insulated neighborhoods. The disruption of CTCF binding sites can lead to dysregulation of contained genes, potentially resulting in the activation of oncogenes and promoting cancer development. It is important to shed light on the impact of mutations occurring at CTCF binding sites. So we develop a novel method, DeepCBS, to analyze the impact of mutations occurring at CTCF binding sites. Our analysis has identified three potential oncogenes, ARHGEF39, UBE2C, and DQX1 of liver cancer. All three genes play an oncogenic role in the development of liver cancer. And overexpression of DQX1 is associated with poor prognosis and tumor immune escape. Our findings demonstrate the potential of DeepCBS to analyze the impact of mutations occurring at CTCF binding sites, as well as providing valuable insights for the diagnosis and treatment of liver cancer. Over all, this study emphasizes the importance of understanding the 3D organization of the human genome and its impact on gene regulation, as well as highlights the potential of computational methods to identify new targets for cancer therapy.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

## Ethics statement

The manuscript presents research on animals that do not require ethical approval for their study.

## Author contributions

YW: Methodology, Visualization, Writing–original draft. XG: Conceptualization, Writing–review and editing. ZN: Writing–review and editing. XH: Writing–review and editing. BW: Writing–review and editing. LG: Writing–review and editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (61672407, 62132015, U22A2037, Nos. 62172318) and the Fundamental Research Funds for the Central Universities (ZYTS23208).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1354208/full#supplementary-material

## References

Chou, C.-P., Huang, N.-C., Jhuang, S.-J., Pan, H.-B., Peng, N.-J., Cheng, J.-T., et al. (2014). Ubiquitin-conjugating enzyme UBE2C is highly expressed in breast microcalcification lesions. *PLoS One* 9, e93934. doi:10.1371/journal.pone.0093934

Cook, D. R., Rossman, K. L., and Der, C. J. (2014). Rho guanine nucleotide exchange factors: regulators of Rho GTPase activity in development and disease. *Oncogene* 33, 4021–4035. doi:10.1038/onc.2013.362

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science* 295, 1306–1311. doi:10.1126/science.1067799

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 16, 1299–1309. doi:10.1101/gr.5571506

Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* 159, 374–387. doi:10.1016/j.cell.2014.09.030

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462, 58–64. doi:10.1038/nature08497

Gao, J., and Jia, W.-D. (2019). Expression of rho guanine nucleotide exchange factor 39 (ARHGEF39) and its prognostic significance in hepatocellular carcinoma. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* 25, 7826–7835. doi:10.12659/MSM.918270

Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M. A. (2016). gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* 32, 2205–2207. doi:10.1093/bioinformatics/btw203

Goicoechea, S. M., Awadia, S., and Garcia-Mata, R. (2014). I'm coming to GEF you: regulation of RhoGEFs during cell migration. *Cell adhesion Migr.* 8, 535–549. doi:10.4161/cam.28721

Győrffy, B. (2023). Discovery and ranking of the most robust prognostic biomarkers in serous ovarian cancer. *Geroscience* 45, 1889–1898. doi:10.1007/s11357-023-00742-4

Han, Q., Zhou, C., Liu, F., Xu, G., Zheng, R., and Zhang, X. (2015). MicroRNA-196a post-transcriptionally upregulates the UBE2C proto-oncogene and promotes cell proliferation in breast cancer. *Oncol. Rep.* 34, 877–883. doi:10.3892/or.2015.4049

Hnisz, D., Day, D. S., and Young, R. A. (2016a). Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell* 167, 1188–1200. doi:10.1016/j.cell.2016.10.024

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., et al. (2016b). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 351, 1454–1458. doi:10.1126/science.aad9024

Ji, X., Dadon, D. B., Powell, B. E., Fan, Z. P., Borges-Rivera, D., Shachar, S., et al. (2016). 3D chromosome regulatory landscape of human pluripotent cells. *Cell stem Cell* 18, 262–275. doi:10.1016/j.stem.2015.11.007

Katainen, R., Dave, K., PitkäNEN, E., Palin, K., Kivioja, T., VäLIMäKI, N., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* 47, 818–821. doi:10.1038/ng.3335

Kuang, S., and Wang, L. (2021). Deep learning of sequence patterns for CCCTC-binding factor-mediated chromatin loop formation. *J. Comput. Biol.* 28, 133–145. doi:10.1089/cmb.2020.0225

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* 326, 289–293. doi:10.1126/science.1181369

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2. 0 for analysis of tumor-infiltrating immune cells. *Nucleic acids Res.* 48, W509–W514. doi:10.1093/nar/gkaa407

Ng, P. (2017). dna2vec: consistent vector representations of variable-length k-mers. *arXiv preprint arXiv:1701.06279.*

Oh, S., Oh, C., and Yoo, K. H. (2017). Functional roles of CTCF in breast cancer. *BMB Rep.* 50, 445–453. doi:10.5483/bmbrep.2017.50.9.108

Simonis, M., Klous, P., Homminga, I., Galjaard, R.-J., Rijkers, E.-J., Grosveld, F., et al. (2009). High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. *Nat. methods* 6, 837–842. doi:10.1038/nmeth.1391

Sun, D., Yang, K.-S., Chen, J.-L., and Wang, Z.-B. (2021). Identification and validation of an immune-associated RNA-binding proteins signature to predict clinical outcomes and therapeutic responses in colon cancer patients. *World J. Surg. Oncol.* 19, 314–413. doi:10.1186/s12957-021-02411-2

Trieu, T., Martinez-Fundichely, A., and Khurana, E. (2020). DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. *Genome Biol.* 21, 79–11. doi:10.1186/s13059-020-01987-4

Umer, H. M., Cavalli, M., Dabrowski, M. J., Diamanti, K., Kruczyk, M., Pan, G., et al. (2016). A significant regulatory mutation burden at a high-affinity position of the CTCF motif in gastrointestinal cancers. *Hum. Mutat.* 37, 904–913. doi:10.1002/humu.23014

Wang, H., Li, M., Tao, X., Qian, Y., Chen, L., and Tao, G. (2018). ARHGEF39 promotes gastric cancer cell proliferation and migration via Akt signaling pathway. *Mol. Cell. Biochem.* 440, 33–42. doi:10.1007/s11010-017-3153-3

Wang, H., Li, Y., Wang, Y., Han, Z.-G., and Cai, B. (2012). C9orf100, a new member of the Dbl-family guanine nucleotide exchange factors, promotes cell proliferation and migration in hepatocellular carcinoma. *Mol. Med. Rep.* 5, 1169–1174. doi:10.3892/mmr.2012.783

Xiong, Y., Lu, J., Fang, Q., Lu, Y., Xie, C., Wu, H., et al. (2019). UBE2C functions as a potential oncogene by enhancing cell proliferation, migration, invasion, and drug resistance in hepatocellular carcinoma cells. *Biosci. Rep.* 39, 82384. doi:10.1042/BSR20182384

Yu, M., and Ren, B. (2017). The three-dimensional organization of mammalian genomes. *Annu. Rev. Cell Dev. Biol.* 33, 265–289. doi:10.1146/annurev-cellbio-100616-060531

Zhang, R., Wang, Y., Yang, Y., Zhang, Y., and Ma, J. (2018). Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics* 34, i133–i141. doi:10.1093/bioinformatics/bty248

Zhou, H., Cai, L., Zhang, X., Li, A., Miao, Y., Li, Q., et al. (2018). ARHGEF39 promotes tumor progression via activation of Rac1/P38 MAPK/ATF2 signaling and predicts poor prognosis in non-small cell lung cancer patients. *Lab. Investig.* 98, 670–681. doi:10.1038/s41374-018-0022-y

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 13 March 2024
DOI 10.3389/fgene.2024.1367731

Check for updates

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Advait Balaji,
Occidental Petroleum Corporation,
United States
Wei Lan,
Guangxi University, China

*CORRESPONDENCE
Le Zhang,
✉ zhangle06@scu.edu.cn

RECEIVED 09 January 2024
ACCEPTED 08 February 2024
PUBLISHED 13 March 2024

CITATION
Xiao M, Xiao Y, Yu J and Zhang L (2024), PCGIMA: developing the web server for human position-defined CpG islands methylation analysis.
*Front. Genet.* 15:1367731.
doi: 10.3389/fgene.2024.1367731

COPYRIGHT
© 2024 Xiao, Xiao, Yu and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# PCGIMA: developing the web server for human position-defined CpG islands methylation analysis

Ming Xiao[1,2], Yi Xiao[1], Jun Yu[3,4] and Le Zhang[1,5,6]*

[1]College of Computer Science, Sichuan University, Chengdu, China, [2]Tianfu Engineering-oriented Numerical Simulation and Software Innovation Center, Chengdu, China, [3]CAS Key Laboratory of Genome Sciences and Information, Chinese Academy of Sciences, Beijing Institute of Genomics, Beijing, China, [4]University of Chinese Academy of Sciences, Beijing, China, [5]Key Laboratory of Systems Biology, Chinese Academy of Sciences, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China, [6]Key Laboratory of Systems Health Science of Zhejiang Province, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China

**Introduction:** CpG island (CGI) methylation is one of the key epigenomic mechanisms for gene expression regulation and chromosomal integrity. However, classical CGI prediction methods are neither easy to locate those short and position-sensitive CGIs (CpG islets), nor investigate genetic and expression pattern for CGIs under different CpG position- and interval-sensitive parameters in a genome-wide perspective. Therefore, it is urgent for us to develop such a bioinformatic algorithm that not only can locate CpG islets, but also provide CGI methylation site annotation and functional analysis to investigate the regulatory mechanisms for CGI methylation.

**Methods:** This study develops Human position-defined CGI prediction method to locate CpG islets using high performance computing, and then builds up a novel human genome annotation and analysis method to investigate the connections among CGI, gene expression and methylation. Finally, we integrate these functions into PCGIMA to provide relevant online computing and visualization service.

**Results:** The main results include: (1) Human position-defined CGI prediction method is more efficient to predict position-defined CGIs with multiple consecutive (d) values and locate more potential short CGIs than previous CGI prediction methods. (2) Our annotation and analysis method not only can investigate the connections between position-defined CGI methylation and gene expression specificity from a genome-wide perspective, but also can analysis the potential association of position-defined CGIs with gene functions. (3) PCGIMA (http://www.combio-lezhang.online/pcgima/home.html) provides an easy-to-use analysis and visualization platform for human CGI prediction and methylation.

**Discussion:** This study not only develops Human position-defined CGI prediction method to locate short and position-sensitive CGIs (CpG islets) using high performance computing to construct MR-CpGCluster algorithm, but also a novel human genome annotation and analysis method to investigate the connections among CGI, gene expression and methylation. Finally, we integrate them into PCGIMA for online computing and visualization.

KEYWORDS

**position-defined CGIs, DNA methylation, genome annotation, high performance computing, genome analysis**

## 1 Introduction

CpG island (CGI) methylation is one of the key epigenomic mechanisms for gene expression regulation and chromosomal integrity (Dor and Cedar, 2018). Especially, recent studies have reported that position-sensitive CGI co-methylation mechanism is essential for such functions that are related to histone modification (Ming, et al., 2021). However, it is neither easy for current commonly used classical CGI island prediction methods (Gardinergarden and Frommer, 1987; Han et al., 2008; Takahashi et al., 2017) to locate those short and position-sensitive CGIs which called CpG islets (Hackenberg et al., 2006) due to the length limitation, nor investigate relationship among CGI density, methylation, and gene expression specificity. Therefore, it is urgent for us to develop such a bioinformatic algorithm that not only can locate short and position-sensitive CGIs (CpG islets), but also provide CGI methylation site annotation and functional analysis to investigate the regulatory mechanisms for CGI methylation (http://www.combio-lezhang.online/pcgima/home.html).

For CGI perdition method, we usually employ the unsupervised clustering methods such as CpGCluster (Hackenberg et al., 2006) and CPG_MI (Su et al., 2009) to locate CGIs with shorter length than the supervised (Bock et al., 2007; Ning et al., 2017), since these unsupervised algorithms do not need consider the constraints of CGI length and content ratio (Hackenberg et al., 2010). However, these methods are not only time-consuming for the big dataset, but also cannot investigate the genetic characteristics of CGIs under different CpG interval parameters. Therefore, our first scientific question is how to develop a novel CGI prediction method with CpG interval parameters selective feature and high-performance computing, and investigate the differences in genetic characteristics such as CpG coverage, CGI length, and GC content of CGIs under various CpG interval parameters.

Several previous studies have interrogated the connections between methylation and CGI (Reik, 2007; Smith et al., 2012; Liu et al., 2016; El-Maarri, 2019; Acton et al., 2021). For example, Ziller et al. (2013) have turned out that not only the hypermethylation of promoter CGI is related to gene expression, but also CGI methylation in the gene body is positively correlated with gene expression. However, these studies usually interrogate the methylation characteristics of CGI from partial sequence regions rather than genome-wide perspective. Meanwhile, although our previous studies (Zhang et al., 2018; Zhang et al., 2021a) have analyzed the relationship between CGI density and gene expression after annotating genome-wide CGI-related genes (CGI+) into high-CGI (HCGI), intermediate-CGI (ICGI), and low-CGI (LCGI) genes based on the classification of CGI density (Weber et al., 2007; Zhu et al., 2008), we are still unclear the relationship between CpG methylation and gene expression. Thus, our second scientific question is how to build up a human genome-wide CGI-based methylation and gene expression annotation and analysis method to investigate the relationship among CGI density, methylation, and gene expression specificity.

Meanwhile, although several CpG methylation online service are already available (Raney et al., 2010; Di et al., 2018; Xiong et al., 2019), most of them only focus on CpG island prediction and data downloading, but not provide visualization and analysis for the distribution of CGI in different sequence regions and the connections between methylation status of CGIs and gene expression. Therefore, our third scientific question is how to establish an easy-to-use web service for fast CGIs prediction and visualization of the connections between CGIs and methylation.

For these reasons, we propose three major innovations to answer the above scientific questions.

Firstly, we develop an unsupervised clustering-based CGI prediction method (Human position-defined CGI prediction), which not only employs high performance computing technology to accelerate its predictive speed, but also offers a parameter selective option that can help us to locate short CGIs (position-defined CGIs) with unique location- or sequence-sensitive features and explore the differences in the genetic characteristics of CGIs under various CpG interval parameters.

Secondly, we build up a novel human genome annotation and analysis function (Human position-defined CGI annotation and analysis), which not only can study the methylation characteristics of CGIs from a genome-wide perspective by computing the methylation level of all CpG sites in the human genome, but also improve the previous CpG-Island-based human gene expression annotation and analysis method (Zhang et al., 2021a) by integrating genome-wide methylation annotation to further investigate the connections among CGI density, gene expression and methylation.

Thirdly, we establish an easy-to-use web service "Position-defined CGI methylation analysis (PCGIMA)" with relevant CGI prediction, annotation, and data analysis functions, which provides us an online platform for further study on the regulation mechanism of CGI and methylation.

In conclusion, we develop a bioinformatic algorithm and web service to investigate the regulatory mechanism of CGI methylation. The main results include: 1) Human position-defined CGI prediction method is more efficient to predict position-defined CGIs with multiple consecutive (d) values and locate more potential short CGIs than previous CGI prediction methods; 2) Our annotation and analysis method not only can investigate the connections between position-defined CGI methylation and gene expression specificity from a genome-wide perspective, but also can analyze the potential association of position-defined CGIs with gene functions; (3) PCGIMA provides an easy-to-use analysis and visualization platform for human CGI prediction and methylation.

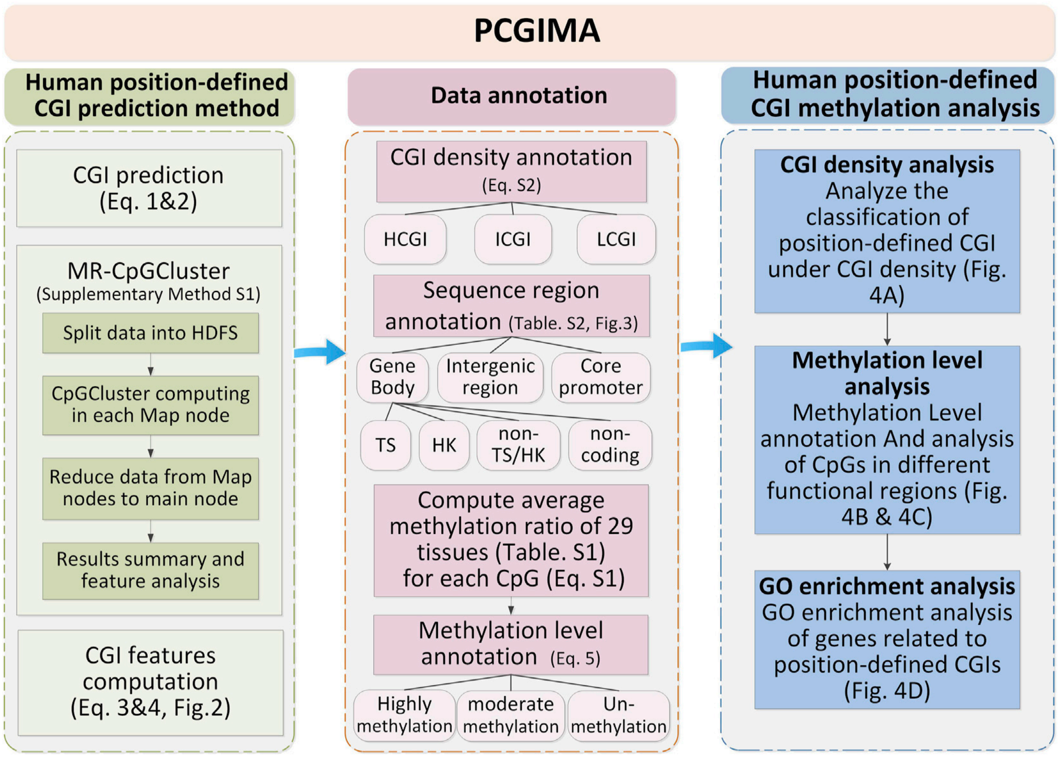

PCGIMA
Human position-defined CGI prediction method
Data annotation
Human position-defined CGI methylation analysis
CGI prediction (Eq. 1&2)
MR-CpGCluster (Supplementary Method S1)
Split data into HDFS
CpGCluster computing in each Map node
Reduce data from Map nodes to main node
Results summary and feature analysis
CGI features computation (Eq. 3&4, Fig.2)
CGI density annotation (Eq. S2)
HCGI
ICGI
LCGI
Sequence region annotation (Table. S2, Fig.3)
Gene Body
Intergenic region
Core promoter
TS
HK
non-TS/HK
non-coding
Compute average methylation ratio of 29 tissues (Table. S1) for each CpG (Eq. S1)
Methylation level annotation (Eq. 5)
Highly methylation
moderate methylation
Un-methylation
CGI density analysis
Analyze the classification of position-defined CGI under CGI density (Fig. 4A)
Methylation level analysis
Methylation Level annotation And analysis of CpGs in different functional regions (Fig. 4B & 4C)
GO enrichment analysis
GO enrichment analysis of genes related to position-defined CGIs (Fig. 4D)

FIGURE 1
Workflow of the study.

## 2 Materials and methods

This study downloads human genome data from GRCh38 assembly (Schneider et al., 2017) at NCBI (Pruitt et al., 2005). To classify CGIs into density-defined and position-defined groups, we download human CGIs data and annotations from UCSC (Casper et al., 2018). Next, we use human genome annotated data (release 24) in GenBank GBFF format (Clark et al., 2016) from GENCODE (Wright et al., 2016) to define different sequence regions. Finally, to study the methylation level of CpG sites in different sequence regions, we obtain all CpG methylation data of 29 human tissues (Supplementary Table S1), including heart, spleen, lung and esophagus, from ENCODE databases (Harrow et al., 2006). In order to ensure data consistency, the above-listed annotation and methylation data are all annotated according to GRCh38 (Schneider et al., 2017). Figure 1 describes the workflow of the study with three essential steps: Human position-defined CGI prediction (left side of Figure 1), Data annotation (right side of Figure 1), and Human position-defined CGI methylation analysis (Bottom side of Figure 1).

Here, we describe the key equations as follows:

(1) CGI prediction: We employed Eq. 1 to define CpGs clusters (Hackenberg et al., 2006) at the start. Next, we consider these CpG clusters with small p-values (Eq. 2) as CGIs (Hackenberg et al., 2006).

$$d_i = x_{i+1} - x_i - 1 \tag{1}$$

Here, x and I represent the position and index of a CpG, respectively.

$$P(d) = (1 - p)^{d-1} p \tag{2}$$

P(d) represents the probability to find a distance d between neighboring CpGs. p corresponds to the probability of CpGs in the sequence. Since our previous studies (Zhang et al., 2018; Zhang et al., 2021a) has led to a conclusion that LAUPs (Lineage-associated underrepresented permutations) are closely related to CGIs and the shortest LAUPs of mammals range from 10bp to 14bp in length, here we use the intermediate value of d = 12bp.

(2) MR-CpGCluster: We develop a MR-CpGCluster algorithm (Supplementary Figure S1) to speed up CGI predict procedure based on MapReduce (Dittrich and Quiané-Ruiz, 2012) and Hadoop Streaming (Dede et al., 2016) techniques detailed by Supplementary Method S1 for Human position-defined CGI prediction method. Finally, our method computes the CGI features of the position-defined CGIs for subsequent analysis.
(3) CGI features computation: To compare the CGIs under different CpG distance intervals (Eq. 1), we compute CGI length, CG content, CpG O/E ratio (Gardinergarden and Frommer, 1987) (Eq. 3) and CpG density (Eq. 4) for each CGI (Hackenberg et al., 2006).

$$\mathrm{O/E} = \frac{\mathrm{CpNum}}{\mathrm{CNum} \times \mathrm{GNum}} \times \mathrm{N} \tag{3}$$

$$\mathrm{CpGdensity} = \frac{\mathrm{CpNum}}{\mathrm{N}} \tag{4}$$

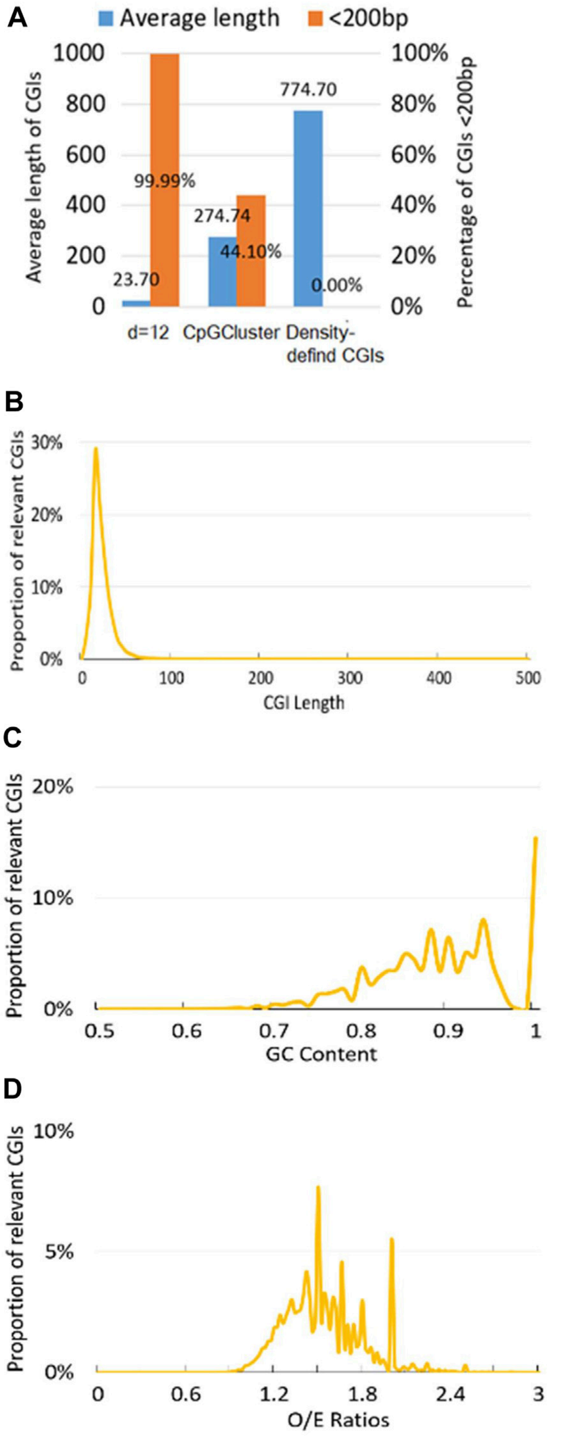


FIGURE 2
Position-defined CGI prediction and analysis. **(A)** CGIs comparative analysis. The proportion distribution of CGI at different **(B)** CGI length, **(C)** GC content, and **(D)** CpG O/E ratios (Eq. 3) under CpG interval d = 12bp.

Here, N is the length of the CGI, CpGNum, CNum and GNum represent the number of CpG, number of C, number of G respectively.

(4) Methylation level annotation: Eq. 5 classifies methylation ratio into three levels with respect to the definition (Ziller et al., 2013).

$$\mathbf{Methylation\ level}(chr, p) = \begin{cases} \mathbf{1, highly\ mehylated} & \mathbf{methylation\_ratio}(chr, p) > 0.75 \\ \mathbf{2, unmethylated} & \mathbf{methylation\_ratio}(chr, p) < 0.1 \\ \mathbf{3, moderate\ methylated} & \mathbf{otherwise} \end{cases} \tag{5}$$

Here, chr and p represent the chromosome and position of a CpG site, respectively.

## 3 Results

### 3.1 Human position-defined CGI prediction method

Indicated by previous study (Hackenberg et al., 2006), we consider CGIs as potentially functionally short islands (CpG islets) if length of CGIs is less than 200bp. Here, Figure 2A demonstrates that Human position-defined CGI prediction method not only can locate the shortest average length (23.7bp) under CpG interval d = 12bp, but also the percentage of CGIs <200bp for Human position-defined CGI prediction method are greater than both CpGCluster method (Hackenberg et al., 2006) and density-defined CGI prediction method (Weber et al., 2007; Zhang et al., 2021a).

Also, since proportion distribution of CGI features is closely related to the regulatory mechanisms for CGI methylation (Hackenberg et al., 2010), Human position-defined CGI prediction method can describe the proportional distribution of the predicted CGIs at different CGI length (Figure 2B), GC content (Figure 2C), and O/E (Figure 2D). Here, we employ default setup for CpG interval, d = 12bp (Zhang et al., 2018; Zhang et al., 2021a).

It should be noted that Human position-defined CGI prediction method can parallel carry out position-defined CGI prediction and comparative analysis for multiple CpG intervals (d) by MR-CpGCluster.

### 3.2 Data annotation

Data annotation is described by the right side of Figure 1. Firstly, the position-defined CGIs are classified into different densities by Supplementary Eq. S2. And then, we classify each CpG methylation site of CGIs into different gene functional regions by Supplementary Table S2. Lastly, we classify the CpG sites into three methylation levels by Eq.5.

Data annotation can help us investigate the distribution of all CpG sites in different structural and functional categories of genome sequences (Figure 3; Supplementary Table S3). For example, we not only can compare the distribution of the number of CpG sites in each region of the predicted CGIs under different CpG interval(d) (Figure 3A), but also visualize the density of CpG sites in different functional regions (Figure 3B).

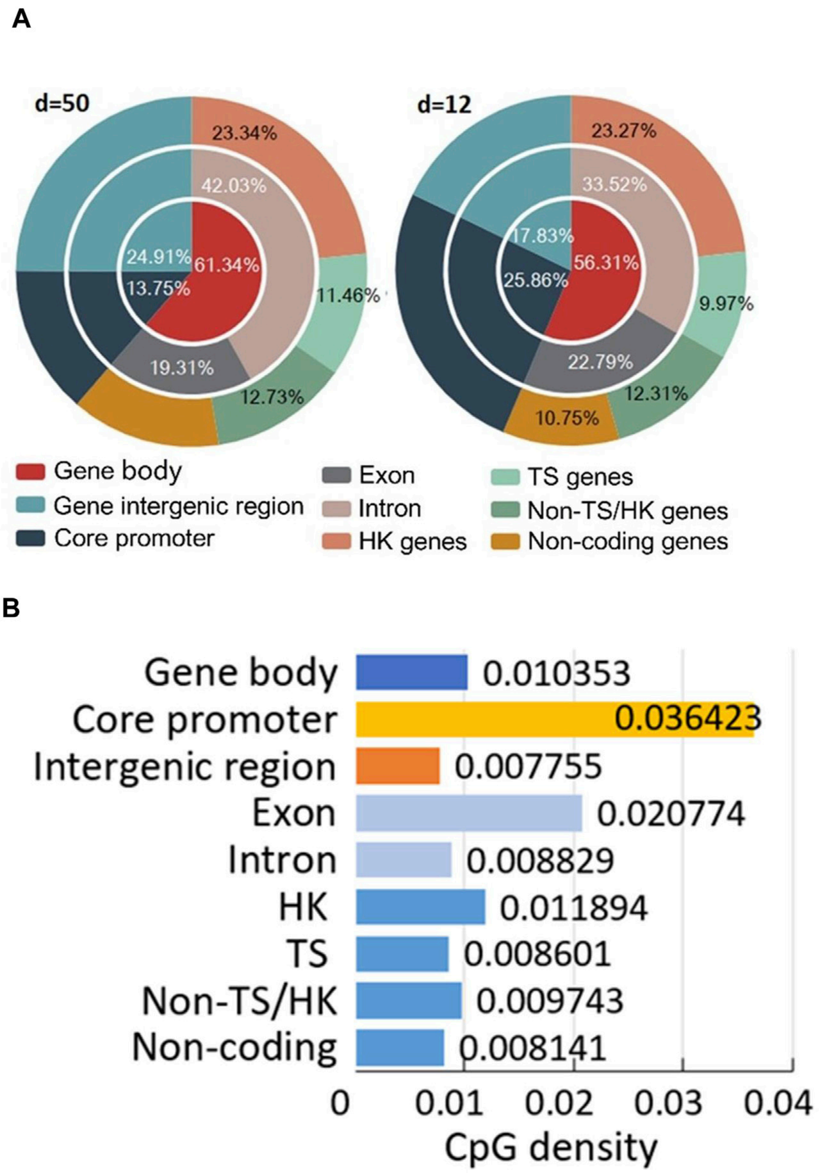


FIGURE 3
Position-defined CGI annotation results. **(A)** Distribution of all CpG sites in different structural and functional categories of genome sequences. **(B)** CpG density (Eq. 4) of different gene and sequence categories.

## 3.3 Human position-defined CGI methylation analysis

The position-defined CGI methylation analysis is described by the bottom side of Figure 1 with three functions.

First is "CGI density analysis" (Figure 4A), which is used to analyze the classification of position-defined CGI under various CGI density (Weber et al., 2007; Zhu et al., 2008) and CpG interval (d).

Second is "Methylation level analysis," which not only can analyze the specificity of methylation level for CpG sites under different annotation categories and CpG interval (d) (Figure 4B), but also allows the visualization and comparative analysis of methylation level of position-defined CGIs at the genome-wide perspective (Figure 4C).

The third is "GO enrichment analysis," which employs clusterProfiler (Yu et al., 2012) to make GO enrichment analysis (Liu et al., 2020) for the CGI + genes (Coding genes that at least one of its TSSs is located in the CGI) (Weber et al., 2007; Zhang et al., 2021a) of position-defined CGIs. Here, Figure 4D shows GO enrichment analysis for the CGI + genes under CpG interval d = 12bp.

## 3.4 Algorithm performance comparison

Firstly, As shown in Figure 5; Supplementary Figure S2, we compare the computing speed for Human position-defined CGI prediction method with MR-CpGCluster and this method without MR-CpGCluster with three commonly used

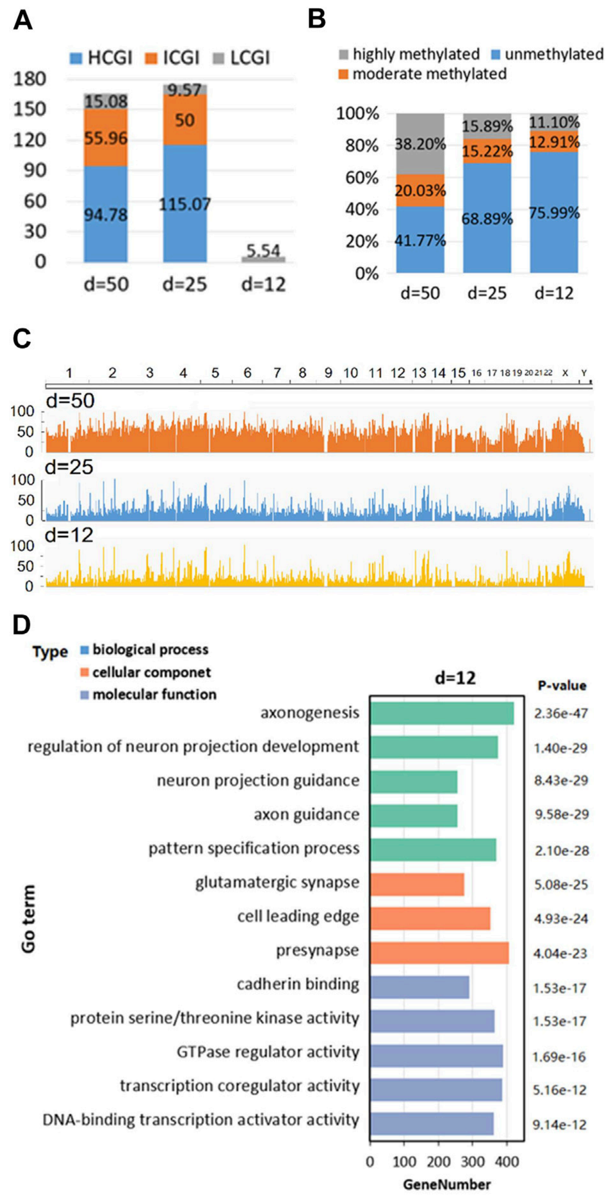


FIGURE 4
Human position-defined CGI methylation analysis. **(A)** The classification of position-defined CGI under CGI density. **(B)** CpG density of different gene and sequence categories. **(C)** Comparison of the methylation ratio of position-defined CGIs. The horizontal and vertical axes represent genomic chromosome position and the methylation rate of the CpG site at corresponding position, respectively. **(D)** GO enrichment analysis.

standards: Speedup, Scaleup and Sizeup (Schatz, 2009). Figure 5 shows that the Speedup is positively related to the number of nodes and the size of dataset. For example, when using 8 nodes for a 3920 MB dataset, the ratio between the actual and ideal Speedup is 6.00/8 = 75%, while with 6 nodes for a 980 MB dataset, this ratio is 2.88/6 = 49.17%.

Next, we compare the computing efficiency for Human position-defined CGI prediction method with commonly used density-defined CGIs prediction method (Weber et al., 2007; Zhang et al., 2021a) and another two classical distance-based CGI prediction methods such as WordCluster (Hackenberg et al., 2011) and CpGProD (Ponger and Mouchiroud, 2002) by CGI length, GC content, and O/E ratio (Eq. 3), which are three broadly used standards (Wang and Leung, 2004; Hackenberg et al., 2010).

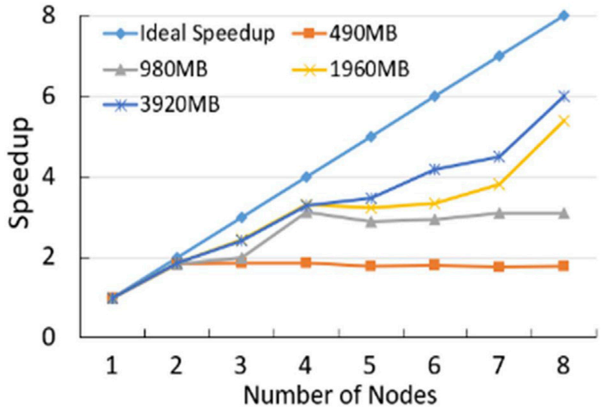


FIGURE 5
The speedup ratio of MR-CpGCluster.

Table 1; Supplementary Figure S5 not only demonstrate that the average length of CGIs of Human position-defined CGI prediction method (23.7 ± 11.5bp) is statistically shorter, but also the average GC content (89.3% ± 7.5%) and O/E value (1.54 ± 0.27) of Human position-defined CGI prediction method are statistically greater than other prediction methods by statistical test (Zhang et al., 2021b; Zhang et al., 2021d; Gao et al., 2021; Liu et al., 2021; Lai et al., 2022; Song et al., 2022).

Note: Here, we employ default setup for CpG interval, d = 12bp (Zhang et al., 2018; Zhang et al., 2021a).

## 3.5 Web service construction

Figure 6 shows the technical architecture of PCGIMA (http://www.combio-lezhang.online/pcgima/home.html), which consists of three modules: “Human position-defined CGI prediction,” “CpG sites annotation analysis,” and “CGI methylation analysis.”

PCGIMA employs MR-CpGCluster to predict the position-defined CGI for multiple consecutive (d) values. To compare and analyze the CpG methylation levels in different genome regions, we integrate the JavaScript version of IGV (Integrative Genomics Viewer) (Thorvaldsdottir et al., 2013) into our Web service. PCGIMA also imports the genome annotation information and analysis results into the MySQL database (Xia et al., 2010) and use eCharts (Bond and Goguen, 2002) to visualize CGI-related analysis results.

“Human position-defined CGI prediction” module provides two functions (Figure 2). One is “Position-defined CGI prediction,” which can online predict position-defined CGI for the human genome or a particular chromosome with multiple consecutive (d) values. The other is “Position-

TABLE 1 CGI prediction methods comparison.

| CGI prediction methods | CGI number | Average length ± standard deviation | Average GC ± standard deviation | Average O/E ± standard deviation | Average CpG Density ± standard deviation |
|---|---|---|---|---|---|
| Human position-defined CGI prediction method | 89,063 | 23.7 ± 11.5 | 89.3% ± 7.5% | 1.54 ± 0.27 | 0.294 ± 0.066 |
| CpGCluster | 198,445 | 274.7 ± 249.8 | 63.8% ± 7.6% | 0.86 ± 0.27 | 0.087 ± 0.04 |
| WordCluster | 198,703 | 273.2 ± 246.4 | 63.8% ± 7.5% | 0.86 ± 0.27 | 0.087 ± 0.04 |
| CpGProD | 76,793 | 1,043.8 ± 761.7 | 54.6% ± 6.1 | 0.64 ± 0.1 | 0.047 ± 0.016 |
| Density-defined CGIs | 30,477 | 774.7 ± 826.9 | 66.5% ± 4.7% | 0.86 ± 0.14 | 0.094 ± 0.018 |

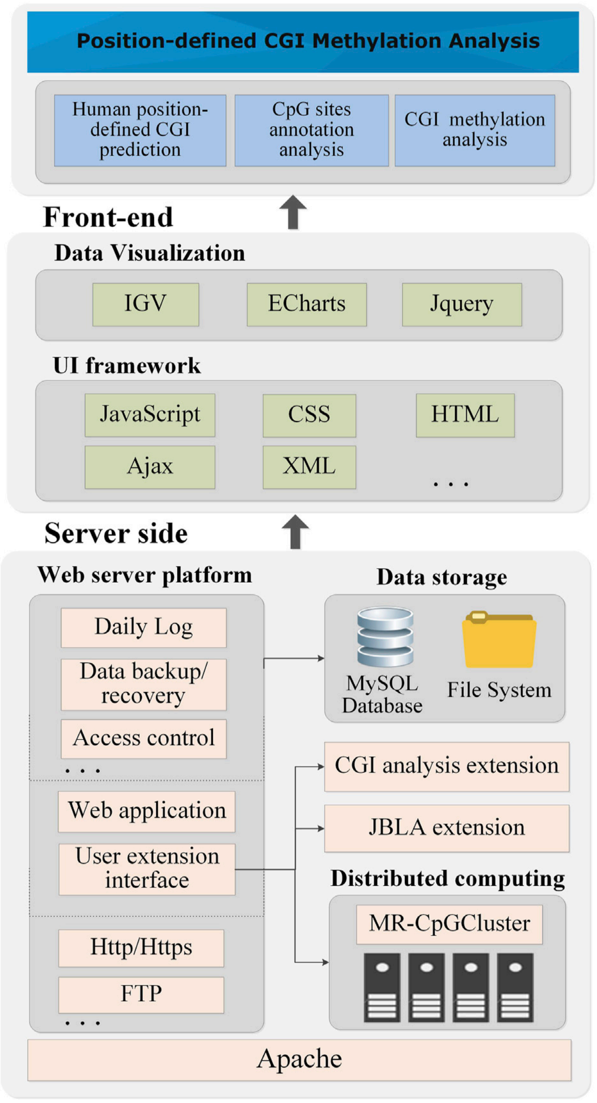


FIGURE 6
The technical architecture of PCGIMA.

defined CGI features analysis," which can describe the connection between the proportion distribution of CGI and CGI features.

"CpG sites annotation analysis" module consists of two functions. First is "Human CpG sites Distribution analysis," which can analyze the distribution of CpG methylation sites in different structural and functional categories of genomic sequences (Figure 3). Second is "Human CpG sites permutation analysis" module (Supplementary Method S4), which can analyze the CpG permutation patterns (Zhang et al., 2018) of density- and position-defined CGIs.

"CGI methylation analysis" module also provides two functions. One is "Position-defined CGI methylation analysis," which can analyze the specificity of methylation level for CpG sites under different annotation categories (Figure 4B). The other is "GO enrichment analysis," which can make GO enrichment analysis for the CGI + genes of position-defined CGIs (Figure 4D). Meanwhile, PCGIMA also provides related source code and data download services. The function descriptions are detailed in Supplementary Method S4.

## 4 Discussion and conclusion

This study not only develops Human position-defined CGI prediction method to locate short and position-sensitive CGIs (CpG islets) using high performance computing to construct MR-CpGCluster algorithm (Figure 1), but also a novel human genome annotation and analysis method to investigate the connections among CGI, gene expression and methylation. Finally, we integrate them into PCGIMA for online computing and visualization.

For Human position-defined CGI prediction method, it not only can efficiently locate CpG islets (Figure 2A; Table 1), but also it can parallel predict position-defined CGIs with multiple consecutive (d) values and investigate the genetic characteristics of position-defined CGIs under different CpG interval parameters (Figures 2B–D; Supplementary Datas S1–S3).

For annotation method, it can investigate the connections between position-defined CGI methylation and gene expression specificity from a genome-wide perspective by considering functional regions (core promoters and gene bodies) and the distribution of methylation sites of genes for different expression breadth (Figure 3). Our annotation method (Figure 3A) reveals that the distribution proportion of methylation sites in TS genes for short positional-defined CGIs ($d = 12$) is 9.97%, which is less than that for long positional-defined CGIs ($d = 50$, 11.46%).

For Human position-defined CGI methylation analysis, not only CGI density analysis (Figure 4A) finds an interesting phenomena that short position-defined CGIs (CpG islets) are closer to LCGI by classifying the position-defined CGI under various CGI density (Weber et al., 2007; Zhu et al., 2008) and CpG interval (d), but also methylation levels analysis demonstrates that the average methylation levels are obviously low for CpG islets from overall scale and genome-wide perspective, respectively (Figures 4B, C) as well as Go enrichment analysis implies that the position-defined CGI-related genes could be associated with unique gene regulatory functions (Figure 4D; Supplementary Figure S4).

For Algorithm performance comparison, Figure 5 turns out that MR-CpGCluster method is faster than classical CpGCluster for the big dataset, which implies Human position-defined CGI prediction method can parallel process the big CGI data.

Moreover, previous studies indicate that CGIs with length less than 200 bp may be functional CpG islets (Hackenberg et al., 2006) and high GC content and O/E values represent enrichment of methylation sites (Gardinergarden and Frommer, 1987; Takai and Jones, 2002). Since Table 1 demonstrates that the average CGI length of the Human position-defined CGI prediction method is much less than 200bp (column 3 of Table 1), and the average GC content and O/E value are statistically greater than other prediction methods (column 4 and 5 of Table 1), we can conclude that Human position-defined CGI prediction method can locate more potential short CGIs with special functions than previous CGI prediction methods (Takai and Jones, 2002; Takahashi et al., 2017).

Lastly, Figure 6 shows that since we utilize the MR-CpGCluster to speed up CGI prediction and incorporate extensive visualization methods to increase user usability, PCGIMA provides an easy-to-use analysis and visualization platform for human CGI prediction and methylation. It should be noted that since the human genome annotation and analysis results have been computed and imported into the database in advance, it is fast (about 2–3 min) for PCGIMA to show the analysis results except the “Human position-defined CGI prediction.”

Although our study already made great progress in CGI prediction, annotation, analysis, and visualization, it still needs further improving. Firstly, we should make detail annotations for human position-defined CGIs in terms of functional and structural features. Secondly, we should interrogate the lineage-based and function-based subsets for CGIs and their regulatory implications (Blackledge et al., 2013). Finally, we should employ advanced high performance computing technology (Jiang et al., 2015; Zhang et al., 2021c; Xiao et al., 2021) to improve PCGIMA in the distant future.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

## Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants’ legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

## Author contributions

MX: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Visualization, Writing–original draft, Writing–review and editing. YX: Writing–original draft, Writing–review and editing, Formal Analysis, Visualization, Data curation, Methodology, Resources. JY: Writing–original draft, Writing–review and editing, Conceptualization, Investigation, Methodology, Project administration, Supervision. LZ: Writing–original draft, Writing–review and editing, Formal Analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by grants from National Science and Technology Major Project (2021YFF1201200, China), National Science and Technology Major Project (2018ZX10201002, China), National Natural Science Foundation of China (Grant No. 62372316, China), China Postdoctoral Science Foundation (2020M673221, China), Fundamental Research Funds for the Central Universities (2020SCU12056, China), Sichuan Science and Technology Program (2022YFS0048), Chongqing Technology Innovation and Application Development Project (CSTB2022TIAD-KPX0067).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1367731/full#supplementary-material

# References

Acton, R., Yuan, W., Gao, F., Xia, Y., Bourne, E., Wozniak, E., et al. (2021). The genomic loci of specific human tRNA genes exhibit ageing-related DNA hypermethylation. *Nat. Commun.* 12, 2655. doi:10.1038/s41467-021-22639-6

Blackledge, N. P., Thomson, J. P., and Skene, P. J. (2013). CpG island chromatin is shaped by recruitment of ZF-CxxC proteins. *Cold Spring Harb. Perspect. Biol.* 5 (11), a018648. doi:10.1101/cshperspect.a018648

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG island mapping by epigenome prediction. *PLOS Comput. Biol.* 3 (6), e110. doi:10.1371/journal.pcbi.0030110

Bond, G. W., and Goguen, H. (2002). "ECharts: balancing design and implementation", in: Proceedings of the 6 th IASTED International Conference on Software Engineering and Applications, 149–155.

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., et al. (2018). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* 46, D762–D769. Database issue). doi:10.1093/nar/gkx1020

Clark, K., Karschmizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). *GenBank. Nucleic Acids Res.* 44, D67–D72. Database issue). doi:10.1093/nar/gkv1276

Dede, E., Sendir, B., Kuzlu, P., Weachock, J., Govindaraju, M., and Ramakrishnan, L. (2016). Processing cassandra datasets with hadoop-streaming based approaches. *IEEE Trans. Serv. Comput.* 9 (1), 46–58. doi:10.1109/tsc.2015.2444838

Di, l., Linna, Z., Zhaoyang, W., Xu, Z., Xiuzhao, F., Yong, L., et al. (2018). EWASdb: epigenome-wide association study database. *Nucleic Acids Res.* D1, D1. doi:10.1093/nar/gky942

Dittrich, J., and Quiané-Ruiz, J. A. (2012). Efficient big data processing in Hadoop MapReduce. *Proc. Vldb Endow.* 5 (12), 2014–2015. doi:10.14778/2367502.2367562

Dor, Y., and Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. *Lancet* 392 (10149), 777–786. doi:10.1016/s0140-6736(18)31268-6

El-Maarri, O., Olek, A., Balaban, B., Montag, M., van der Ven, H., Urman, B., et al. (2019). Methylation levels at selected CpG sites in the factor VIII and FGFR3 genes, in mature female and male germ cells: implications for male-driven evolution. *Am. J. Hum. Genet.* 63 (4), 1001–1008. doi:10.1086/302065

Gao, J., Liu, P., Liu, G. D., and Zhang, L. (2021). Robust needle localization and enhancement algorithm for ultrasound by deep learning and beam steering methods. *J. Comput. Sci. Technol.* 36 (2), 334–346. doi:10.1007/s11390-021-0861-7

Gardinergarden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* 196 (2), 261–282. doi:10.1016/0022-2836(87)90689-9

Hackenberg, M., Barturen, G., Carpena, P., Luque-Escamilla, P., Previti, C., and Oliver, J. (2010). Prediction of CpG-island function: CpG clustering vs sliding-window methods. *BMC genomics* 11, 327. doi:10.1186/1471-2164-11-327

Hackenberg, M., Carpena, P., Bernaola-Galván, P., Barturen, G., Alganza, Á. M., and Oliver, J. L. (2011). WordCluster: detecting clusters of DNA words and genomic elements. *Algorithms Mol. Biol.* 6 (1), 2. doi:10.1186/1748-7188-6-2

Hackenberg, M., Previti, C., Luqueescamilla, P. L., Carpena, P., Martínezaroza, J., and Oliver, J. L. (2006). CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinforma.* 7 (1), 446. doi:10.1186/1471-2105-7-446

Han, L., Su, B., Li, W. H., and Zhao, Z. (2008). CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol.* 9 (5), R79–R12. doi:10.1186/gb-2008-9-5-r79

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., et al. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 7 (Suppl. 1), 1–9. doi:10.1186/gb-2006-7-s1-s4

Jiang, B., Dai, W., Khaliq, A., Zhou, X., and Zhang, L. (2015). Novel 3D GPU based numerical parallel diffusion algorithms in cylindrical coordinates for health care simulation. *Math. Comput. Simul.* 109, 1–19. doi:10.1016/j.matcom.2014.07.003

Lai, X., Zhou, J., Wessely, A., Heppt, M., Maier, A., Berking, C., et al. (2022). A disease network-based deep learning approach for characterizing melanoma. *Int. J. Cancer* 150 (6), 1029–1044. doi:10.1002/ijc.33860

Liu, B., Du, Q., Chen, L., Fu, G., Li, S., Fu, L., et al. (2016). CpG methylation patterns of human mitochondrial DNA. *Sci. Rep.* 6 (1), 23421. doi:10.1038/srep23421

Liu, G. D., Li, Y. C., Zhang, W., and Zhang, L. (2020). A brief review of artificial intelligence applications and algorithms for psychiatric disorders. *Engineering* 6 (4), 462–467. doi:10.1016/j.eng.2019.06.008

Liu, S., You, Y., Tong, Z., and Zhang, L. (2021). Developing an embedding, koopman and autoencoder technologies-based multi-omics time series predictive model (EKATP) for systems biology research. *Front. Genet.* 12, 761629. doi:10.3389/fgene.2021.761629

Ming, X., Zhu, B., and Li, Y. (2021). Mitotic inheritance of DNA methylation: more than just copy and paste. *J Genet Genomics.* 48 (1), 1–13. doi:10.1016/j.jgg.2021.01.006

Ning, Y., Xuan, G., Alexander, Z., and Pan, Y. (2017). GaussianCpG: a Gaussian model for detection of CpG island in human genome sequences. *BMC Genomics* 18 (S4), 392. doi:10.1186/s12864-017-3731-5

Ponger, L. C., and Mouchiroud, D. (2002). CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* 18 (4), 631–633. doi:10.1093/bioinformatics/18.4.631

Pruitt, K. D., Tatiana, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33, D501–D504. Database issue. doi:10.1093/nar/gki025

Raney, B. J., Cline, M. S., Rosenbloom, K. R., Dreszer, T. R., Katrina, L., Barber, G. P., et al. (2010). ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.* 39, D871–D875. doi:10.1093/nar/gkq1017

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447 (7143), 425–432. doi:10.1038/nature05918

Schatz, M. C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* 25 (11), 1363–1369. doi:10.1093/bioinformatics/btp236

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., et al. (2017). Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 27, 849–864. doi:10.1101/gr.213611.116

Smith, Z. D., Chan, M. M., Mikkelsen, T. S., Gu, H., Gnirke, A., Regev, A., et al. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* 484 (7394), 339–344. doi:10.1038/nature10960

Song, H., Chen, L., Cui, Y., Li, Q., Wang, Q., Fan, J., et al. (2022). Denoising of MR and CT images using cascaded multi-supervision convolutional neural networks with progressive training. *Neurocomputing* 469, 354–365. doi:10.1016/j.neucom.2020.10.118

Su, J., Zhang, Y., Lv, J., Liu, H., Tang, X., Wang, F., et al. (2009). CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. *Nucleic acids Res.* 38, e6. doi:10.1093/nar/gkp882

Takahashi, Y., Wu, J., Suzuki, K., Martinez-Redondo, P., Li, M., Liao, H. K., et al. (2017). Integration of CpG-free DNA induces *de novo* methylation of CpG islands in pluripotent stem cells. *Science* 356 (6337), 503–508. doi:10.1126/science.aag3260

Takai, D., and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. U. S. A.* 99 (6), 3740–3745. doi:10.1073/pnas.052410099

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform* 14 (2), 178–192. doi:10.1093/bib/bbs017

Wang, Y., and Leung, F. C. C. (2004). An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics* 20 (7), 1170–1177. doi:10.1093/bioinformatics/bth059

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., et al. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* 39 (4), 457–466. doi:10.1038/ng1990

Wright, J. C., Mudge, J., Weisser, H., Barzine, M. P., Gonzalez, J. M., Brazma, A., et al. (2016). Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat. Commun.* 7, 11778. doi:10.1038/ncomms11778

Xia, X. Q., McClelland, M., and Wang, Y. (2010). TabSQL: a MySQL tool to facilitate mapping user data to public databases. *BMC Bioinforma.* 11, 342. doi:10.1186/1471-2105-11-342

Xiao, M., Liu, G., Xie, J., Dai, Z., Wei, Z., Ren, Z., et al. (2021). 2019nCoVAS: developing the web service for epidemic transmission prediction, genome analysis, and psychological stress assessment for 2019-nCoV. *IEEE/ACM Trans. Comput. Biol. Bioinform* 18 (4), 1250–1261. doi:10.1109/TCBB.2021.3049617

Xiong, Z., Li, M., Yang, F., Ma, Y., Sang, J., Li, R., et al. (2019). EWAS Data Hub: a resource of DNA methylation array data and metadata. *Nucleic acids Res.* 48, D890–D895. doi:10.1093/nar/gkz840

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R Package for comparing biological Themes among gene clusters. *Omics-a J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, L., Dai, Z., Yu, J., and Xiao, M. (2021a). CpG-island-based annotation and analysis of human housekeeping genes. *Briefings Bioinforma.* 22 (1), 515–525. doi:10.1093/bib/bbz134

Zhang, L., Liu, G., Kong, M., Li, T., Wu, D., Zhou, X., et al. (2021b). Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model. *Bioinformatics* 37 (11), 1554–1561. doi:10.1093/bioinformatics/btz542

Zhang, L., Xiao, M., Zhou, J., and Yu, J. (2018). Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a Jellyfish-based LAUPs analysis application (JBLA). *Bioinformatics* 34 (21), 3624–3630. doi:10.1093/bioinformatics/bty392

Zhang, L., Zhang, L., Guo, Y., Xiao, M., Feng, L., Yang, C., et al. (2021c). MCDB: a comprehensive curated mitotic catastrophe database for retrieval, protein sequence alignment, and target prediction. *Acta Pharm. Sin. B* 11 (10), 3092–3104. doi:10.1016/j.apsb.2021.05.032

Zhang, L., Zhao, J., Bi, H., Yang, X., Zhang, Z., Su, Y., et al. (2021d). Bioinformatic analysis of chromatin organization and biased expression of duplicated genes between two poplars with a common whole-genome duplication. *Hortic. Res.* 8 (1), 62. doi:10.1038/s41438-021-00494-2

Zhu, J., He, F., Song, S., Wang, J., and Yu, J. (2008). How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* 9 (1), 172. doi:10.1186/1471-2164-9-172

Ziller, M. J., Gu, H., Muller, F., Donaghey, J., Tsai, L. T., Kohlbacher, O., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500 (7463), 477–481. doi:10.1038/nature12433

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 08 April 2024
DOI 10.3389/fdata.2024.1346958

Check for updates

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Milan Deepak Parmar,
The University of Tennessee, United States
Xuming Han,
Jinan University, China
Hongwei Ge,
Dalian University of Technology, China

*CORRESPONDENCE
Renchu Guan
✉ guanrenchu@jlu.edu.cn

RECEIVED 05 December 2023
ACCEPTED 25 March 2024
PUBLISHED 08 April 2024

CITATION
Li X, Han X, Wei S, Liang Y and Guan R (2024) Acupuncture and tuina knowledge graph with prompt learning. *Front. Big Data* 7:1346958. doi: 10.3389/fdata.2024.1346958

COPYRIGHT
© 2024 Li, Han, Wei, Liang and Guan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Acupuncture and tuina knowledge graph with prompt learning

Xiaoran Li[1], Xiaosong Han[1], Siqing Wei[1], Yanchun Liang[1,2] and Renchu Guan[1]*

[1]Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, College of Computer Science and Technology Jilin University, Changchun, China, [2]Zhuhai Laboratory of Key Laboratory for Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology of Zhuhai College of Science and Technology, Zhuhai, China

**Introduction:** Acupuncture and tuina, acknowledged as ancient and highly efficacious therapeutic modalities within the domain of Traditional Chinese Medicine (TCM), have provided pragmatic treatment pathways for numerous patients. To address the problems of ambiguity in the concept of Traditional Chinese Medicine (TCM) acupuncture and tuina treatment protocols, the lack of accurate quantitative assessment of treatment protocols, and the diversity of TCM systems, we have established a map-filling technique for modern literature to achieve personalized medical recommendations.

**Methods:** (1) Extensive acupuncture and tuina data were collected, analyzed, and processed to establish a concise TCM domain knowledge base. (2)A template-free Chinese text NER joint training method (TemplateFC) was proposed, which enhances the EntLM model with BiLSTM and CRF layers. Appropriate rules were set for ERE. (3) A comprehensive knowledge graph comprising 10,346 entities and 40,919 relationships was constructed based on modern literature.

**Results:** A robust TCM KG with a wide range of entities and relationships was created. The template-free joint training approach significantly improved NER accuracy, especially in Chinese text, addressing issues related to entity identification and tokenization differences. The KG provided valuable insights into acupuncture and tuina, facilitating efficient information retrieval and personalized treatment recommendations.

**Discussion:** The integration of KGs in TCM research is essential for advancing diagnostics and interventions. Challenges in NER and ERE were effectively tackled using hybrid approaches and innovative techniques. The comprehensive TCM KG our built contributes to bridging the gap in TCM knowledge and serves as a valuable resource for specialists and non-specialists alike.

KEYWORDS

**prompt learning, knowledge graph, Traditional Chinese Medicine, Named Entity Recognition, Entity Relationship Extract**

## 1 Introduction

Acupuncture and Tuina stand as foundational therapies within Traditional Chinese Medicine (TCM), boasting centuries of esteemed practice and theoretical development. Their origins trace back to ancient China, where detailed theories and practical methodologies were documented in medical classics like the Huangdi Neijing over two millennia ago (Unschuld, 2016). Since the Neolithic era, these practices have gradually become integrated into human life (Cao, 2008) and are widely employed in clinical treatments due to their simplicity and rapid effectiveness. In recent years, acupuncture and tuina have gained increased recognition in various countries, including the

United States and Europe. In 2020, a research article featured in the journal Neuron, authored by a team led by Professor Qiufu Ma from Harvard Medical School, demonstrates that acupuncture, through targeted stimulation of specific acupoints on the body's surface, can activate a range of "somatic-sensory-autonomic-target-organ" reflex pathways. This activation is capable of effecting immune-inflammatory modulation (Liu et al., 2020). In 2021, the team led by Qiufu Ma further affirmed the scientific validity of acupuncture treatments by shedding light on the neuroanatomy involved in acupuncture's activation of specific signaling pathways (Liu et al., 2021). Moreover, the World Health Organization (WHO) acknowledges acupuncture's efficacy in addressing a wide range of conditions, including but not limited to chronic pain, nausea, and certain neurological disorders. Such recognition further underscores the significance and widespread applicability of acupuncture within contemporary medical practices. In the era of the Internet's evolution, deep learning's pervasive use in daily life has become evident. Its applications span essay recommendations to text data analysis, serving as an essential component (Wang et al., 2018; Guan et al., 2020). The integration of deep learning into the medical domain is equally noteworthy, synergizing neural networks and medical technology to advance science and technology. The surge in regional health informatization and medical technology has amassed substantial medical data. Extracting and effectively utilizing this information is vital for intelligent medical support (Hou et al., 2018). Knowledge graph (KG) technology, as an emerging technology for information organization and processing, demonstrates the ability to efficiently integrate and analyze vast amounts of data and information. By systematically consolidating both ancient and modern literature into a knowledge base, knowledge graphs offer decision-making support for medical practitioners. In the realm of TCM, this technology presents new opportunities for development.

The construction of knowledge graphs is a fundamental concern within the field of knowledge graph research (Liu et al., 2016). The knowledge graph architecture encompasses both its logical structure and technical components. In this paper, we concentrate on two vital aspects of knowledge graph construction: the technical architecture, specifically entity extraction, and relationship extraction.

Nevertheless, constructing high-quality medical knowledge graphs faces significant challenges, particularly due to the distinct characteristics and requirements of medical terminology compared to common terms. Such endeavors often demand substantial human and material resources. In particular, the development of Traditional Chinese Medicine (TCM) knowledge graphs has received relatively less research attention compared to general medical knowledge graphs. Furthermore, detailed research specific to each branch of TCM remains limited.

In the field of acupuncture and tuina, there are still many problems.

- Many acupuncture and tuina treatment plans have many ambiguous concepts, such as the location of acupuncture, the strength of tuina, etc.
- There is a lack of specific quantitative assessment in the implementation of treatment protocols.
- The extant TCM systems are heterogeneous. In the case of acupuncture alone, 26 mainstream schools exist (Zhang and Xia, 2018).

All of the above problems pose a major obstacle to the development of personalized treatment plans. In our previous work, a knowledge graph of acupuncture and tuina was constructed using ancient Chinese medical literature, but there is still no graph filling technique for modern literature (Han et al., 2021).

In the context of Prompt Tuning for Few-shot Named Entity Recognition (NER), the majority of existing methods have primarily focused on English text, resulting in a relative scarcity of methods specifically designed for Chinese text. Chinese sentences tend to be longer compared to English sentences, often spanning tens or even hundreds of Chinese characters. This increased sentence length significantly expands the search space for templates and poses challenges in finding suitable templates for Chinese text. Moreover, Chinese entities typically consist of at least two characters, and in domains like Chinese medicine, entities can even comprise nearly ten words. The diverse meanings of Chinese characters introduce the possibility of unintended partially nested entities within longer entities. All this makes the previous methods do not fit well in Chinese texts.

In order to solve the above problems, in this paper, we hope to accomplish the structured storage and retrieval of acupuncture and tuina knowledge in modern literature with the help of knowledge graphs, and at the same time provide a basis for personalized medical solution recommendation. To summarize the contribution of this work:

- We have collected a large amount of knowledge related to the field of acupuncture and tuina and have built a small domain knowledge base based on this knowledge. It contributes to the construction of the KG later, and also facilitates the needs of other researchers.
- The development of the KG revolves around two primary facets: NER and ERE. For ERE, we opted for the conventional rule-based method for relationship extraction. In contrast, for NER, we devised a hybrid entity extraction model that combines Trie tree-based techniques with deep learning. In particular, we proposed the TemplateFC model, which becomes a more adaptable template-free prompt tuning method for Chinese text by adding BiLSTM layer and CRF layer for joint training.
- We built a knowledge graph of acupuncture and tuina based on modern literature. This graph encompasses 10,346 entities and 440,919 relationships. Additionally, we have developed a user-friendly entity query interface and UI, enhancing accessibility and usability.

The study of this paper is shown in Figure 1.

## 2 Related work

### 2.1 Named Entity Recognition (NER)

Entity extraction, also known as Named Entity Recognition (NER), holds significant importance in the field of natural language

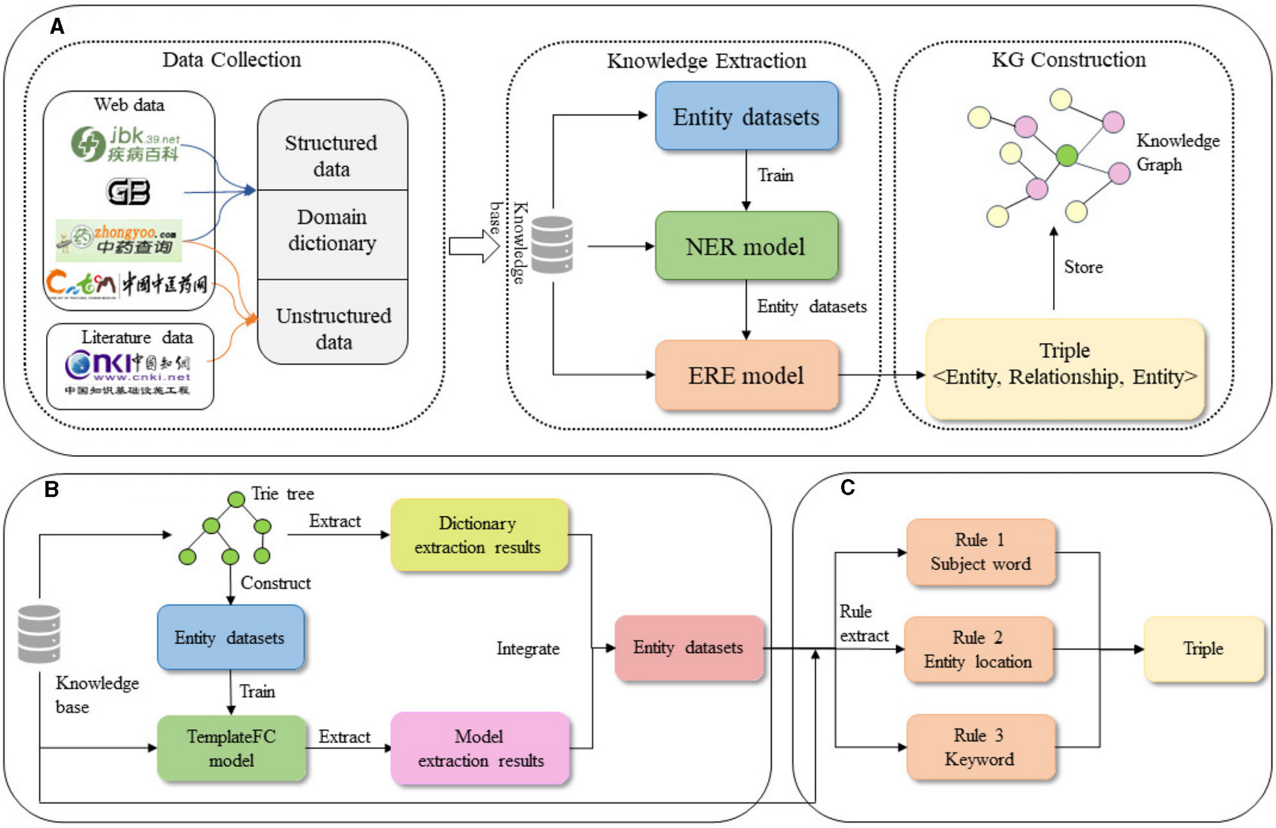


FIGURE 1
Flow chart of knowledge graph construction. **(A)** The flow of knowledge graph construction. **(B)** Entity extraction. **(C)** Relationship extraction.

processing. Currently, two primary architectural approaches are prevalent in NER research: Neural Network-Conditional Random Field (NN-CRF) and sliding windows. In the NN-CRF architecture, a sequence labeling model, such as CRF, is employed, utilizing a neural network for text representation and feature extraction. Dai et al. (2019) utilized the BERT-BiLSTM-CRF architecture to successfully perform entity extraction for Chinese medical records. Conversely, the sliding window architecture treats NER as a classification task, identifying and classifying potential entities within a sentence. Eberts and Ulges (2019) employed syntactic features to classify Spans created by splitting the sentence. In practical industrial applications, NER techniques often employ hybrid models that combine multiple methods to achieve improved results. Meituan technicians, for instance, employed a combination of lexicon and deep learning models for entity recognition, as documented in the literature. In this paper, we also employ a similar hybrid model.[1]

## 2.2 Prompt tuning for few-shot NER

Prompt learning refers to changing the downstream task to the text generation task by adding some hints to the input of the model without significantly changing the structure and parameters of the pre-trained language model. This approach has gained popularity, particularly in low-resource scenarios of Named Entity Recognition (NER). Numerous authors have proposed their own ideas during this period, leveraging templates tag words, and other techniques. For instance, Lee et al. (2021) focused on selecting specific high-quality examples, while Cui et al. (2021) explored the process of identifying suitable templates. Additionally, Chen et al. (2023) introduced the concept of weighted averaging of multiple prompts using mask-reducible prompts. These advancements have contributed to establishing a strong correlation between the quality of search results and the output of prompt learning techniques.

## 2.3 Entity Relationship Extract (ERE)

Entity Relationship Extraction (ERE) serves as the successor task to entity extraction. Traditional approaches primarily rely on rule-based methods, which identify relationship expressions by detecting specific patterns and using regular expressions. Conversely, deep learning-based methods for relationship extraction closely resemble entity extraction and necessitate a substantial amount of annotated data for supervision. Liu and Zhao (2020) proposed a neural network architecture namely BERT-CNN-BiLSTM-CRE, to achieve relationship extraction in

---

[1] Exploration and Practice of NER Technology in Meituan Search. https://tech.meituan.com/2020/07/23/ner-in-meituan-nlp.html. Acessed:2024-01-10.

the medical domain. Their architecture was successfully applied to extract relationships from a medical corpus. Moscato et al. (2023) used three biomedical datasets and a multi-task learning framework for relationship extraction. Milošević and Thielemann (2023) have developed a new rule-based method for relationship extraction. The method relies on vocabularies for relationship trigger words, negation cues, speculation cues, mode of action (MoA) cues, and grammar pattern rule set.

## 2.4 Knowledge graph of TCM

The construction of the Traditional Chinese Medicine (TCM) knowledge graph has attracted the involvement of numerous domestic and international internet companies. Yu et al. leveraged the linguistic system of TCM as a foundation and integrated a series of TCM-related databases to develop a comprehensive TCM knowledge graph. This knowledge graph was subsequently embedded and utilized in a TCM knowledge service platform. Furthermore, Yu et al. (2015) capitalized on the digital resources accumulated in the field of TCM to construct a knowledge graph specifically tailored to TCM healthcare. Currently, one of the most notable TCM knowledge graphs is developed by the Institute of Traditional Chinese Medicine Information at the Chinese Academy of Traditional Chinese Medicine. This knowledge graph encompasses 127 semantic types and 58 semantic relationships, establishing it as a valuable resource in the field (Cui et al., 2014). The global acupuncture clinical trial research is booming, and clinical evidence for acupuncture is emerging. Nenggui Xu's team applied artificial intelligence analysis technology to complete the "linking" of original research and 332 systematic evaluations of evidence in 20 disease areas, comprehensively improved the clinical evidence matrix of acupuncture therapy in the Epistemonikos database, and formulated the world's first clinical evidence atlas for acupuncture (Lu et al., 2022).

# 3 Construction of the knowledge graph schema layer

Before the knowledge graph is constructed, it's critical to have a deep understanding of domain requirements. In the field of acupuncture and tuina, the core problem faced is disease and treatment, and the main questions to be addressed are as follows.

- What are the possible symptoms of a disease? What treatment techniques should be used and which acupoints should be targeted?
- What are the possible diseases associated with certain symptoms? What treatment techniques should be used and which acupoints should be targeted to relieve symptoms?
- What are the common therapies of acupuncture and tuina? Which acupoints are commonly used for a particular therapy? What kind of functions will it have?
- What are the common acupoints used for acupuncture and tuina? What are the effects of stimulating a particular acupoint?

TABLE 1 Examples of various types of entities.

| Entity | Abbreviation | Example |
|---|---|---|
| Disease | DIS | Hypertension, Heart disease |
| Symptom | SYM | Runny nose, Nasal congestion |
| Acupoint | XW | Weiling, Sanjiao |
| Therapy | OPE | Moxibustion, Push |
| Function | FUN | Reduce fever, Hemostasis |

TABLE 2 Example of each type of entity relationship.

| Category | Abbreviation | Example |
|---|---|---|
| Associations | DIS-SYM | Symptom |
| Treatment | DIS-OPE | Common therapies for treatment |
| | DIS-XW | Commonly used acupoints for treatment |
| | OPE-XW | Commonly used acupoints for therapy |
| Function | OPE-FUN | Therapy has the function |
| | XW-FUN | Stimulation of acupoints has the function |

After analyzing the aforementioned issues, we have identified five main categories of entities that are crucial to the field of acupuncture and tuina. These categories include disease, symptom, acupoint, therapy, and function. The specific descriptions of each category are summarized in Table 1. In Section 5 of our paper, we specifically address the challenge of dealing with these issues.

The distinction between diseases and symptoms lacks a clear demarcation. For instance, the entity "headache" can be regarded both as a disease and as a symptom of a disease. To ensure consistency and prevent conflicts during the construction of the knowledge graph, we impose limitations on the scope of diseases. Specifically, all disease entities are required to align with either the International Classification of Diseases, 10th Revision (ICD-10) for [Classification in Health (Australia), 2004] or the Clinical Terminology for Chinese Medicine, Disease Section (GB/T 16751.1-1997). Consequently, if a suspected entity corresponds to a disease or symptom and fulfills both of the mentioned criteria, it is classified as a disease entity; otherwise, it is categorized as a symptom entity.

Both acupoints and meridians play crucial roles in TCM, representing significant theoretical relationships. Meridians are characterized by their linear distribution throughout the body, while acupoints are specific points or zones along these meridians. It can be considered that acupoints are attributed to meridians, including the existence of extra-meridian points. Given the limited number of meridians, for the purpose of this paper, we consider meridians and acupoints as entities of the same type.

Then, this paper also defines six relationships according to the requirements, as described in Table 2.

After the above definition, the specific framework design of the acupuncture and tuina knowledge graph schema layer is shown in Figure 2.

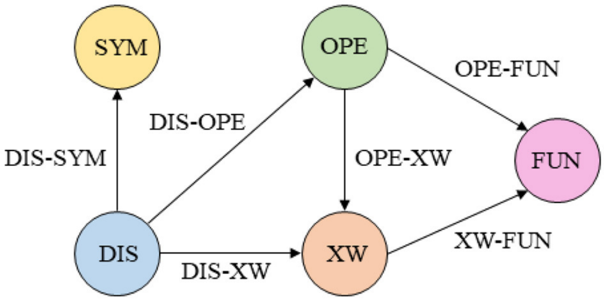


FIGURE 2
Acupuncture and Tuina knowledge graph entity relationship diagram.

TABLE 3 Sample from disease data.

| Disease | Section | Site of disease | Symptom |
|---|---|---|---|
| Measles | Infectious diseases section; pediatrics | Whole body | Fever |
| Rhinitis | Otolaryngology | Nose | Runny nose; Stuffy nose; Loss of smell; |
| Liver cysts | Hepatobiliary surgery | Liver | Abdominal; masses; stomachache |

# 4 Construction of the domain knowledge base

The cold-start problem poses a significant challenge in the construction of domain knowledge graphs at the current stage. Both entity extraction and relationship extraction rely heavily on annotated data, necessitating the creation of comprehensive datasets. Furthermore, annotators involved in dataset annotation are required to possess substantial domain-specific knowledge.

Hence, prior to constructing the data layer, this paper advocates the establishment of a comprehensive domain knowledge base. This knowledge base encompasses essential components, namely the domain dictionary, disease information base, acupoint information base, and acupuncture and tuina literature base. The integration of these resources aims to address the challenges posed by the limited domain expertise of personnel and mitigates the complexities associated with data annotation processes. We also hope to improve the structure of the graph with the domain knowledge base, including increasing the types of relationships and supplementing entity attributes.

## 4.1 Structured knowledge base

According to the above, we mainly collected data for two types of entities, disease, and acupoint. For the disease entities, information on aliases, onset sites, registered departments, and typical symptoms of common diseases was crawled. A total of 2011 common diseases were crawled, and the sample data (only some attributes of some records were extracted) are shown in Table 3.

TABLE 4 Sample from acupoint data.

| Acupoint | Attributed meridians | Functions | Method of operation |
|---|---|---|---|
| Atrium | Conception vessels | Unblocking of veins; Tranquility | Flat prick 0.3 0.5 inch |
| Zuqiaoyin | Gallbladder meridian | Invigorate the meridians: Migraine, Tinnitus | Shallow prick 0.1 inch or prick blood |
| Yangxi | Large intestine channel of hand yangming | Clearing heat and benefiting the throat; Heartburn; Cataracts of the eyes | Straight prick 0.3 0.5 inch |

For the acupoint entity, we crawled the data related to acupoint and meridian in the Chinese medicine network.[2] A total of 366 common acupoints and meridians were crawled, and the sample data (only some attributes of some records were extracted) are shown in Table 4.

In order to make the knowledge base scalable and more efficient for retrieval, the obtained data are stored in database in the form of tables to establish a structured knowledge base.

After crawling the structured data, we directly populate it into the knowledge graph as entities or their corresponding relationships. This approach enables us to include some data directly in the initial graph, effectively addressing the cold start problem. As an illustration, consider the second row of data in Table 3, we discern that the symptoms associated with rhinitis include runny nose, stuffy nose, and loss of smell. This information was directly incorporated into the KG, accompanied by the relevant relationships, resulting in the creation of a node graph representing the interconnections among entities linked to rhinitis.

## 4.2 Domain dictionary

Before proceeding with formal entity extraction, it is necessary to collect a representative sample of each entity type in order to build a comprehensive entity dictionary. The entities were primarily sourced from Chinese medicine websites[3] and other relevant sources. An example dataset for the dictionary is presented in Table 5. To maintain consistency, the length of all entities was limited to 10 characters. It should be noted that not all entities in the knowledge base or domain dictionary were added to the knowledge graph in order to avoid creating a large number of "isolated nodes" and connected components during entity extraction. If the current entity appears in the corpus then it is added to the knowledge graph. Conversely, if it only appeared within the domain dictionary and not in the corpus, it means that it is likely to have no relational triples and is isolated, then it will not be added to the knowledge graph. Finally, only about 48% of the entities were added to the graph.

---

[2] Acupuncture and Tuina - Acupuncture techniques, acupuncture treatment, meridians and acupoints, TCM Tuina. http://www.pharmnet.com.cn/tcm/zjdq/. Acessed:2022-05-09.

[3] Complete List of Human Acupuncture Points—Chinese Medicine Search. http://www.zhongyoo.com/xw/. Acessed:2022-05-04.

TABLE 5 Example data for the dictionary.

| Entity | Number | Example |
|---|---|---|
| Disease | 14,691 | Enteritis, Pneumoconiosis |
| Symptom | 5,142 | Intestinal tinnitus, Dry lips |
| Acupoint | 594 | Guanchong, Zhongwan |
| Therapy | 579 | Slap, Twist, Push |
| Function | 357 | Toning the spleen, stomach, liver and kidneys |

## 4.3 Unstructured literature base

The corpus data used for information extraction is mainly obtained from three parts: China National Knowledge Infrastructure (CNKI),[4] Chinese Medicine and Chinese Herbs Network,[5] and Chinese Medicine Network (see text footnote 3). literature for subsequent analysis and knowledge graph construction.

CNKI is a comprehensive repository of knowledge resources in China, covering a wide range of subject areas. In this website, we are able to obtain valuable and authoritative literature related to acupuncture and tuina. However, due to copyright restrictions, we can only download a limited number of relevant documents. To further analyze the literature, we performed optical character recognition (OCR) on the PDF documents, converting them into TXT format to obtain editable text. This allowed us to extract and process the textual content of the literature for subsequent analysis and knowledge graph construction.

Chinese Medicine and Chinese Herbs Network and Chinese Medicine Network are similar open websites for Chinese medicine, with a large number of medical post related to acupuncture and tuina. These posts serve as valuable sources for extracting entities and relationships required for constructing the knowledge graph. To compensate for the limited corpus available on the Knowledge Network, we collected web texts from these medical websites. In total, we crawled 3,236 posts relevant to acupuncture and tuina.

The collected literature underwent basic preprocessing, including the filtering of special symbols. The title, content, and source information of the literature were then stored in the database, establishing a comprehensive library of acupuncture and tuina literature.

# 5 Construction of the graph data layer

## 5.1 Entity extraction based on fusion models

We used the above corpus for entity extraction. It's a combination of the Trie tree model and deep learning model, as depicted in Figure 1B of a diagram. The Trie tree construction relied on the previously built domain dictionary and primarily extracted known entities. The extraction results from the Trie tree served as the dataset for training the deep learning model. The trained model was then used for a second round of entity extraction, focusing on identifying potential entities that were not initially discovered. Finally, the extraction results from both methods were merged to obtain the final entity set.

### 5.1.1 Entity extraction based on Trie tree

Trie tree, which is also known as dictionary tree and prefix tree. Figure 3 shows a Trie tree and the list of words it contains. For this tree, a query for the entity "rhinitis" would follow the path "1-2-5".

In this study, the construction of Trie trees was implemented using a Dictionary built in Python. The key-words of the Trie tree are strings, and the Trie tree stores each keyword in a path instead of a node. With different node paths, different keywords were gotten, i.e. entities. In addition, two keywords with a common prefix have the same path in the prefix part of the Trie tree. Here, we modeled a total of five Trie trees, each representing a different entity type, for entity extraction. The algorithm for constructing the Trie trees was described in Algorithm 1.

After obtaining the Trie tree, entity extraction and data annotation were performed for the 819 literature samples. The data annotation followed the BIO annotation format. For each literature, it was divided into sentences, and Algorithm 2 was applied to generate candidate words and match them with the Trie tree. The entities that matched successfully were annotated with the corresponding BI tags, while unsuccessful candidates were annotated with the O tag. Entity extraction was conducted for all 819 samples. The number of entity samples obtained for each entity type is presented in Table 6, and the annotation samples can be seen in Figure 4.

### 5.1.2 TemplateFC based entity extraction

#### 5.1.2.1 Dataset pre-processing

The BERT embedding models have a maximum input sequence length of 512 tokens. Considering the two special characters [CLS] and [SEP], the original sequence length should be less than 510 tokens. Since most of the literature in the previous section exceeds this length, we needed to divide the literature.

We opted to divide the literature into blocks while aiming to preserve as much information as possible about the headings at each level. The blocks were created to be as large as possible but still smaller than 510 tokens. Additionally, we retained the requirement of dividing the literature by sentence. As a result, a total of 2505 literature pieces were obtained after the block division. Subsequently, Trie tree-based entity extraction was conducted on the divided literature to obtain the initial training dataset for the deep learning model.

#### 5.1.2.2 TemplateFC model

In the template-free prompt tuning method, NER (Named Entity Recognition) was reimagined as a language modeling (LM) task. However, instead of relying on predefined templates, a new objective called Entity-oriented LM (EntLM) was introduced to

---

[4] Search—China Knowledge Network. https://kns.cnki.net/kns/brief/default_result.aspx. Acessed:2022-05-09.

[5] TCM.com_TCM.com_Excellent Chinese Medicine Portal. https://www.zhzyw.com/. Acessed: 2022-05-09.

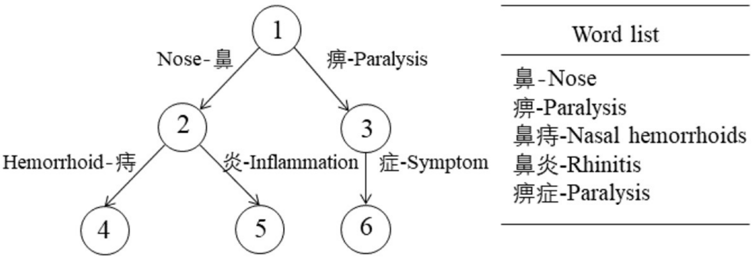


FIGURE 3
The example of a Trie Tree.

```
Input: $root : Node, word : String$
Output: $root : Node$
$curNode \Leftarrow root$
for $c$ in $word$ do
    if $cnotincurNode.child$ then
        $curNode.child[c] \Leftarrow Node$
    end if
    $curNode \Leftarrow curNode.child[c]$
    $N \Leftarrow n$
    $curNode.end \Leftarrow True$
end for
return $root$
```

**Algorithm 1. AddTrieNode.**

```
Input: $root : Node, sentence : String$
Output: $entities : List$
$curNode \Leftarrow root; index \Leftarrow 0; maxLength \Leftarrow 10; entities \Leftarrow [\ ]$
while $index < len(sentence)$ do
    $j \Leftarrow maxLength$
    while $j \neq 0$ do
        $word \Leftarrow sentence[index : index + j]$
        if $SearchWord(root, word) = True$ then
            $index \Leftarrow index + j - 1; entities.add(word);$
        end if
        $j \Leftarrow j - 1$
    end while
    $index \Leftarrow index + 1$
end while
return $entities$
```

**Algorithm 2. FindEntity.**

TABLE 6 Trie tree extraction entity sample statistics.

| Entity | Total | DIS | SYM | XW | OPE | FUN |
|---|---|---|---|---|---|---|
| Number | 10,346 | 3,671 | 3,252 | 1,491 | 1,330 | 602 |

fine-tune NER without reusing the LM objective (as done in previous approaches using templates) (Ma et al., 2021). In this method, the LM was trained to predict a label word at the position of the entity when given input text, serving as an indication of the entity's label. For non-entity words like "was" the LM continued to predict the original word.

However, the template-free prompt tuning method also has some disadvantages when applied to the NER task, particularly in the context of Chinese text. Chinese entities often exhibit continuity, leading to situations where a single entity is identified as multiple entities. For instance, in the case of "gastrointestinal neurosis", the term "gastrointestinal" might be identified as an acupoint entity, while "neurosis" is recognized as a disease entity.This highlights the challenge of accurately identifying and disambiguating Chinese entities due to their structural uniqueness. Another limitation is that the EntLM model is primarily designed for English text and may not be well-suited for Chinese language processing. In English, most entities are typically represented by a single token, whereas Chinese entities often span multiple tokens. This difference in tokenization and entity representation increases the likelihood of inaccurate identification when applying the EntLM model to Chinese text.

In this paper, we proposed an NER model that combines the benefits of template-free prompt learning with enhanced applicability to Chinese text (TemplateFC). We continued to employ a template-free entity-level LM fine-tuning process, but we introduced additional components, namely Bi-directional Long Short-Term Memory (BiLSTM) and Conditional Random Field (CRF) layers, to improve the accuracy of Chinese entity recognition. By incorporating these layers into the training process, the model can learn useful constraints that facilitate better fine-tuning of the pre-trained model.

First, we selected the appropriate label words for the NER model. Due to the limited availability of few-shot data, the label words were chosen randomly. However, to improve the generalizability of the model, we selected label words from a pool of 1,060 documents in the literature base. The selection of label words was performed using the best method from the EntLM model, which combined both Data and LM search. The process of selecting label words for the B-DIS label was illustrated in the Figure 5.

Figure 6 is the general structure of the model. This model first gives the input text $X = \{x_1, ..., x_n\}$, and its corresponding label sequence is $Y = \{y_1, ..., y_n\}$. Here, we set the set of label words to $V_l$, and it was connected to the set of task labels with a mapping function $M : Y \rightarrow V_l$. Next, the target sequence we want to obtain is $X^{Ent} = \{x_1, ..., M(y_i), ..., x_n\}$ ( Here, we assume that the word at $i$ is the entity label). The loss function is shown in Equation 1.

$$L_{word} = -\sum_{i=1}^{n} \log P(x_i = x_i^{Ent} \mid X). \tag{1}$$

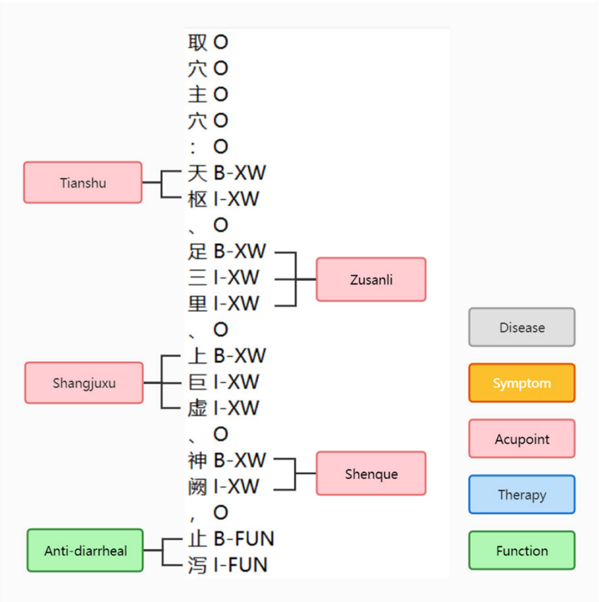


FIGURE 4
BIO labeling example.

In the following, we set the embedding vector $W = \{w_1, \ldots, w_n\}$ obtained after the BERT pre-training model. Next, the embedding vector representation $W$ gets the Emission score matrix $X_{r,l} \in R^{n\times l}$ ($n$ is the number of tokens, $l$ is the number of label types) of the current token location about each label through the BiLSTM layer. Each token may be a different label, so there are $n \times l$ possible paths, and only one true path exists among them. Therefore, after passing through the CRF layer, a new loss function was obtained as shown in Equation 2.

$$L_{entity} = \frac{P_{realpath}}{\sum_{i=1}^{n} P_i}. \tag{2}$$

$P_i = e^{S_i}$ is the score of the $i$-th path. The exact calculation of $S_i$ is shown in Equation 3.

$$S_n = EmissionScore + TransitionScore. \tag{3}$$

*EmissionScore* can be obtained from the Emission score matrix, and *TransitionScore* is composed of parameters in the CRF layer. Finally, we combined the two loss functions to get the new Loss value and performed joint training, the new Loss value is shown in Equation 4.

$$L_{all} = L_{word} + L_{entity}. \tag{4}$$

After the input text $X$ got the embedding vector $W$ by the pre-training model, the label corresponding to the current token can be obtained directly by the argmax function, as shown in Equation 5.

$$Y = argmax(x_i = x_i^{Ent} \mid X). \tag{5}$$

Alternatively, the corresponding label can be obtained by decoding through the CRF layer.

### 5.1.3 Fusion of entity extraction results

When the entity extraction based on the Trie tree and ALBERT-CRF were completed, the results of the two extractions needed to be fused. Drawing on the literature, and Meituan's design of the entity extraction model, we used dictionary matching to mine discovered entities and deep learning models to mine potential entities. The following fusion rules were developed.

- When the Trie tree extraction results agree with the model extraction results, the results are fused directly.
- When the Trie tree extraction result is "O" and the model extraction result is "B-" or "I-", the fusion result will be based on the model extraction result.
- When the Trie tree extraction result is "B-" or "I-" and the model extraction result is "O", the fusion result will be based on the Trie tree extraction result.

The specific fusion algorithm is shown in Algorithm 3.

Finally, a total of 10346 entities were extracted from all the literature for the fusion model. Among them, 3671 are DIS entities, 3252 are SYM entites, 149 are XW entities, 1330 are OPE entities, 602 are FUN entities.

## 5.2 Rule-based relationship extraction

Acupuncture and tuina constitute an ancient and specialized medical domain characterized by a stable terminology and conceptual framework. Given the absence of an initial relational dataset and the challenge of acquiring extensive labeled data, opting for a rule-based approach becomes advantageous. This approach involves extracting relational patterns based on expert experience and existing literature, thereby enhancing alignment with the specialized nature of acupuncture and massage. Additionally, it serves to alleviate the burden associated with data labeling. To establish an expression paradigm for the relationships, a substantial amount of text was analyzed and summarized. Subsequently, we devised a set of coherent matching rules to facilitate the extraction of relationships. The formulation of rules primarily encompassed subject word-based extraction, entity location-based extraction, and keyword-based extraction strategies. Our rules were built based on a large amount of text and facts. For example, "Body acupuncture and massage with acupuncture operation: Quchi, Hegu and matching acupoints", our model identified body acupuncture and massage as OPE entities, identified Quchi and Hegu as XW entities, and extracted the corresponding OPE-XW relationships.

This paper combines three common rule-based approaches and an analysis of literature data to develop rules as shown in Table 7.

A total of 40,919 relationships were extracted from all the literature. Among them, 5566 are DIS-SYM relationships, 6705 are DIS-OPE relationships, 6412 are DIS-XW relationships, 6808 are OPE-XW relationships, 6824 are OPE-FUN relationships, 8604 are XW-FUN relationships.

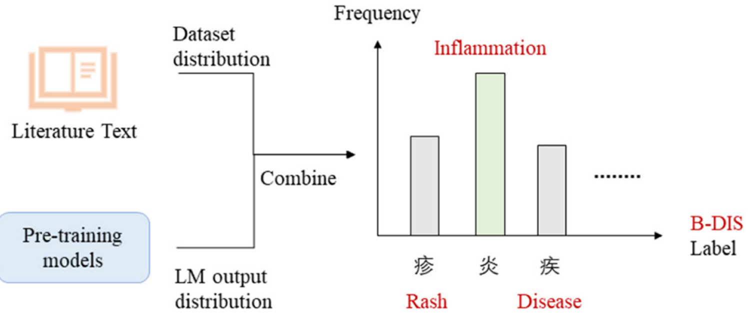


FIGURE 5
Label words selection.

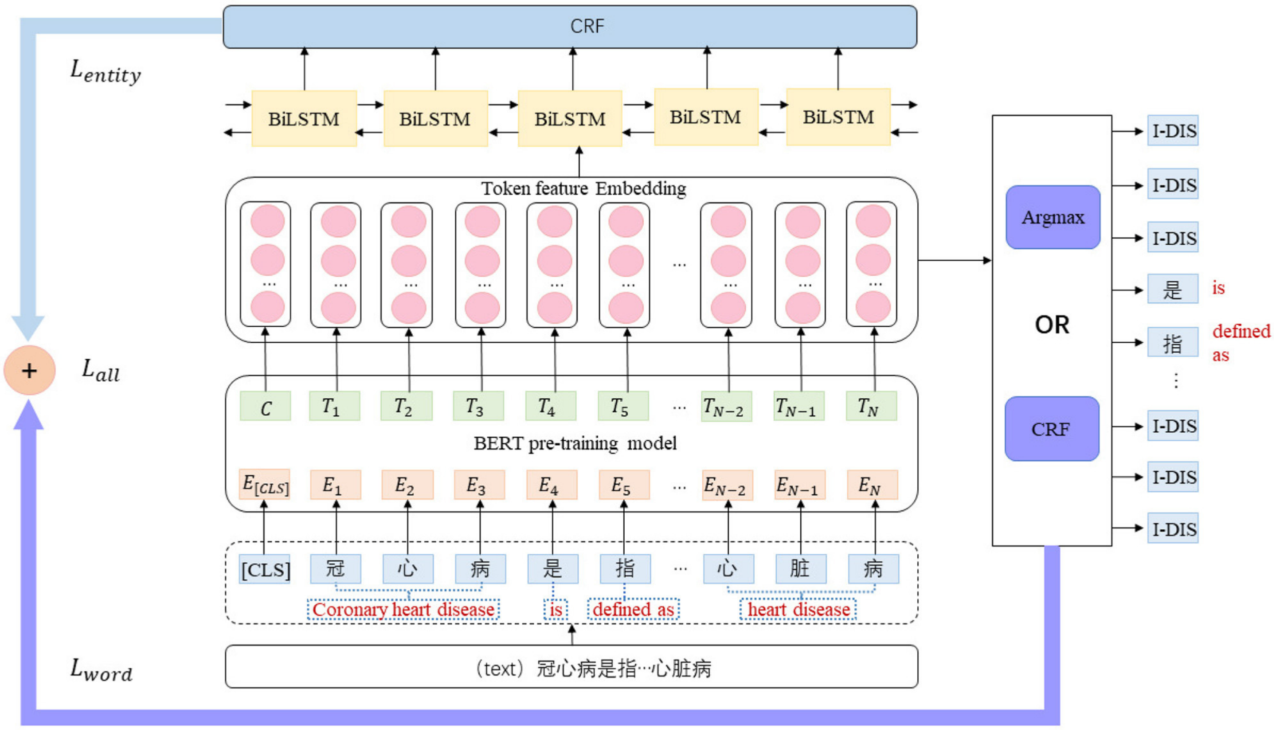


FIGURE 6
TemplateFC model architecture diagram.

# 6 Experiments and results

## 6.1 Datasets and implementation details

This paper addressed the limitation of resources in the TCM domain by conducting two experiments: one in a resource-rich setting and another in a few-shot setting. To evaluate our approach, we utilized two datasets from distinct domains, the CoNLL2003 dataset (Sang and De Meulder, 2003) sourced from the newswire domain, and a self-built database specific to the acupuncture and tuina domain within TCM. The details of these experiments were provided below.

**Multiple-shot NER Dataset:** The dataset used in this study was primarily derived from the CoNLL2003 dataset. The textual content of this dataset focused on the People's Daily, which is one of the most influential newspaper publications in China. The dataset comprised three common entity types: Person (PER), Place (LOC), and Organization (ORG) with 25862 training sets, 4671 test sets, and 2,385 validation sets.

**Fewshot-NER Dataset:** This dataset is a 10-shot dataset extracted manually from the above Multiple-shot NER Dateset.

**ZJTA Dateset:** This dataset was constructed based on the previous method for NER and ERE. It comprises five entity types and six relationship types, resulting in a total of 10,346 entities and

```
Input: $entity : String, label1 : String, label2 : String$
Output: $entity : String, label : String$
$curEntity \Leftarrow entity; curLabel1 \Leftarrow label1; curLabel2 \Leftarrow label2$
if $curLlabel1 = curLabel2$ then
    $label = curLlabel1$
else if $curLabel1 = \backslash O'' and (curLabel2 = \backslash B -'' or curLabel2 == \backslash I-'')$ then
    $label = curLlabel2$
else if $(curLabel1 = \backslash B -'' or curLabel1 = \backslash I-'') and curLabel2 = \backslash O''$ then
    $label = curLlabel1$
else
    $break$
end if
return $label$
```

Algorithm 3. Blind.

TABLE 7 Relationship extraction rules.

| Category | Rule | The relationship of extraction |
|---|---|---|
| Subject word | Disease therapy literature | DIS-SYM, DIS-OPE, DIS-XW |
| | Introduction to therapy literature | DIS-OPE, OPE-XW, OPE-FUN |
| Entity location | Proximity of hand techniques and acupoints | OPE-XW |
| Keyword | "efficacy", "effect", "function" | OPE-FUN, XW-FUN |

40,919 relationships. It is important to note that due to limited availability of public information and restricted access to Chinese medicine acupuncture and tuina, the dataset may be limited in size. Consequently, for the few-shot experiment, we selected a subset of this data. Specifically, the few-shot experiment utilized 10 training texts and 498 test texts.

Finally, We selected the F1 score as the evaluation metric for our experiments. The F1 score, which is a balanced measure of precision and recall, effectively reflects the performance of our experiments in a fair and comprehensive manner.

## 6.2 Baselines and proposed models

In our experiments, we have chosen the richer baselines with the following details.

**Bert + HMM:** A classical model which combines BERT and HMM for NER. It learns the representation of the input text using BERT, which converts each word or subword into a high-dimensional vector representation capturing its rich semantic information. And then models the sequence of BERT representations using the HMM model, which can be used for NER.

**TemplateNER** (Cui et al., 2021): A Prompt Learning approach: using generative BART models for sequence annotation tasks and exploring the potential of BART models for few-shot scenarios using a Template-based approach.

TABLE 8 Comparison of F1 score results for different resources under two data domains.

| Method | Multiple-shot | Few-shot | ZJTA |
|---|---|---|---|
| Bert+HMM | 0.8879 | 0.1862 | 0.2025 |
| TemplateNER | 0.8942 | 0.2945 | 0.3193 |
| Two-tower | 0.9102 | 0.2893 | 0.3107 |
| EntLM | 0.9041 | 0.3373 | 0.3871 |
| EntLM-CRF | 0.9039 | 0.3543 | 0.4016 |
| TemplateFC (Ours) | **0.9174** | **0.3746** | **0.4268** |

**Two-tower** (Ma et al., 2022): The Few-shot Named Entity Recognition (NER) problem was addressed using a two-tower model. The model comprises two BERT encoders: one encoder is responsible for encoding the representation of each token, while the other encoder encodes the natural language form of the BIO label to obtain the label representation. Subsequently, the model predicts the similarity between each token and all the label representations within the text. Finally, the label with the highest similarity is assigned to the token.

**EntLM** (Ma et al., 2021): A approach which abandoned the template and used NER as a language model task. In NER task, the position of the entity was predicted as label word, and the non-entity position was predicted as the original.

**EntLM-CRF:** Add CRF layer decoding on top of the above model.

**TemplateFC:** The proposed method.

Table 8 presents the results of the proposed method and the baseline approaches across different settings. Here, the bold values represent the highest experimental results in the current dataset. Based on the findings from the table, the following observations can be made.

1. Our model demonstrates higher scores on the common dataset, both in multiple-shot and few-shot conditions. There are several reasons why the TemplateFC model demonstrates superiority. Firstly, the incorporation of BiLSTM and CRF layers guides BERT to learn more nuanced representations by leveraging sequence label information. This aids BERT in comprehending the structural and semantic nuances within input sequences, particularly when data is scarce, thereby providing supplementary supervisory signals. Secondly, the BiLSTM and CRF layers enable the back-propagation of gradients from label prediction errors, facilitating BERT in adapting its representation learning process based on erroneous predictions. This joint training methodology expedites BERT's convergence to enhanced representations with fewer samples. Lastly, the BiLSTM and CRF layers contribute to BERT's improved understanding of context and sequence continuity, resulting in more coherent and semantically enriched representations.
2. In relation to the self-created dataset in this paper, our model exhibits superior performance, surpassing EntLM-CRF by 2.5%. This underscores the advantages of integrating the BiLSTM and CRF layers for joint training, leading to enhanced model

fitting speed, improved NER accuracy, and heightened model generalization.

3. Even in scenarios with abundant resources, our model maintains an advantage of 0.72% compared to the highest-scoring baseline model. This highlights the versatility of our approach, which is applicable across both few-shot and multi-shot scenes. Despite the potential risk of overfitting in multi-shot scenarios, our experimental findings indicate that our model continues to enhance model performance and stability.
4. It is important to note that for the last three baselines, we conducted four training runs to obtain four sets of results. Subsequently, we selected the highest F1 score among the results for comparison. Conversely, for the experiments conducted in this paper, we directly decoded the outputs in the multiple-shot scenario. In the few-shot resource scenario, we employed CRF decoding.

## 6.3 Implementation of the knowledge graph

After the completion of entity extraction, the obtained set of entities and relational triples were stored in Neo4j, a graph database. To facilitate querying and exploration, graph queries were executed using the Cipher language. Additionally, to cater to the needs of non-specialists, a Python query interface.and a web-based query user interface (UI) were developed.

Considering the practical applicability in real-life scenarios, queries were primarily conducted for four main types of entities: diseases, symptoms, acupoints, and therapies. The queries were limited to a step size of 1, which corresponds to single-hop queries. This approach enabled efficient retrieval of related information within a single query operation.

After completing the development phase, the website was deployed locally using a personal computer. Subsequently, local testing was conducted to ensure its functionality.

During the entity search, we entered the query “coronary artery disease”, and the results are displayed in Figure 7. The diagram illustrates that coronary artery disease is associated with symptoms such as shortness of breath and sweating. Moreover, commonly employed treatments include pointing, pressing, and kneading. Lastly, specific acupoints such as Zhimen and Zhiyang can be targeted for treatment. This effectively addresses question one as outlined in section two.

To substantiate the credibility of our research, we have opted to engage a reputable medical platform, Hua Tuo GPT. We have meticulously designed a set of three targeted inquiries to be posed and elucidated on this platform:

- What are the main symptoms of coronary heart disease?

  **Answer:** The principal manifestations encompass *angina* or *discomfort in the chest*, sometimes accompanied by secondary symptoms like *dyspnea* and *diaphoresis*. In severe instances, symptoms can escalate to *myocardial infarction*. Urgent medical attention is imperative upon experiencing any of these signs, ensuring prompt mitigation of health risks. Furthermore, proactive lifestyle adjustments, such as smoking cessation, alcohol moderation, weight maintenance, and increased physical activity, can collectively serve as preventative measures against disease onset and progression.
- What are the common therapies for the treatment of coronary heart disease?

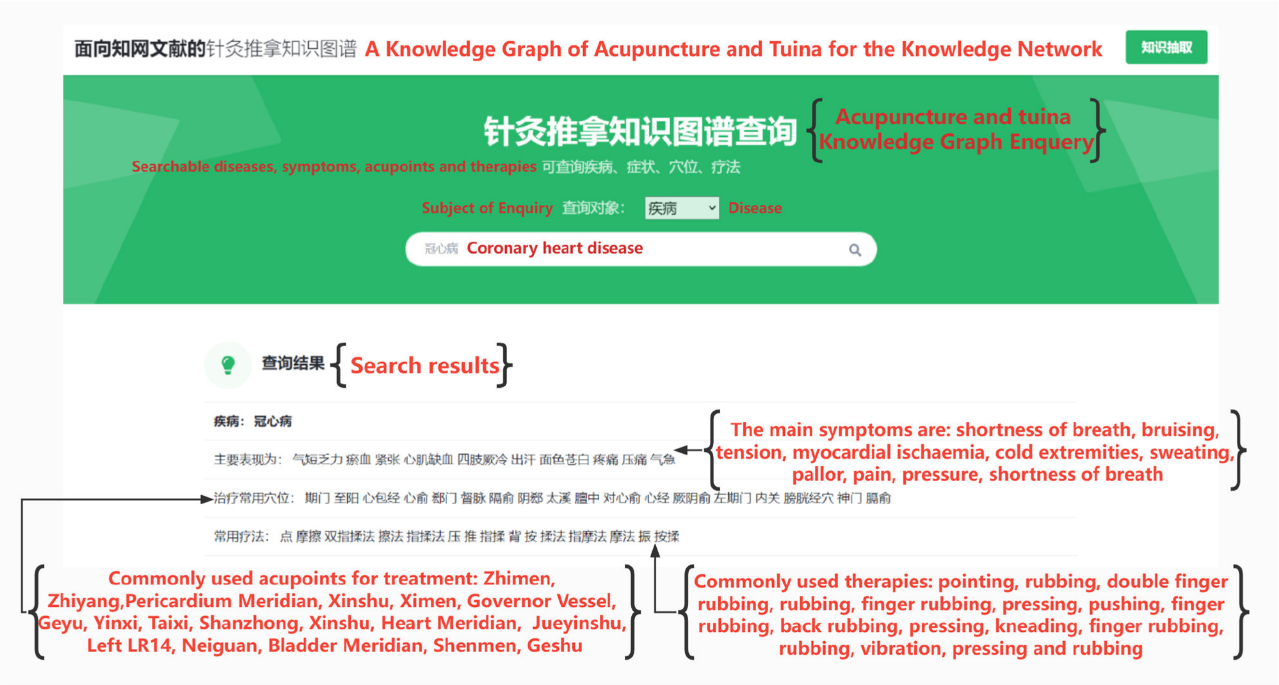
  **Answer:** The treatment paradigm encompasses both pharmacological and non-pharmacological interventions. Pharmacotherapy predominantly targets pain alleviation and cardiovascular function regulation through agents like *anticoagulants*, *lipid-lowering agents*, and *vasodilators*. Conversely, non-pharmacological avenues involve dietary restructuring, appropriate physical exertion, psychological counseling, among others, which collectively alleviate symptoms and foster recuperation. Importantly, tailored application of a multitude of treatment approaches yields optimal outcomes across diverse coronary heart disease presentations and severities.
- What acupoints to be targeted and what techniques to be used to relieve coronary heart disease?

  **Answer:** Acupoints include the *Xinshu* and *Danzhong*. Commonly used acupuncture techniques include *tonifying deficiency*, *clearing the heart*, *calming asthma*, and *warming yang*, while tuina focuses on regulating qi and blood, promoting blood circulation for removing obstruction in collaterals, and other aspects. However, due to the variability of each individual’s body and different disease states, the choice of acupoints and techniques may vary. Therefore, it is recommended that you consult with a professional practitioner and obtain the necessary assessment and guidance before undergoing acupuncture and tuina treatment.

After analyzing the three provided questions and their corresponding answers, several conclusions can be deduced:

- Enhanced Detail and Precision: Our responses exhibit a notable augmentation in both detail and specificity. For instance, in elucidating the symptoms of coronary heart disease, our answers encompass a broader spectrum of symptoms, encompassing pallor and bruising, which the website’s responses did not encompass. This augmentation in symptom coverage enhances the comprehensiveness of our answers.
- Domain-Specific Focus: Our answers distinctly exhibit a focused domain expertise, centered around the realm of acupuncture and tuina. This is reflected in the knowledge graph that this paper presents, which is intrinsically tailored to the acupuncture and tuina domain. Conversely, the website’s responses tend to encompass a more generalized medical perspective, devoid of the intricate nuances specific to acupuncture and tuina. This divergence underscores our paper’s potency in providing in-depth insights within the acupuncture and tuina domain.

In summation, our website notably excels in the sphere of acupuncture and tuina, securing a distinct advantage over the comprehensive medical advice provided by the website. Our work



面向知网文献的针灸推拿知识图谱 A Knowledge Graph of Acupuncture and Tuina for the Knowledge Network 知识抽取

针灸推拿知识图谱查询 {Acupuncture and tuina Knowledge Graph Enquery}

Searchable diseases, symptoms, acupoints and therapies 可查询疾病、症状、穴位、疗法

Subject of Enquiry 查询对象：疾病 Disease

冠心病 Coronary heart disease

查询结果 {Search results}

疾病：冠心病

主要表现为：气短乏力 瘀血 紧张 心肌缺血 四肢厥冷 出汗 面色苍白 疼痛 压痛 气急

治疗常用穴位：期门 至阳 心包经 心俞 郄门 督脉 隔俞 阴郄 太溪 膻中 对心俞 心经 厥阴俞 左期门 内关 膀胱经穴 神门 膈俞

常用疗法：点 摩擦 双指揉法 擦法 指揉法 压 推 指揉 背 按 揉法 指摩法 摩法 振 按揉

{The main symptoms are: shortness of breath, bruising, tension, myocardial ischaemia, cold extremities, sweating, pallor, pain, pressure, shortness of breath}

{Commonly used acupoints for treatment: Zhimen, Zhiyang,Pericardium Meridian, Xinshu, Ximen, Governor Vessel, Geyu, Yinxi, Taixi, Shanzhong, Xinshu, Heart Meridian, Jueyinshu, Left LR14, Neiguan, Bladder Meridian, Shenmen, Geshu}

{Commonly used therapies: pointing, rubbing, double finger rubbing, rubbing, finger rubbing, pressing, pushing, finger rubbing, back rubbing, pressing, kneading, finger rubbing, rubbing, vibration, pressing and rubbing}

FIGURE 7
Coronary heart disease text extraction results.

leverages its domain-specific focus to furnish detailed and precise counsel, aligning with the depth and expertise inherent in the acupuncture and tuina field. This renders our website a preeminent source of tailored advice within the realm of acupuncture and tuina, underscoring its authority and value within this specialized domain.

## 7 Discussion

This paper presents the establishment of a novel knowledge base in the domain of acupuncture and tuina, utilizing modern literature as the foundation. Subsequently, the schema layer of the acupuncture and tuina knowledge graph was designed, considering the requirements for practical applications in TCM. The NER task was accomplished through fused Trie extraction and model extraction techniques, while the ERE task was completed using rule-based methods. Additionally, in the context of few-shot learning, we proposed a TemplateFC model, which becomes a more adaptable template-free prompt tuning method for Chinese text by adding BiLSTM layer and CRF layer for joint training. Lastly, the work encompassed graph storage and querying, enabling the KG of acupuncture and tuina to facilitate doctors' understanding of relevant knowledge and give diagnostic and therapeutic advice.

The method proposed in this paper aims to organize and integrate various knowledge and concepts within the acupuncture and tuina domain, culminating in a structured knowledge graph. This facilitates knowledge sharing and communication among experts and researchers from diverse fields, fostering cross-disciplinary collaboration and discourse. Nonetheless, certain limitations persist in this study. For instance, within ERE, the conventional rule-based approach struggles to encompass all text features adequately, posing challenges for migration. In future work, we intend to explore deep learning-based methods for relationship extraction, tailored to the nuances of Chinese text, thereby advancing TCM development. Concurrently, we aim to delve into the capabilities of Large Language Models (LLMs), aspiring to enhance the precision in the extraction of entities and their interrelationships within textual datasets. The ultimate objective of this exploration is to attain zero-shot learning capabilities, thereby significantly advancing the efficacy and adaptability of our models in understanding and processing complex textual information.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Author contributions

XL: Conceptualization, Funding acquisition, Methodology, Project administration, Writing — original draft, Writing — review & editing, Data curation, Formal analysis, Investigation, Software, Validation, Visualization. XH: Conceptualization, Funding acquisition, Methodology, Project administration, Writing — original draft, Writing — review & editing, Resources, Supervision. SW: Data curation, Software, Validation, Writing — original draft, Writing — review & editing. YL: Project administration, Supervision, Writing — original draft, Writing — review & editing. RG: Conceptualization, Funding acquisition, Writing — original draft, Writing — review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The National Key Research and Development Program of China No. 2021YFF1201200, the National Natural Science Foundation of China Nos. 62372494 and 62172187, the Science and Technology Planning Project of Jilin Province Nos. 20220201145GX, 20200708112YY, and 20220601112FG, the Science and Technology Planning Project of Guangdong Province No. 2020A0505100018, Guangdong Universities' Innovation Team Project No. 2021 KCXTD015, and Guangdong Key Disciplines Project Nos. 2021ZDJS138 and 2022ZDJS139.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Cao, D. (2008). To discuss the chinese medicine external treatment development from the source of acupuncture and manipulation. *J. Henan Coll. Trad. Chin. Med.* 23, 78–79.

Chen, Y., Zheng, Y., and Yang Z. (2023). "Prompt-based metric learning for few-shot NER," in *Findings of the Association for Computational Linguistics: ACL* (Toronto, ON: Association for Computational Linguistics), 7199–7212.

Classification in Health (Australia), N. C. (2004). *The International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Australian Modification (ICD-10-AM)*. Hobart: National Centre for Classification in Health.

Cui, L., Wu, Y., Liu, J., Yang, S., and Zhang, Y. (2021). "Template-based named entity recognition using BART," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP* (Association for Computational Linguistics), 1835–1845.

Cui, M., Jia, L., Yu, T., Yang, S., Zhu, L., Li, J., et al. (2014). "Current status of Traditional Chinese Medicine language system," in *Frontier and Future Development of Information Technology in Medicine and Education: ITME* (Springer), 2287–2292.

Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., and Bai, X. (2019). "Named entity recognition using bert bilstm crf for chinese electronic health records," in *2019 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI)* (Suzhou: IEEE), 1–5.

Eberts, M., and Ulges, A. (2019). "Span-based joint entity and relation extraction with transformer pre-training," in *European Conference on Artificial Intelligence*.

Guan, R., Zhang, H., Liang, Y., Giunchiglia, F., Huang, L., and Feng, X. (2020). Deep feature-based text clustering and its explanation. *IEEE Trans. Knowl. Data Eng.* 34, 3669–3680. doi: 10.1109/TKDE.2020.3028943

Han, X., Li, X., Liang, Y., Wang, X., Xu, D., and Guan, R. (2021). "Acupuncture and tuina knowledge graph for ancient literature of traditional chinese medicine," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Houston: IEEE), 674–677.

Hou, M., Wei, R., Lu, L., Lan, X., and Cai, H. (2018). A review of knowledge graph research and its application in the medical field. *Comp. Res. Dev.* 55, 2587–2599. doi: 10.7544/issn1000-1239.2018.20180623

Lee, D., Kadakia, A., Tan, K., Agarwal, M., Feng, X., Shibuya, T., et al. (2021). "Good examples make a faster learner: Simple demonstration-based learning for low-resource ner," in *arXiv*. doi: 10.18653/v1/2022.acl-long.192

Liu, L., and Zhao, T. (2020). *Research on the Construction Method of Chinese Medical Knowledge Graph Based on Multi Resources* (Master's thesis). Harbin: Harbin Institute of Technology.

Liu, Q., Li, Y., Duan, H., Liu, Y., and Qin, Z. (2016). A review of knowledge graph construction techniques. *Comp. Res. Dev.*53, 582–600. doi: 10.7544/issn1000-1239.2016.20148228

Liu, S., Wang, Z., Su, Y., Qi, L., Yang, W., Fu, M., et al. (2021). A neuroanatomical basis for electroacupuncture to drive the vagal-adrenal axis. *Nature* 598, 641–645. doi: 10.1038/s41586-021-04001-4

Liu, S., Wang, Z., Su, Y., Ray, R. S., Jing, X., Wang, Y., et al. (2020). Somatotopic organization and intensity dependence in driving distinct npy-expressing sympathetic pathways by electroacupuncture. *Neuron* 108, 436–450. doi: 10.1016/j.neuron.2020.07.015

Lu, L., Zhang, Y., Tang, X., Ge, S., Wen, H., Zeng, J., et al. (2022). Evidence on acupuncture therapies is underused in clinical practice and health policy. *BMJ* 376, e67475. doi: 10.1136/bmj-2021-067475

Ma, R., Zhou, X., Gui, T., Tan, Y., Li, L., Zhang, Q., et al. (2021). "Template-free prompt tuning for few-shot ner," in *arXiv*.

Ma, R., Zhou, X., Gui, T., Tan, Y., Li, L., Zhang, Q., et al. (2022). "Template-free prompt tuning for few-shot NER," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle, WA: Association for Computational Linguistics), 5721–5732.

Milovsević, N., and Thielemann, W. (2023). Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *J. Web Semant.* 75, 100756. doi: 10.1016/j.websem.2022.100756

Moscato, V., Napolano, G., Postiglione, M., and Sperlì, G. (2023). Multi-task learning for few-shot biomedical relation extraction. *Artif. Intellig. Rev.* 6, 1–21. doi: 10.1007/s10462-023-10484-6

Sang, E. F., and De Meulder, F. (2003). "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in in *arXiv*. doi: 10.3115/1119176.1119195

Unschuld, P. U. (2016). *Huang Di Nei Jing Ling Shu: The Ancient Classic on Needle Therapy*. California: Univ of California Press.

Wang, D., Liang, Y., Xu, D., Feng, X., and Guan, R. (2018). A content-based recommender system for computer science publications. *Knowl.-Based Syst.* 157, 1–9. doi: 10.1016/j.knosys.2018.05.001

Yu, T., Liu, J., Jia, L., Zhang, Z., Yang, S., Liu, L., et al. (2015). Research on the construction of big knowledge graph for traditional chinese medicine. *China Digital Med.* 10, 80–82. doi: 10.3969/j.issn.1673-7571.2015.3.027

Zhang, L., and Xia, Y. (2018). *Study of Formation Process and Influences of Contemporary Acupuncture Schools* (Master's thesis). Nanjing: Nanjing University of Chinese Medicine.

Frontiers in Genetics

TYPE Original Research
PUBLISHED 06 May 2024
DOI 10.3389/fgene.2024.1393406

Check for updates

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Bolin Chen,
Northwestern Polytechnical University, China
Fei Guo,
Central South University, China
Wei Peng,
Kunming University of Science and Technology, China

*CORRESPONDENCE
Junwei Luo,
✉ luojunwei@hpu.edu.cn

RECEIVED 29 February 2024
ACCEPTED 11 April 2024
PUBLISHED 06 May 2024

CITATION
Shen J, Wang Y and Luo J (2024), CD-Loop: a chromatin loop detection method based on the diffusion model.
*Front. Genet.* 15:1393406.
doi: 10.3389/fgene.2024.1393406

COPYRIGHT
© 2024 Shen, Wang and Luo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# CD-Loop: a chromatin loop detection method based on the diffusion model

Jiquan Shen, Yang Wang and Junwei Luo*

School of Software, Henan Polytechnic University, Jiaozuo, China

**Motivation:** In recent years, there have been significant advances in various chromatin conformation capture techniques, and annotating the topological structure from Hi-C contact maps has become crucial for studying the three-dimensional structure of chromosomes. However, the structure and function of chromatin loops are highly dynamic and diverse, influenced by multiple factors. Therefore, obtaining the three-dimensional structure of the genome remains a challenging task. Among many chromatin loop prediction methods, it is difficult to fully extract features from the contact map and make accurate predictions at low sequencing depths.

**Results:** In this study, we put forward a deep learning framework based on the diffusion model called CD-Loop for predicting accurate chromatin loops. First, by pre-training the input data, we obtain prior probabilities for predicting the classification of the Hi-C contact map. Then, by combining the denoising process based on the diffusion model and the prior probability obtained by pre-training, candidate loops were predicted from the input Hi-C contact map. Finally, CD-Loop uses a density-based clustering algorithm to cluster the candidate chromatin loops and predict the final chromatin loops. We compared CD-Loop with the currently popular methods, such as Peakachu, Chromosight, and Mustache, and found that in different cell types, species, and sequencing depths, CD-Loop outperforms other methods in loop annotation. We conclude that CD-Loop can accurately predict chromatin loops and reveal cell-type specificity. The code is available at https://github.com/wangyang199897/CD-Loop.

KEYWORDS

**chromatin loop, diffusion model, Hi-C contact map, three-dimensional structure, deep learning**

## 1 Introduction

The genome of eukaryotic organisms exists in the form of nuclear chromatin, and the function of chromatin is closely related to its three-dimensional structure. For example, biological functions such as genome replication, transcription, regulation, DNA mutation, the spread of long non-coding RNA, and embryonic development all are completed in the three-dimensional space of the cell nucleus (Bonev and Cavali, 2016). In recent years, with the development of high-throughput chromosome conformation capture (Hi-C) (Lieberman-Aiden et al., 2009) and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) technologies (Fullwood et al., 2009), researchers have discovered that chromosomes can be categorized into chromatin compartments (A/B compartments), topologically associated domains (TADs), and chromatin loops.

Chromatin loops, such as enhancer-promoter loops, explain the regulatory mechanism of enhancers on target genes. Despite the enhancer being far away from the target gene in linear distance, the enhancers and target gene promoters are located on the same chromatin loop in close spatial proximity, regulating the target gene by binding to the promoter (Dekker et al., 2013; Dixon et al., 2015; Gorkin et al., 2014; Rao et al., 2014; Dixon et al., 2012).

The chromatin loop is an advanced structural form of chromatin in eukaryotic organisms. In previous studies, chromatin loops could not be observed, but with the emergence of three-dimensional structures, it is now possible to clearly observe various organizations of genes. Experiments have shown that the chromatin loops are linked to proteins such as CTCF and cohesin. Two genes may be linearly distant from each other, but their spatial arrangement is not linear, and their spatial distance may be very close. Therefore, the two genes may interact with each other. They may be close in spatial proximity, potentially allowing for interactions between the two genes. We call the loop-like structure formed by two genes that are close together and the chromatin segment between them a chromatin loop. During the formation of cancer, the structure of chromatin loops may also undergo changes, leading to alterations in cancer-related genes (Wang et al., 2022). In genome-wide association studies (GWAS), it has been discovered that certain immune-related genetic variations are concentrated in chromatin loops specific to blood cells rather than embryonic cells, indicating that these chromatin loops can help us further understand certain disease variations (Buenrostro et al., 2013; Tang et al., 2015; Szabo et al., 2019; Grubert et al., 2020; Kloetgen et al., 2020).

Although some important progress has been made in the study of chromatin loops, the structure and function of chromatin loops are highly dynamic and diverse and are influenced by various factors. Currently, our comprehension of the structure and role of chromatin loops in the three-dimensional space of the cell nucleus remains limited, rendering it challenging to anticipate the consequences of alterations in the chromatin loop structure on gene mutations. Therefore, it remains a challenging problem to acquire the correlation between the three-dimensional architecture and functionality of the genome and use experimental techniques to detect chromatin loops in cell types or species with unknown 3D structures. At low coverage, due to the limited amount of data and the presence of random noise and biases, the detection of loops will be more challenging. Therefore, more accurate and efficient computational models and methods are needed to address these issues. This will help us better understand the organization, function, and gene regulation mechanisms of chromatin loops.

The methods for predicting chromatin loops are diverse, mainly encompassing the following aspects: (1) prediction of chromatin loops based on statistical methods. For high-throughput chromosome conformation capture (Hi-C), it focuses on the entire cell nucleus, studies the spatial relationships of the entire chromatin across the whole genome, and achieves the capture of interactions between chromatin segments across the entire genome. The corresponding tools are as follows: HiCCUPS (Rao et al., 2014; Durand et al., 2016) integrates nearby background information into its framework and employs a Poisson test in conjunction with an adapted Benjamini–Hochberg procedure to assess the significance of chromatin interactions. The HiCExplorer method (Wolff et al., 2022) uses ongoing negative binomial distribution and the Wilcoxon rank-sum test to ascertain the enrichment of Hi-C interactions by considering the neighborhood of candidate elements and distinguishing significant peaks from background noise. For ChIA-PET technology, using PET sequencing technology to study DNA fragments with nearby connections after immunoprecipitation allows researchers to obtain chromatin interactions; this fundamentally investigates the interactions between DNA fragments. The difference between Hi-C and ChIA-PET lies in the fact that data generated by Hi-C reflect chromatin interactions, including all proteins, while the ChIA-PET technology enrichment of specific protein factors results in data that represent chromatin interactions of a particular protein. Using ChIA-PET technology to develop tools includes the ChIA-PET tool (Li et al., 2010), which employs the hypergeometric distribution to filter noise. Mango software (Phanstiel et al., 2015) establishes a null model by merging the genomic distances and read depths for each anchor point. For the capture Hi-C technique (Mifsud et al., 2015), an additional capture step is introduced on top of the traditional Hi-C library preparation process to capture target fragments for subsequent sequencing. CHiCAGO (Cairns et al., 2016) employs an innovative background correction technique and a two-component convolution background model while addressing multiple testing through a *p*-value weighting approach. The ChiCMaxima method (Ben Zouari et al., 2019) applies loess smoothing to the captured Hi-C reads and transforms the detection of chromatin loops into the search for peaks from the loess-smoothed profiles. HiChIP (Mumbach et al., 2016) is a protein-centric approach for studying chromatin conformation, which synergistically combines Hi-C technology and ChIA-PET technology to extract more detailed three-dimensional chromatin structure information using a reduced dataset. Related tools include HiChIP-Peaks (Shi et al., 2020), which models the background signal as a negative binomial to simulate excessive dispersion and identify enriched signal regions. It also corrects HiChIP specific biases caused by the uneven distribution of restriction enzyme sites. (2) Prediction of chromatin loops based on traditional methods. Lollipop (Kai et al., 2018) is a machine learning framework based on the random forest classifier, which uses genomic and epigenomic features to predict CTCF-mediated interactions. CTCF-MP (Zhang et al., 2018), based on word2vec and boosted trees, accurately predicts loops formed by convergent CTCF motifs using sequence features, CTCF ChIP-seq and DNase-seq. C-Loops (Cao et al., 2020) relies on the clustering algorithm cDBSCAN, which directly examines paired-end tags (PET) to detect potential loops and employs permuted local backgrounds to estimate their significance. However, one of the recent trends in research is to apply computer vision and machine learning techniques to the annotation of topological structures. For example, the SIP method (Rowley et al., 2020) applies Gaussian smoothing, contrast adjustment, morphological white top-hat transformation, and a maximum–minimum filter to an image. After these steps, the corrected image of the interaction is provided, which is used in conjunction with the regional maxima detection algorithm to detect loops. Peakachu (Salameh et al., 2020) uses a classification framework to forecast chromatin loops based on the Hi-C contact map, capable of identifying a unique set of short-range interactions. Chromosight (Matthey-Doret et al., 2020) is a computer vision-based algorithm that takes whole-genome contact matrices as the input and uses a balancing normalization procedure to mitigate experimental biases. The Mustache method (Roayaei Ardakany et al., 2020) represents the interaction matrix using scale-space theory, and we consider the identification of chromatin loops as a problem of detecting spot-like objects. Both of these pattern-based general methods work well with a sufficient number of contact pairs but perform poorly at low sequencing depths. Due to the swift advancement and widespread utilization of deep learning technology, significant progress has been made in bioinformatics. It is not surprising that some work has been achieved in the field of genomics. For example, DeepLUCIA (Yang et al., 2022), a deep learning-based chromatin interaction model, utilizes epigenomic information to forecast chromatin loops in various tissues. The predicted chromatin loops can help enhance our understanding of the genomic structure of human tissues. DeepMILO (Trieu et al., 2020) uses a deep learning framework to anticipate the impacts of mutations on CTCF-mediated insulator loops. DeepLoop (Zhang et al., 2022) discovers noteworthy interactions from Hi-C contact maps using neural networks to denoise and enhance loop signals. RefHiC (Zhang and Blanchette, 2022) is a deep learning method that uses high-quality Hi-C datasets with different cell types to study the topological structure annotation of samples. GILoop (Wang et al., 2022) is a twin-branch neural network that utilizes the image view and graph view to identify interactions in the entire genome. Be-1DCNN (Wu et al., 2023) utilizes a bagging ensemble learning strategy and one-dimensional convolutional neural network (1DCNN) to improve the accuracy and reliability of predictions by integrating multiple 1DCNN models.

Although some progress has been made with the above methods, it remains a significant challenge to fully extract features from Hi-C contact maps and identify chromatin loops in different sequencing depths and cell lines. Recently, the denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) has had good performance in image generation and synthesis tasks by progressively enhancing the quality of the provided image. Furthermore, the diffusion model is able to simulate the propagation and influence of features in an image, which helps better capture local and global features, thus improving classification accuracy (Han et al., 2022). Here, we propose a method, named CD-Loop, based on the diffusion model, which combines CTCF ChIA-PET and H3K27ac HiChIP (Mumbach, 2017) data derived from biologically diverse experiments to label samples. This approach aims to cover a wider range of chromatin loops. Using a conditional generative model based on noise addition and noise reduction, along with a pre-trained conditional mean estimator, we convert the task of identifying chromatin loops into a binary classification task. The results indicate that training the data only on the original sequencing depth is effective for different cell types, sequencing depths, and species with high precision and recall. In comparison to existing methods, our approach successfully identifies a set of distinct chromatin loops.

## 2 Materials and methods

CD-Loop takes a Hi-C contact map as the input and predicts highly reliable chromatin loops. The model can be roughly divided into two parts, as shown in Figure 1. (1) First, CD-Loop pre-trains the input data using the LeNet5 model to obtain the prior probability of predicted classification for the input Hi-C contact map. (2) Then, by combining the denoising process based on the diffusion model and the prior probability obtained by pre-training to predict the candidate chromatin loop, the output includes the probability score, CI confidence, and two-tailed t-test evaluation metrics for each candidate chromatin loop. (3) Finally, the low-scoring candidate chromatin loop is filtered out, and then clustering is performed based on the density algorithm to select representative chromatin loops.

### 2.1 Pretraining phase

The CD-Loop network first applies a pre-trained conditional mean estimator, utilizing the LeNet5 network, with an input of dimension of 2w × 2w, where w is the window size (w = 14). This module consists of two convolutional blocks, three fully connected layers, and two ReLU layers. Each block includes a convolution operation, batch normalization, a ReLU activation function, and an average pooling operation. The input and output of this process are class labels, which we refer to as prior probabilities.

### 2.2 Diffusion model

The second part of the model applies the forward and backward processes of the diffusion model, assuming that the endpoint of our forward process is

$$p\left(y_T \middle| x\right) = \mathcal{N}\left(f_{\varnothing}(x), I\right),$$

where $f_{\varnothing}(x)$ is the prior probability with respect to x and $y_0$. The conditional distribution of the forward process can be defined as follows for all timesteps including t = 1:

$$q\left(y_t \middle| y_{t-1}, f_{\varnothing}(x)\right) = \mathcal{N}\left(y_t; \sqrt{1-\beta_t}\, y_{t-1} + \left(1 - \sqrt{1-\beta_t}\right) f_{\varnothing}(x), \beta_t I\right),$$

which enables a closed sampling distribution with arbitrary timesteps t:

$$q\left(y_t \middle| y_0, f_{\varnothing}(x)\right) = \mathcal{N}\left(y_t; \sqrt{\bar{\alpha}_t}\, y_0 + \left(1 - \sqrt{\bar{\alpha}_t}\right) f_{\varnothing}(x), \left(1 - \bar{\alpha}_t\right) I\right),$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \Pi_t \alpha_t$. The backward process can be defined as follows:

$$\begin{aligned} p\left(y_{t-1} \middle| y_t, y_0, x\right) &= p\left(y_{t-1} \middle| y_t, y_0, f_{\varnothing}(x)\right) \\ &= \mathcal{N}\left(y_{t-1}; \tilde{\mu}\left(y_t, y_0, f_{\varnothing}(x)\right), \tilde{\beta}_t I\right), \end{aligned}$$

where

$$\begin{aligned} \tilde{\mu} &:= \underbrace{\frac{\beta_t \sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_t}}_{\gamma_0} y_0 + \underbrace{\frac{\left(1 - \bar{\alpha}_{t-1}\right)\sqrt{\alpha_t}}{1 - \bar{\alpha}_t}}_{\gamma_1} y_t \\ &\quad + \underbrace{\left(1 + \frac{\left(\sqrt{\bar{\alpha}_t} - 1\right)\left(\sqrt{\alpha_t} + \sqrt{\bar{\alpha}_{t-1}}\right)}{1 - \bar{\alpha}_t}\right)}_{\gamma_2} f_{\varnothing}(x), \\ \tilde{\beta}_t &:= \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t. \end{aligned}$$

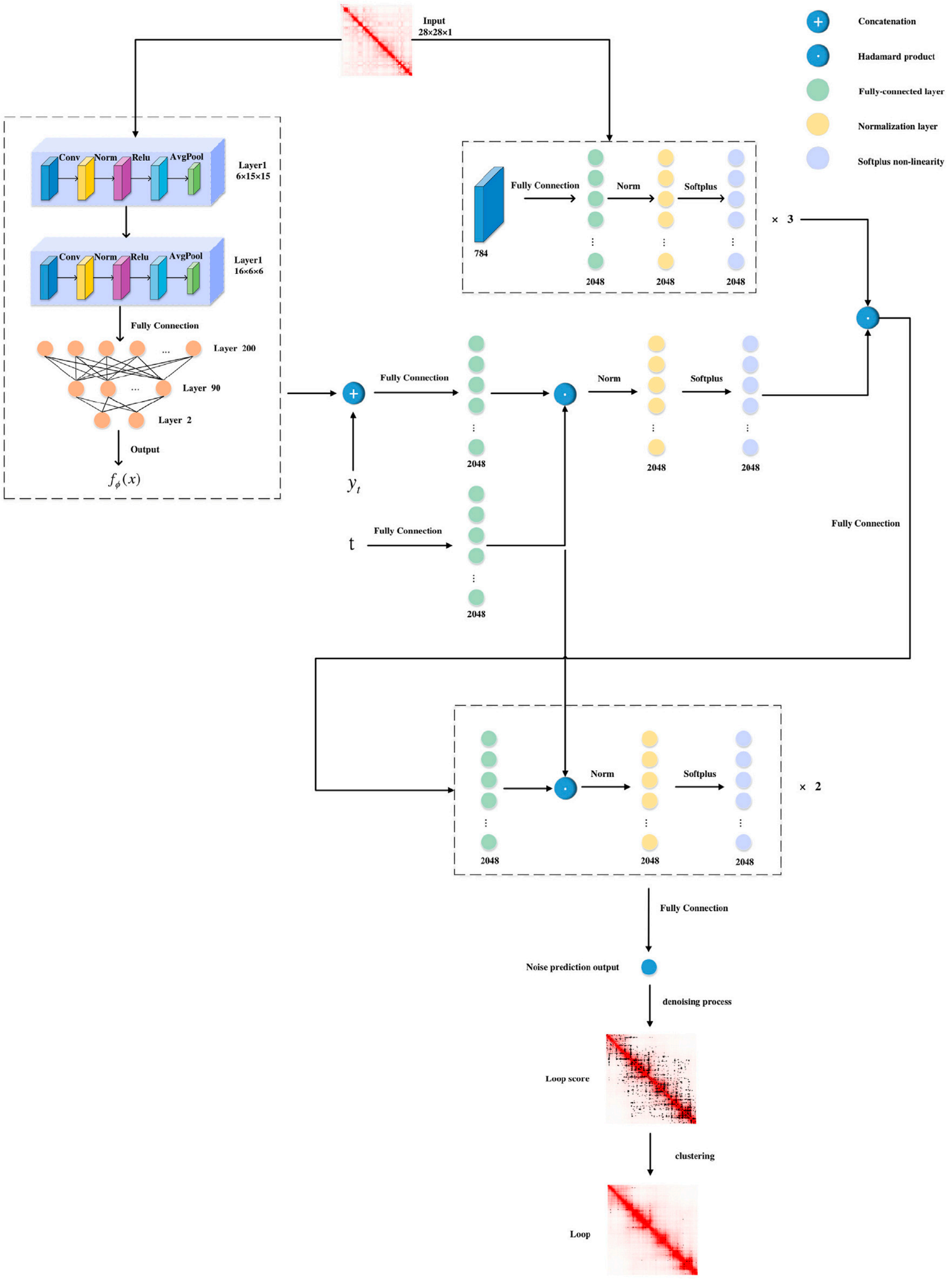


FIGURE 1
CD-Loop architecture. Overview of the CD-Loop neural network for loop.

After organizing the process, the optimization objective is the maximization of the likelihood function:

$$\log \mathrm{p}_\theta\left(y_0 \middle| x\right) = \log \int \mathrm{p}_\theta\left(y_{0:\,T} \middle| x\right) d y_{1;T} \geq \mathbb{E}_{q\left(y_{1:\,T} \middle| y_0, x\right)} \left[\log \frac{\mathrm{p}_\theta\left(y_{0:\,T} \middle| x\right)}{q\left(y_{1:\,T} \middle| y_0, x\right)}\right].$$

We chose the diffusion model as the second part of our model because through the iterative process of the diffusion model, the noise in the Hi-C interaction matrix can be corrected, and the quality is improved. In addition, the diffusion model can simulate the spread and influence of features in the image, helping better capture the local and global features in the Hi-C interaction matrix, thereby improving the accuracy of the classification.

This part of the model framework applies an encoder to a flattened input image to obtain a 2048-dimensional representation. The encoder consists of three fully connected layers, with an output size of 2048. Meanwhile, we concatenate $\mathrm{y}_\mathrm{t}$ and the output $\mathrm{f}_\varnothing\,(\mathrm{x})$ from the first part, applying a fully-connected layer to generate an output vector of 2048 dimensions. To incorporate the timestep information, we apply a Hadamard product between the vector and timestep embedding, generating a response embedding specific to the timestep. Next, we integrate this response embedding with the image embedding through another Hadamard product. The resulting vector is then passed through two additional fully-connected layers. Each layer has 2048 dimensions. Before each layer, a Hadamard product is performed with timestep embedding. At last, a fully-connected layer is employed to predict. It is worth noting that, in addition to the output layer, there is also a batch normalization layer and Softplus non-linearity after each fully connected layer. The architecture is shown in Table 1.

The model framework outputs a noise prediction and then utilizes a denoising process, combined with pre-trained prior probabilities, to obtain the posterior mean and posterior variance. Based on the obtained posterior mean and posterior variance, the predicted label at time T-1 is calculated from time T, and this process is repeated until time 1.

**TABLE 1 Diffusion model network architecture.**

| |
|---|
| input: $\mathrm{x}, \mathrm{y}_\mathrm{t}, \mathrm{f}\varphi\,(\mathrm{x}), \mathrm{t}$ |
| $\mathrm{l}_{1,\mathrm{x}} = \sigma(\mathrm{BN}(\mathrm{g}_{1,\mathrm{x}}(\mathrm{x})))$ |
| $\mathrm{l}_{2,\mathrm{x}} = \sigma(\mathrm{BN}(\mathrm{g}_{2,\mathrm{x}}(\mathrm{x})))$ |
| $\mathrm{l}_{3,\mathrm{x}} = \mathrm{BN}(\mathrm{g}_{1,\mathrm{x}}(\mathrm{x}))$ |
| $\mathrm{l}_{1,\mathrm{y}} = \sigma(\mathrm{BN}(\mathrm{g}_{1,\mathrm{y}}(\mathrm{y}_\mathrm{t} \oplus \mathrm{f}_\varnothing(\mathrm{x})) \odot \mathrm{g}_{1,\mathrm{b}}(\mathrm{t})))$ |
| $\mathrm{l}_1 = \mathrm{l}_{3,\mathrm{x}} \odot \mathrm{l}_{1,\mathrm{y}}$ |
| $\mathrm{l}_2 = \sigma(\mathrm{BN}(\mathrm{g}_{2,\mathrm{a}}(\mathrm{l}_1) \odot \mathrm{g}_{2,\mathrm{b}}(\mathrm{t})))$ |
| $\mathrm{l}_3 = \sigma(\mathrm{BN}(\mathrm{g}_{3,\mathrm{a}}(\mathrm{l}_2) \odot \mathrm{g}_{3,\mathrm{b}}(\mathrm{t})))$ |
| output: $\mathrm{g}_4(\mathrm{l}_3)$ |

⊕: concatenation; ⊙: Hadamard product; σ: Softplus non-linearity; g: a fully-connected layer; and l: a hidden layer output.

## 2.3 Detect loops by density-based clustering

For the window centered around each bin pair (i, j) after model prediction, CD-Loop generates a probability score s (i, j) for each bin pair. A higher score value indicates a higher likelihood of the bin pair being a loop. Therefore, we retain bin pairs that are predicted as loops and have a score greater than 0.5, and these bin pairs (i, j) are referred to as candidate loops. If there are fewer than 15 candidate loops within a 5-bin by 5-bin square centered around (i, j), it is referred to as an isolated prediction. These isolated predictions are likely to be false positives and are therefore excluded. Then, we use a density-based clustering algorithm to cluster the remaining candidate loops. First, we use the nearest neighbors (Abeywickrama et al., 2016) method to compute the local density of each candidate (i , j). To achieve a fast nearest neighbor search, we use the K-D tree data structure, and the distance metric used is Chebyshev distance. We then calculated and recorded the indices and distances of the nearest neighbors for each candidate (i, j). By iterating over the nearest neighbors of each candidate (i, j), find the nearest neighbor with a higher density than itself. If the nearest neighbor with a higher density than the current point is found, we record its index and distance as the delta value. If no such point is found, meaning that the candidate (i, j) has the highest local density within the current range, we set the delta value to a distance greater than that of the neighboring nodes. We repeat this process, increasing the query radius until the nearest neighbors of all candidate (i, j) pairs are found. Finally, we discard candidate loops with delta values less than 5, as they may represent redundant predictions. The remaining candidate loops after filtering constitute our final predicted loops. The same parameters are used in different datasets and different coverages, and these parameters perform well in the final prediction.

## 2.4 Composition of training samples

Selection of positive samples: CD-Loop selects the combination of CTCF ChIA-PET data and H3K27ac HiChIP data and then removes all interaction pairs outside the range of 30 kb to 3 Mb as positive sample data. Because in loop annotation, CTCF ChIA-PET data contain long-range interactions, while H3K27ac HiChIP data contain shorter-range interactions, combining the two can cover a wider range of loop types.

Selection of negative samples: Due to a large unbalance in the number of positive and negative samples, we selected different types of negative samples three times based on the genomic distance characteristics of the positive samples, ensuring that the genomic distance of each negative sample falls within the range of 30 kb to 3 Mb. This can reduce the number of negative samples and consider all the characteristics of negative samples as much as possible. (1) For each positive interaction pair, two negative interaction pairs with the same genomic distance are randomly selected from the entire genome. (2) For all possible genomic distances of positive samples, randomly select a genomic distance each time, and then a negative interaction pair with the same genomic distance is randomly selected from the entire genome until the number of generated negative samples equals that of positive samples. (3) For the largest genomic distance among positive samples, a value greater

TABLE 2 Different sources of datasets.

| Deposited data | Source | Identifier | Link |
|---|---|---|---|
| GM12878 Hi-C | GEO | GSE63525 | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525 |
| K562 Hi-C | GEO | GSE63525 | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525 |
| IMR90 Hi-C | GEO | GSE63525 | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525 |
| mESC Hi-C | 4D Nucleome | 4DNFIU8AF5ZY | https://data.4dnucleome.org/experiment-set-replicates/4DNESUCLJAZ8 |
| GM12878 CTCF ChIP-Seq | ENCODE | ENCFF963PJY | https://www.encodeproject.org/files/ENCFF963PJY |
| K562 CTCF ChIP-Seq | ENCODE | ENCFF085HTY | https://www.encodeproject.org/files/ENCFF085HTY |
| IMR-90 CTCF ChIP-Seq | ENCODE | ENCFF453XKM | https://www.encodeproject.org/files/ENCFF453XKM |
| mESC CTCF ChIP-Seq | ENCODE | ENCFF508CKL | https://www.encodeproject.org/files/ENCFF508CKL |
| K562 CTCF ChIA-PET | ENCODE | ENCFF001THV | https://www.encodeproject.org/files/ENCFF001THV |
| K562 RAD21 ChIA-PET | ENCODE | ENCFF002ENT | https://www.encodeproject.org/files/ENCFF002ENT |
| mESC CTCF ChIA-PET | ENCODE | ENCFF550QMW | https://www.encodeproject.org/files/ENCFF550QMW |
| GM12878 CTCF ChIA-PET | Reference (Tang et al., 2015) | Tang, Z. et al. (2015) | https://doi.org/10.1016/j.cell.2015.11.024 |
| GM12878 RAD21 ChIA-PET | Reference (Heidari et al., 2014) | Heidari et al. (2014) | https://doi.org/10.1101/gr.176586.114 |
| GM12878 H3k27ac HiChIP | Reference (Mumbach et al., 2017) | Mumbach et al. (2017) | https://doi.org/10.1038/ng.3963 |
| GM12878 SMC1 HiCHIP | Reference (Mumbach et al., 2016) | Mumbach et al. (2016) | https://doi.org/10.1038/nmeth.3999 |

than that distance is randomly chosen as the genomic distance for negative samples, and a negative interaction pair is randomly selected from the entire genome until the number of generated negative samples equals that of positive samples.

We selected positive and negative samples within the gene distance range of 30 kb to 3 Mb for the following two reasons: first, interactions that are far apart in the genome are more likely to have sequencing errors, resulting in chromatin loops that are detected in distant genomes having high error rates. Interactions that are relatively close together are generally caused by physical interactions rather than true chromatin loops. Second, in other chromatin loop prediction methods, most of the predicted chromatin loops are in the range of 30 kb to 3 Mb. So, considering these two factors, we chose the range of 30 kb to 3 Mb to filter other samples.

Data preprocessing: The input Hi-C contact map is divided into bins at a resolution of 5 kb. Due to the existing sampling bias and technical noise, Knight–Ruiz (KR) normalization (Knight and Ruiz, 2013) is used for correction. If a positive sample is represented by two or more pixels in the contact map, each pixel represents a positive interaction pair. After obtaining the positive and negative interaction pairs, in the Hi-C contact map, with each interaction pair as the center, 13 bins are selected upward and to the right, and 14 bins are selected to the right and downward, forming a 28*28 matrix. The matrices from the negative sample matrix that consist entirely of zero elements are removed. The reasons why we delete matrices with all 0 elements in the negative sample matrix are as follows: first, consider that in the Hi-C interaction matrix, all elements are 0, which means that there is no interaction between one region of the chromosome and another region. Such a matrix does not contain any meaningful information. Second, the model cannot extract effective features from these matrices with all 0 elements. There are too many such matrices in the training set, which will only increase the training time.

## 2.5 Model training and prediction

During the model training and testing process, Hi-C contact maps from the GM12878 dataset are adopted. For validation, we used chr11 and chr12, while chr15, chr16, and chr17 were used for testing and prediction. The remaining chromosomes were used for training. During the model training process, data augmentation was performed by flipping the 28*28 matrices generated from the positive samples horizontally and vertically. In the model prediction process, since it involves taking every bin pair of an entire chromosome as the input, the amount of data is very large. Therefore, we performed the following three preprocessing steps on the chromosome to be predicted. (1) The genomic distance threshold between bin pairs: since the distribution of chromatin loops in the genome ranges from approximately 30 kb to 3 Mb, we remove the predicted bin pairs with a genomic distance greater than 3 Mb or less than 30 kb. (2) Interaction frequency threshold of bin pairs: by observing the interaction frequency of each bin pair in the positive samples, we found that 99% of the positive samples have an interaction frequency greater than 1. Therefore, we remove the bins with an interaction frequency less than 1. (3) Threshold of the number of zero elements in matrices: after counting the number of zero elements in the 28*28 matrices of positive samples, it was found that 90% of the positive sample matrices have less than 200 zero elements. When making predictions for downsampling data, the same processing is applied.

# 3 Results

CD-Loop is trained on the original sequencing depth of GM12878, which is not specific to this sequencing depth or cell type. Next, we will demonstrate the superiority of this model by

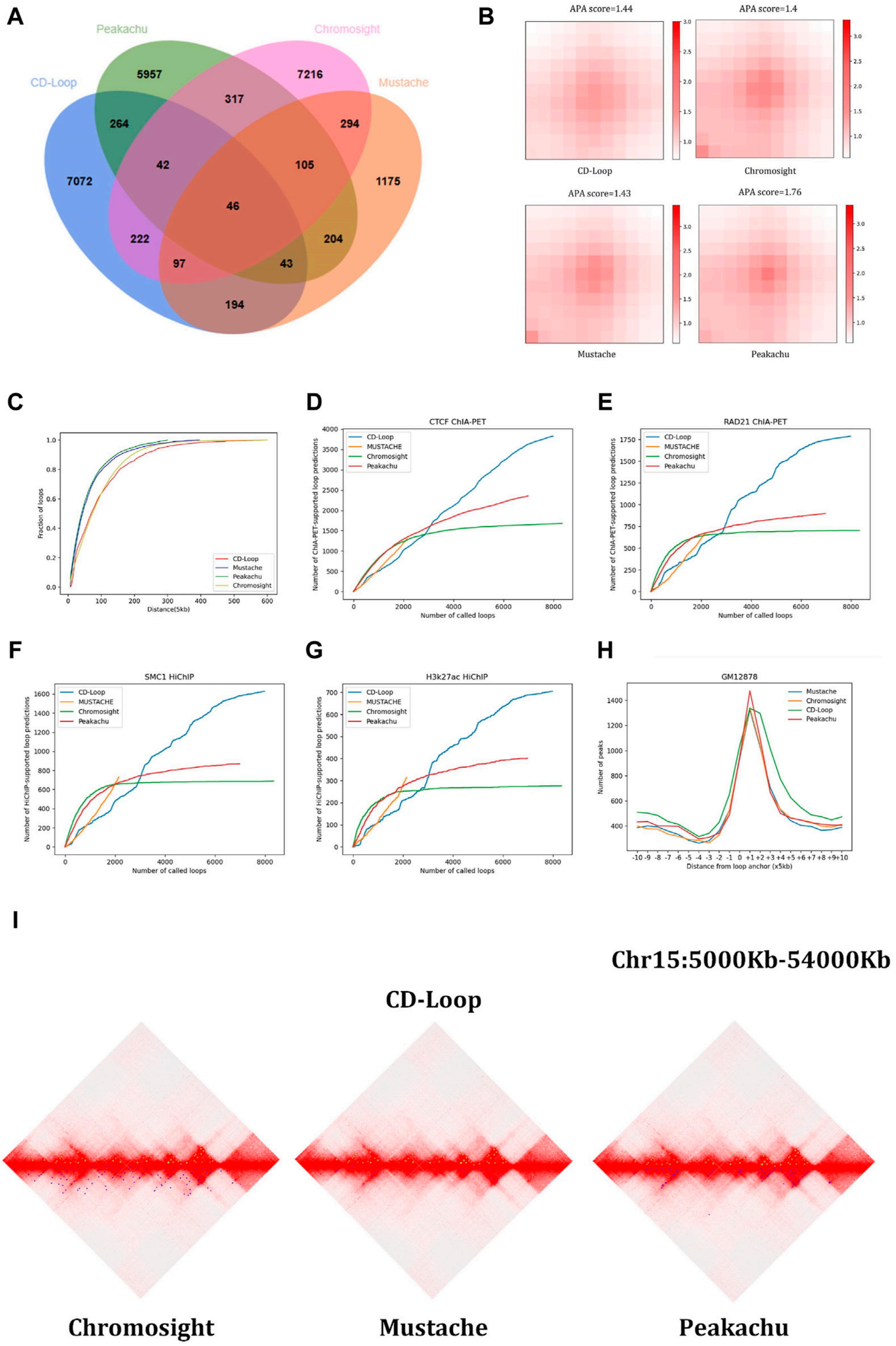


FIGURE 2
Comparison based on the GM12878 dataset: **(A)** Venn diagram, **(B)** aggregated peak analysis, and **(C)** cumulative distance distribution. The chromatin loop genome distance distribution predicted by CD-Loop is highly similar to that predicted by Chromosight. **(D–G)** Supporting loops validated by CTCF ChIA-PET **(D)**, RAD21 ChIA-PET **(E)**, SMC1 HiChIP **(F)**, and H3k27ac HiChIP **(G)** enrichment experiments for loops predicted by CD-Loop and other tools. The loop predictions by CD-Loop align better with these experimental data compared to other tools on the testing chromosomes. **(H)** Function depicting the distance from predicted loop anchors to CTCF-binding sites identified by ChIP-seq signals. **(I)** Visualization example of loop identification. The upper half of the three diamond plots display green dots, which represent CD-Loop. On the other hand, the lower half of the plots consists of blue dots, which represent Chromosight, Mustache, and Peakachu, respectively.

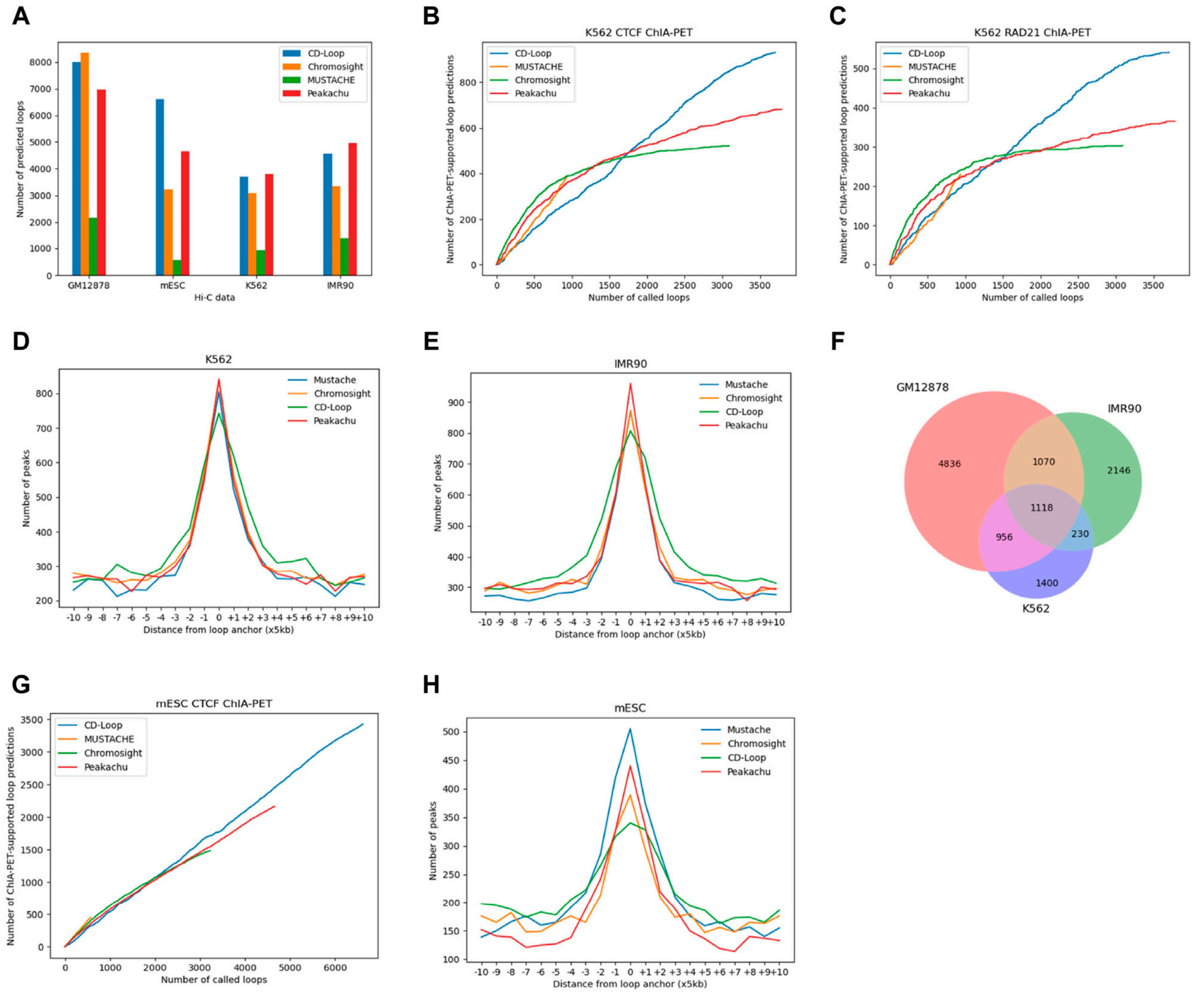


FIGURE 3
Comparison based on Hi-C data from human K562, IMR90, and mouse ESC. **(A)** The number of loops present. **(B–C)** Overlap between the chromatin loops predicted by CD-Loop and other tools on the K562 Hi-C contact map with CTCF ChIA-PET **(B)** and RAD21 ChIA-PET **(C)** enrichment experiments on testing chromosomes chr15-17. **(D)** Function depicting the distance from predicted loop anchors to CTCF-binding sites in K562 cells identified by ChIP-seq signals. **(E)** Function depicting the distance from predicted loop anchors to CTCF-binding sites in IMR90 cells identified by ChIP-seq signals. **(F)** Venn diagram showing the chromatin loops identified by CD-Loop across three cell lines: GM12878, IMR90, and K562. **(G)** Overlap between the chromatin loops predicted by CD-Loop and other tools on the mESC with CTCF ChIA-PET. **(H)** Function depicting the distance from predicted loop anchors to CTCF-binding sites in mESC cells identified by ChIP-seq signals.

demonstrating the ability of the same model weights to annotate loops in the same cell type at different sequencing depths. Meanwhile, we will highlight that the model can accurately annotate loops in numerous other cell types without requiring retraining. Furthermore, the same trained model can also be used to annotate loops in mouse Hi-C contact maps. This demonstrates that the CD-Loop model has good generalization and adaptability and is applicable to annotation tasks of different sequencing depths and cell types. In our experiment, chromosomes 11 and 12 are used for validation, chromosomes 15–17 are used for testing, and the remaining chromosomes are used for training. The reported human gene results only apply to the three test chromosomes, while the results for mice apply to all chromosomes.

This method is trained and predicted on RTX4090 GPU and requires at least 15 GB of space to load samples during prediction. The runtime of chromatin loop identification depends on the sequencing depth of Hi-C data. For example, prediction can be completed within 325 min on Hi-C data containing 500 M valid read pairs. The data used in the experiments of this paper are shown in Table 2.

## 3.1 GM12878 experimental results

We first evaluated the prediction accuracy of chromatin loops by CD-Loop on the original sequencing depth Hi-C dataset (2600 M valid read pairs) from the human GM12878 cell line. Simultaneously, we compared it with several popular methods, including Chromosight, Peakachu, and Mustache. To ensure a fair comparison, we evaluated the chromatin loops at a 5 kb resolution and annotated them from the same dataset using default parameters. Additionally, we applied a consistent 5% FDR cutoff for all tools.

### 3.1.1 Quantitative analysis

Loops predicted by different tools vary significantly. First, among the four methods, we controlled the genomic distance of all predicted chromatin loops between 30 kb and 3 Mb. To compare the performance with all methods, we ensured that the anchoring points of two chromatin loops fully matched and overlapped. As shown in Figure 2A, we found that CD-Loop identified almost as many chromatin loops as Chromosight and Peakachu. For the results, 89% of CD-Loop, 87% of Chromosight, 85% of Peakachu, and 54% of Mustache were unique.

### 3.1.2 Aggregation peak analysis

The aggregation peak analysis of the four methods in the GM12878 cell line is shown in Figure 2B. The APA score quantifies the loop pattern of detected peaks by comparing the number of reads in the center point bin to the average number of reads in the lower left corner of the matrix. We only consider the top 2,000 loops with high scores. Compared with Peakachu, the loops detected by Chromosight, CD-Loop, and Mustache have more dispersed loop centers, and the three methods have a similar APA score. Since we split the chromatin loop spanning multiple pixel points into multiple chromatin loops, each pixel is regarded as a single chromatin loop, and the most representative pixel among them is not selected as a positive sample. So, the boundary range of chromatin loops is expanded, and the detected chromatin loops have dispersed loop centers. Next, we compared the genomic distances of loop anchors predicted by the four methods. As shown in Figure 2C, the distance distributions between chromatin loop anchors predicted by CD-Loop and Chromosight are similar and have larger genetic distances, while Peakachu and Mustache predict more short-range interactions.

### 3.1.3 Enrichment experimental analysis

Then, we compared the chromatin loops predicted by different methods on different datasets. Different enrichment experimental data include CTCF ChIA-PET, RAD21 ChIA-PET, SMC1 HiChIP, and H3k27ac HiChIP. We make predictions for the three chromosomes chr15, chr16, and chr17 of the test set and compare them with three other methods to evaluate these loops (allowing an error of 5 kb). As shown in Figures 2D–G, among the four methods, Mustache predicts the smallest number of loops and the least number of overlaps with enrichment experiments. The remaining three methods predict almost the same number of loops, but CD-Loop has the largest number of correct predictions across different enrichment experiments and has the highest recall rates. CD-Loop predicted a total of 7,980 loops, with 3,821 correctly predicted loops in the CTCF dataset, yielding an accuracy of 48%. Peakachu predicted 6,978 loops, with 2,351 correctly predicted loops in the CTCF dataset, resulting in an accuracy of 33%. Chromosight predicted 8,993 loops, with 1,676 correctly predicted loops in the CTCF dataset, giving an accuracy of 20%. Mustache predicted 2,158 loops, with 1,293 correctly predicted loops in the CTCF dataset, achieving the highest accuracy of 59%. CD-Loop had the highest number of successfully predicted loops, ranking second in accuracy, while Mustache, with the highest accuracy, had the fewest successfully predicted loops, only one-third of CD-Loop's count. The specific data for RAD21, SMC1, and H3K27ac can be found in Figures 2E–G, where CD-Loop demonstrates good performance in both accuracy and recall.

### 3.1.4 CTCF-binding site analysis

We next performed this by visualizing the CTCF ChIP-Seq and H3k27ac HiChIP-binding signals on the flanking regions around the loop anchors. As shown in Figure 2H, the predicted loop anchors detected by CD-Loop showed a clear enrichment effect in CTCF, and the H3k27ac-binding motif proves that CD-Loop can not only identify loops related to CTCF but also loops related to H3k27ac.

### 3.1.5 Hi-C heat map analysis

The genome-wide analysis described above demonstrates the good ability of CD-Loop to identify loops in Hi-C contact maps. We used the Juicebox tool (Durand et al., 2016) to visualize chromatin loops for the purpose of visual representation, demonstrating that CD-Loop can detect more chromatin loops and unique chromatin loops undetectable by other methods. As shown in Figure 2I, the upper part is the visual representation of chromatin loops detected by CD-Loop in the Hi-C interaction matrix, and the lower part is the detection of the remaining three methods (Peakachu, Chromosight, and Mustache) for the visual representation of chromatin loops in the Hi-C interaction matrix. The green dots represent the position of the chromatin loop detected by CD-Loop in the Hi-C interaction matrix, and the blue dots represent the position of the chromatin loop detected by Peakachu, Chromosight, and Mustache in the Hi-C interaction matrix. We can find that the results of CD-Loop mostly overlap with the results of other methods, but some are unique.

Taken together, these results show that CD-Loop has better overall prediction accuracy for GM12878 data (2600 M read pairs) than other methods.

## 3.2 Experimental results on other cells and species

Our method was trained on the original test depth data of the human GM12878 cell line, but our findings reveal that the trained model demonstrates better performance across various cell types. To further verify the performance of CD-Loop, we compared CD-Loop and other methods using K562 and IMR90 cell lines from humans (only chromosomes 15–17 test) and mouse embryonic stem cells (mESCs) (all chromosomes).

### 3.2.1 Number analysis

As shown in Figure 3A, in different cell lines of both humans and mice, CD-Loop and Peakachu predicted the most loops, indicating that the CD-Loop method is more reliable for predicting chromatin loops, regardless of sequencing coverage. When applied to the complete set of autosomes with 124 M read pairs from mESC data, the CD-Loop model trained with the GM12878 original sequencing depth was used to predict mouse cell lines. CD-Loop identified a higher number of loops compared to other tools in low-coverage data.

### 3.2.2 K562 and IMR90 cell line enrichment experiments and CTCF-binding site analysis

Different enrichment experiments were used to reveal loops and evaluate the accuracy of these tools. Like GM12878 before, we controlled the FDR of other methods to 5%. For K562 data, as shown in Figure 3B, CD-Loop predicted a total of 3,704 loops in the K562 cell line, with 1,169 correctly predicted loops in the CTCF dataset, yielding an accuracy of 32%. Peakachu predicted 3,785 loops, with 830 correctly predicted loops in the CTCF dataset, resulting in an accuracy of 22%. Chromosight predicted 3,089 loops, with 665 correctly predicted loops in the CTCF dataset, achieving an accuracy of 22%. Mustache predicted 939 loops, with 543 correctly predicted loops in the CTCF dataset, attaining the highest accuracy of 58%. CD-Loop had the highest number of successfully predicted loops and ranked second in accuracy, while Mustache, with the highest accuracy, had the fewest successfully predicted loops, only half of CD-Loop's count. Specific data for RAD21 can be found in Figure 3C. CD-Loop has advantages over other tools, being able to identify more loops supported by CTCF and RAD21, demonstrating good performance in both precision and recall.

Stacking analysis of surrounding CTCF-binding sites at predicted chromatin loop anchor locations is shown in Figure 3D, indicating that the chromatin loops predicted by these four methods are rich in CTCF-binding motifs and have little difference, indicating that the same training model can not only identify CTCF motifs in GM12878 cell lines but also be applicable in K562 cell lines. Similar results were obtained on IMR90 data (Figure 3E). Whether it is the K562 cell line or IMR90 cell line, the number of CTCF-binding sites at the left and right anchor points of the chromatin loop detected by CD-Loop is less than that of the other three methods. However, the number of CTCF-binding sites present was higher than other methods, within a 50-Kb range of the left and right anchor points of the chromatin loop. The reason may be that we represent a chromatin loop connected by multiple pixels as a single pixel as a single chromatin loop. Due to the expansion of the range of the anchor point of the chromatin loop, the CTCF-binding sites present on the anchor point are also within a certain range float.

### 3.2.3 Specificity analysis

In addition, to further illustrate the differences between cell lines, we conducted a comparison of chromatin loop overlap among three cell lines; to enhance fault tolerance, we allowed partial matches (±5 kb) between any anchors in two bins. As shown in Figure 3F, even when the overlap range was increased, the extent of the chromatin loop overlap was relatively low among the three cell lines, suggesting that the chromatin loops are specific to each cell type.

### 3.2.4 mESC cell line enrichment experiments and CTCF-binding site analysis

Enrichment experiments and CTCF ChIP-Seq signal analysis for mESC data are shown in Figure 3G, H. In the CTCF ChIA-PET enrichment experiment, CD-Loop predicted the most loops among the four methods, and the number of overlaps increased linearly with the increasing number of predicted loops. The number of perfectly matched CTCF-binding sites is slightly lower compared with the other three methods, but the number of CTCF-binding sites around the anchor fluctuation range was higher than the other three methods.

In conclusion, the research results show that CD-Loop has achieved superior performance in human K562 and IMR90 cell lines and mouse cell types.

## 3.3 Experimental results at different sequencing depths

### 3.3.1 Quantity, F1-score, and enrichment experimental analysis

To evaluate the ability of CD-Loop at different sequencing depths, we conducted downsampling experiments using the FAN-C method (Kruse et al., 2020)on the original sequencing depth of 2,600 M valid read pairs. We performed downsampling at various percentages, including 90%, 70%, 50%, 20%, and 10%. The corresponding effective read pairs for each downsampling were 2,300 M, 1,800 M, 1,200 M, 500 M, and 250 M. Using default parameters for different loop prediction tools, we observed a decrease in predicted chromatin loops as the sequencing depth decreased, as shown in Figure 4A. CD-Loop and Peakachu predicted the highest number of chromatin loops. However, in enrichment experiments (Figures 4C–F), CD-Loop consistently achieved the highest F1-score among the four methods. The F1-score decreased with decreasing sequencing depth but remained at its highest level. The enrichment experiments for different methods at different sequencing depths are shown in Supplementary Figure S1 in Supplementary Material.

### 3.3.2 Robustness analysis

We assessed the overlap between the loops predicted by CD-Loop at different sequencing depths and the loops present in the original sequencing depth matrix. As depicted in Figure 4B, the overlap rates were 76%, 68%, 66%, 54%, and 51% for downsampling matrices with 2,300 M, 1,800 M, 1,200 M, 500 M, and 250 M valid read pairs, respectively. This high overlap rate indicates that CD-Loop not only predicts a significant number of chromatin loops but also detects more loops at low sequencing depth. Moreover, it will not cause more false positives, highlighting the robustness of CD-Loop.

### 3.3.3 Sensitivity analysis

CD-Loop efficiently identifies a significant quantity of loop structures within sparse data without increasing the number of false positives. As shown in Figures 4G–J, when evaluating loops mediated by CTCF, RAD21, SMC1, and H3K27ac in low-depth datasets, CD-Loop maintains a high level of accuracy. This implies that the predictions made for low-sequencing depth data are almost as accurate as predictions on complete data, with lower sensitivity.

Overall, CD-Loop outperforms other tools in terms of accuracy at all sequencing depths. These results highlight the superior robustness, accuracy, and reliability of CD-Loop.

## 3.4 Hyperparameters and resolution analysis

In order to prove the generalization ability of CD-Loop, we conducted different experiments on the three hyperparameters of the optimizer, batch size and epoch, and used chromosome 15 as the test set to verify the optimal hyperparameters of the model. The

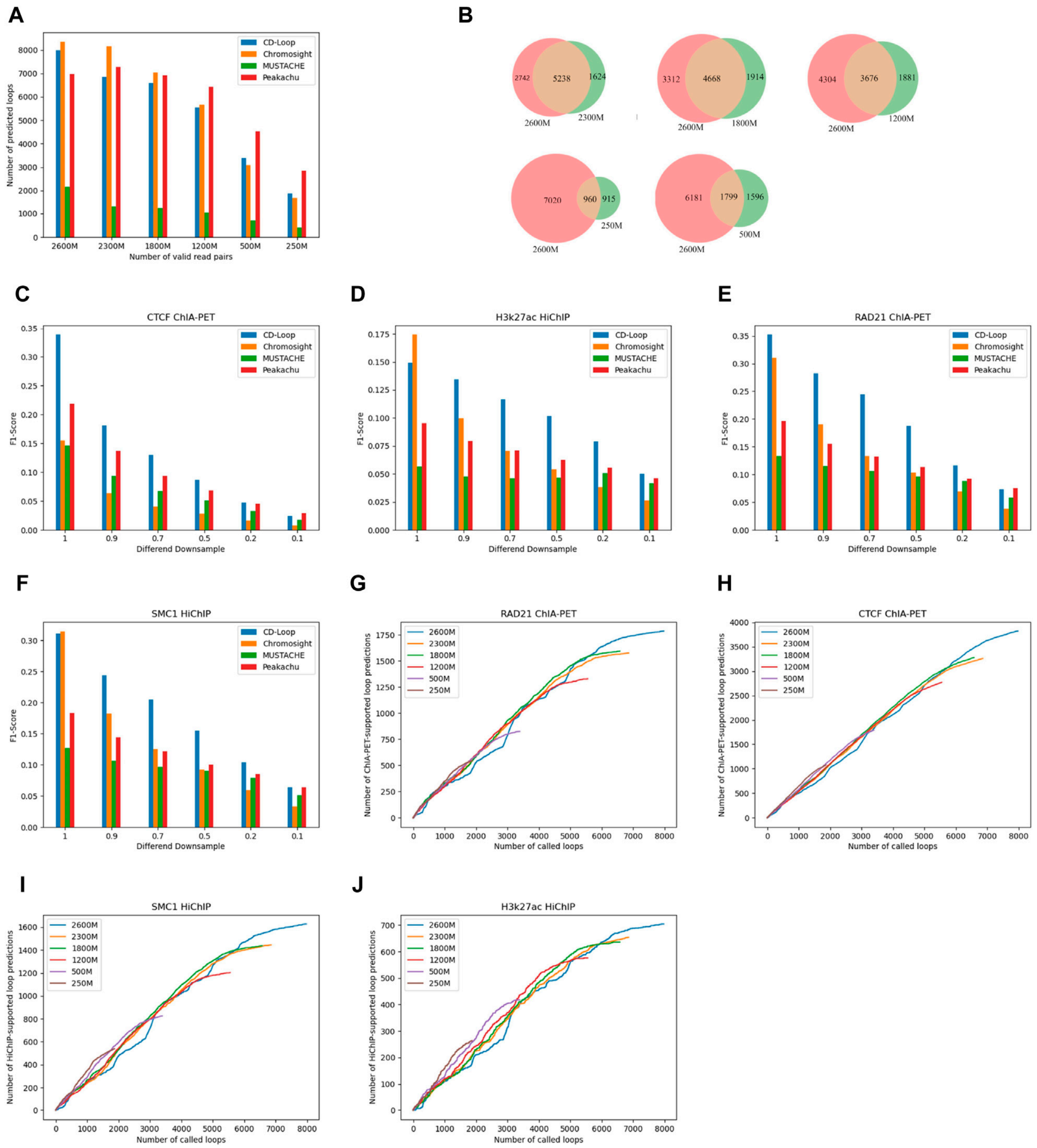


FIGURE 4
Evaluation at different sequencing depths. **(A)** The number of chromatin loops predicted by different methods decreases as the number of effective chromosome read pairs decreases. **(B)** Venn diagrams at different sequencing depths. **(C–F)** F1-scores of different enrichment experiments, including CTCF ChIA-PET **(C)**, H3k27ac HiChIP **(D)**, RAD21 ChIA-PET **(E)**, and SMC1 HiChIP **(F)**, in GM12878 cells at different sequencing depths. **(G–J)** Number of supports on different enrichment data [RAD21 ChIA-PET **(G)**, CTCF ChIA-PET **(H)**, SMC1 HiChIP **(I)**, and H3k27ac HiChIP **(J)**] for predicted chromatin loops at different sequencing depths.

experimental results are shown in Supplementary Tables S1–S3 in Supplementary Material.

In addition, we used chr15 as the test set and conducted experiments at three resolutions: 5 KB, 10 KB, and 25 KB. The experimental results are shown in Supplementary Table S4 in Supplementary Material.

## 4 Discussion

Here, we propose CD-Loop, a deep learning-based method that uses diffusion models to predict the chromatin loops from a given Hi-C contact map. Our extensive evaluations indicate that CD-Loop outperforms existing tools in loop annotation for datasets with various sequencing coverages.

The main contributions of CD-Loop are as follows: 1) the development of a deep learning framework that first conducts pre-training to obtain prior probabilities and then utilizes the denoising process of the diffusion model and a pre-trained estimative model for forecasting chromosomal loops in Hi-C contact matrices, resulting in improved accuracy for genome-wide chromatin loop recognition; 2) the use of data augmentation by flipping the interested parts of the Hi-C matrices in all four directions, which increases the diversity of training data and improves the generalization ability of the model, allowing for the training of a unified framework designed for processing Hi-C datasets from different sequencing depths, cell types, and species. A series of experimental results demonstrate that CD-Loop can effectively improve chromatin loop recognition accuracy compared to other methods and identify a range of unique chromatin loops. The overlap rate between different sequencing depths within the same cell line is relatively high, while the overlap rate between different cell lines is relatively low. Finally, equally important is that CD-Loop exhibits good robustness and stability on different biological cells and sequencing depths.

Although CD-Loop has superior performance compared to other methods, there are still areas that need optimization and improvement: 1) when predicting the entire Hi-C matrix, the prediction time is long. It can be improved by processing the data to reduce the waiting time. 2) CD-Loop can also be extended to analyze data at a higher resolution, but this would require optimizing the data processing procedure to reduce memory usage and IO time.

CD-Loop is a method that implements three-dimensional genome data analysis based on diffusion model classification. It enables accurate prediction of Hi-C contact maps at medium sequencing depth and improves the accuracy of its analysis even at low sequencing depth. With the continuous increase in high-quality Hi-C datasets, we expect that the capabilities of CD-Loop will be further improved and developed.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author/s.

## Author contributions

JS: data curation, investigation, methodology, resources, software, writing–original draft, and writing–review and editing. YW: data curation, formal analysis, investigation, methodology, resources, software, validation, writing–original draft, and writing–review and editing. JL: conceptualization, funding acquisition, investigation, methodology, supervision, and writing–review and editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported by the National Natural Science Foundation of China (grant no. 61972134, 62372156), Young Backbone Teachers of Henan Province (grant no. 2020GGJS050), Doctoral Fund of Henan Polytechnic University (grant no. B2018-36), Innovative Research Team of Henan Polytechnic University (grant no. T2021-3), and Henan Provincial Department of Science and Technology Research Project (grant no. 232102211046).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1393406/full#supplementary-material

## References

Abeywickrama, T., Cheema, M. A., and Taniar, D. (2016). *K-nearest neighbors on road networks: a journey in experimentation and in-memory implementation*. arXiv preprint arXiv:1601.01549. doi:10.48550/arXiv.1601.01549

Ben Zouari, Y., Molitor, A. M., Sikorska, N., Pancaldi, V., and Sexton, T. (2019). ChiCMaxima: a robust and simple pipeline for detection and visualization of chromatin looping in Capture Hi-C. *Genome Biol.* 20 (1), 102–119. doi:10.1186/s13059-019-1706-3

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. *Nat. Rev. Genet.* 17 (11), 661–678. doi:10.1038/nrg.2016.112

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. methods* 10 (12), 1213–1218. doi:10.1038/nmeth.2688

Cairns, J., Freire-Pritchett, P., Wingett, S. W., Várnai, C., Dimond, A., Plagnol, V., et al. (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* 17 (1), 127–217. doi:10.1186/s13059-016-0992-2

Cao, Y., Chen, Z., Chen, X., Ai, D., Chen, G., McDermott, J., et al. (2020). Accurate loop calling for 3D genomic data with cLoops. *Bioinformatics* 36 (3), 666–675. doi:10.1093/bioinformatics/btz651

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14 (6), 390–403. doi:10.1038/nrg3454

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature* 518 (7539), 331–336. doi:10.1038/nature14222

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485 (7398), 376–380. doi:10.1038/nature11082

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell. Syst.* 3 (1), 95–98. doi:10.1016/j.cels.2016.07.002

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462 (7269), 58–64. doi:10.1038/nature08497

Gorkin, D. U., Leung, D., and Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. *Cell. stem Cell.* 14 (6), 762–775. doi:10.1016/j.stem.2014.05.017

Grubert, F., Srivas, R., Spacek, D. V., Kasowski, M., Ruiz-Velasco, M., Sinnott-Armstrong, N., et al. (2020). Landscape of cohesin-mediated chromatin loops in the human genome. *Nature* 583 (7818), 737–743. doi:10.1038/s41586-020-2151-x

Han, X., Zheng, H., and Zhou, M. (2022). *CARD: classification and regression diffusion models*. doi:10.48550/arXiv.2206.07275

Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., et al. (2014). Genome-wide map of regulatory interactions in the human genome. *Genome Res.* 24 (12), 1905–1917. doi:10.1101/gr.176586.114

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Adv. neural Inf. Process. Syst.* 33, 6840–6851. doi:10.48550/arXiv.2006.11239

Kai, Y., Andricovich, J., Zeng, Z., Zhu, J., Tzatsos, A., and Peng, W. (2018). Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. *Nat. Commun.* 9 (1), 4221. doi:10.1038/s41467-018-06664-6

Kloetgen, A., Thandapani, P., Ntziachristos, P., Ghebrechristos, Y., Nomikou, S., Lazaris, C., et al. (2020). Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. *Nat. Genet.* 52 (4), 388–400. doi:10.1038/s41588-020-0602-9

Knight, P. A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA J. Numer. Analysis* 33 (3), 1029–1047. doi:10.1093/imanum/drs019

Kruse, K., Hug, C. B., and Vaquerizas, J. M. (2020). FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol.* 21 (1), 303–319. doi:10.1186/s13059-020-02215-9

Li, G., Fullwood, M. J., Xu, H., Mulawadi, F. H., Velkov, S., Vega, V., et al. (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.* 11, R22–R13. doi:10.1186/gb-2010-11-2-r22

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* 326 (5950), 289–293. doi:10.1126/science.1181369

Matthey-Doret, C., Baudry, L., Breuer, A., Montagne, R., Guiglielmoni, N., Scolari, V., et al. (2020). Computer vision for pattern detection in chromosome contact maps. *Nat. Commun.* 11 (1), 5795. doi:10.1038/s41467-020-19562-7

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* 47 (6), 598–606. doi:10.1038/ng.3286

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., et al. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. methods* 13 (11), 919–922. doi:10.1038/nmeth.3999

Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* 49 (11), 1602–1612. doi:10.1038/ng.3963

Phanstiel, D. H., Boyle, A. P., Heidari, N., and Snyder, M. P. (2015). Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics* 31 (19), 3092–3098. doi:10.1093/bioinformatics/btv336

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* 159 (7), 1665–1680. doi:10.1016/j.cell.2014.11.021

Roayaei Ardakany, A., Gezer, H. T., Lonardi, S., and Ay, F. (2020). Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome Biol.* 21, 256–317. doi:10.1186/s13059-020-02167-0

Rowley, M. J., Poulet, A., Nichols, M. H., Bixler, B. J., Sanborn, A. L., Brouhard, E. A., et al. (2020). Analysis of Hi-C data using SIP effectively identifies loops in organisms from *C. elegans* to mammals. *Genome Res.* 30 (3), 447–458. doi:10.1101/gr.257832.119

Salameh, T. J., Wang, X., Song, F., Zhang, B., Wright, S. M., Khunsriraksakul, C., et al. (2020). A supervised learning framework for chromatin loop detection in genome-wide contact maps. *Nat. Commun.* 11 (1), 3428. doi:10.1038/s41467-020-17239-9

Shi, C., Rattray, M., and Orozco, G. (2020). HiChIP-Peaks: a HiChIP peak calling algorithm. *Bioinformatics* 36 (12), 3625–3631. doi:10.1093/bioinformatics/btaa202

Szabo, Q., Bantignies, F., and Cavalli, G. (2019). Principles of genome folding into topologically associating domains. *Sci. Adv.* 5 (4), eaaw1668. doi:10.1126/sciadv.aaw1668

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., et al. (2015). CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell.* 163 (7), 1611–1627. doi:10.1016/j.cell.2015.11.024

Trieu, T., Martinez-Fundichely, A., and Khurana, E. (2020). DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. *Genome Biol.* 21, 79–11. doi:10.1186/s13059-020-01987-4

Wang, F., Gao, T., Lin, J., Zheng, Z., Huang, L., Toseef, M., et al. (2022). GILoop: robust chromatin loop calling across multiple sequencing depths on Hi-C data. *Iscience* 25 (12), 105535. doi:10.1016/j.isci.2022.105535

Wang, M., Sunkel, B. D., Ray, W. C., and Stanton, B. Z. (2022). Chromatin structure in cancer. *BMC Mol. Cell. Biol.* 23 (1), 35–10. doi:10.1186/s12860-022-00433-6

Wolff, J., Backofen, R., and Grüning, B. (2022). Loop detection using Hi-C data with HiCExplorer. *Gigascience* 11, giac061. doi:10.1093/gigascience/giac061

Wu, H., Zhou, B., Zhou, H., Zhang, P., and Wang, M. (2023). Be-1DCNN: a neural network model for chromatin loop prediction based on bagging ensemble learning. *Briefings Funct. Genomics, elad015* 22, 475–484. doi:10.1093/bfgp/elad015

Yang, D., Chung, T., and Kim, D. (2022). DeepLUCIA: predicting tissue-specific chromatin loops using deep learning-based universal chromatin interaction annotator. *Bioinformatics* 38 (14), 3501–3512. doi:10.1093/bioinformatics/btac373

Zhang, R., Wang, Y., Yang, Y., Zhang, Y., and Ma, J. (2018). Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics* 34 (13), i133–i141. doi:10.1093/bioinformatics/bty248

Zhang, S., Plummer, D., Lu, L., Cui, J., Xu, W., Wang, M., et al. (2022). DeepLoop robustly maps chromatin interactions from sparse allele-resolved or single-cell Hi-C data at kilobase resolution. *Nat. Genet.* 54 (7), 1013–1025. doi:10.1038/s41588-022-01116-w

Zhang, Y., and Blanchette, M. (2022). Reference panel guided topological structure annotation of Hi-C data. *Nat. Commun.* 13 (1), 7426. doi:10.1038/s41467-022-35231-3

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 May 2024
DOI 10.3389/fgene.2024.1404415

# LcDel: deletion variation detection based on clustering and long reads

Yanan Yu, Runtian Gao and Junwei Luo*

School of Software, Henan Polytechnic University, Jiaozuo, China

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Hailin Chen,
East China Jiaotong University, China
Minzhu Xie,
Hunan Normal University, China

*CORRESPONDENCE
Junwei Luo,
✉ luojunwei@hpu.edu.cn

RECEIVED 21 March 2024
ACCEPTED 25 April 2024
PUBLISHED 10 May 2024

CITATION
Yu Y, Gao R and Luo J (2024), LcDel: deletion variation detection based on clustering and long reads.
*Front. Genet.* 15:1404415.
doi: 10.3389/fgene.2024.1404415

COPYRIGHT
© 2024 Yu, Gao and Luo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

**Motivation:** Genomic structural variation refers to chromosomal level variations such as genome rearrangement or insertion/deletion, which typically involve larger DNA fragments compared to single nucleotide variations. Deletion is a common type of structural variants in the genome, which may lead to mangy diseases, so the detection of deletions can help to gain insights into the pathogenesis of diseases and provide accurate information for disease diagnosis, treatment, and prevention. Many tools exist for deletion variant detection, but they are still inadequate in some aspects, and most of them ignore the presence of chimeric variants in clustering, resulting in less precise clustering results.

**Results:** In this paper, we present LcDel, which can detect deletion variation based on clustering and long reads. LcDel first finds the candidate deletion sites and then performs the first clustering step using two clustering methods (sliding window-based and coverage-based, respectively) based on the length of the deletion. After that, LcDel immediately uses the second clustering by hierarchical clustering to determine the location and length of the deletion. LcDel is benchmarked against some other structural variation detection tools on multiple datasets, and the results show that LcDel has better detection performance for deletion. The source code is available in https://github.com/cyq1314woaini/LcDel.

KEYWORDS

**deletion, structural variation, long read, clustering, hierarchical-clustering**

## 1 Introduction

Genome sequences are very different between species, even within the same species. Genome variation refers to heritable changes in the composition or arrangement of base pairs at the molecular level of a gene, including single nucleotide variants, indels, structural variants (He et al., 2009), and copy number variants. A single nucleotide variation refers to the variation of one nucleotide base to another under the influence of certain factors; indels refers to the addition or subtraction of a small fragment to the genome that occurs within 50 bp of the length of the small fragment; Copy Number Variation refers to a rearrangement of the genome that has occurred and generally refers to an increase or decrease in the copy number of a genomic segment that is 1 kb or more in length; Structural Variations refer to mutations that occur on chromosomes in segments larger than 50 bp, including forms such as insertions, deletions, duplications, and inversions (Figure 1 below). Deletions account for a certain proportion of structural variants and have a large impact on the human body. Deletions in some genomes may lead to disease (Beyter et al., 2021), for example, deletions of genes related to the nervous system may lead to Huntington's chorea, and deletions of key genes may lead to cystic fibrosis and autism (Aganezov et al., 2020). Therefore, the detection of deletion variants can provide more precise information for the diagnosis, treatment and prevention of diseases.

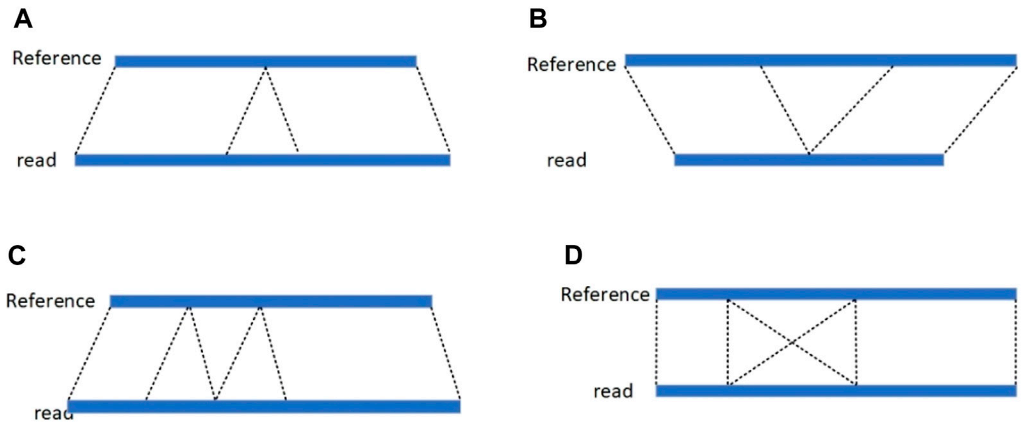


FIGURE 1
**(A)** insertion; **(B)** deletion; **(C)** duplication; **(D)** inversion (deletion is the loss of a portion of a chromosome; an insertion is the insertion of a portion of a chromosome; a duplication is the repetition of a portion of a chromosome; and an inversion is the reverse complementation operation of a portion of a chromosome.).

Genome sequencing technology has a significant impact on the detection of structural variants, and sequencing technology has gone through the first generation of sequencing technology, the second generation of sequencing technology, and the third generation of sequencing technology. The first-generation sequencing technology is known as Sanger sequencing technology (Sanger et al., 1977) and is widely used for genome sequencing. The read length of the first-generation sequencing technology can reach 1,000 bp with an accuracy of 99.99%, but the shortcomings of high sequencing cost and low throughput restrict its further application. Second-generation sequencing technology is also known as high-throughput sequencing (Maxam and Gilbert, 1992), which has the advantages of low cost, low sequencing error rate and high throughput, but the sequencing reads are shorter in length, which is more suitable for the detection of shorter structural variations, and the detection of structural variations in repetitive regions and regions with a high GC content has some difficulties. Third-generation sequencing technology, also known as single-molecule real-time technology (Korlach et al., 2010), is capable of directly sequencing longer DNA fragments and providing more comprehensive genomic information, but the sequencing error rate is high and therefore structural variants are not detected in sufficiently accurate locations. Cycle-consistent sequencing technology (Wenger et al., 2019) can sequence highly accurate long reads that cover repetitive and GC-rich regions of the genome, and can therefore be well suited for detecting structural variants.

Hi-C sequencing (de Wit and de Laat, 2012) is a high-throughput sequencing technology used to study the three-dimensional structure of chromosomes and genome interactions, which joins DNA fragments from different chromosomal regions by enzymatic cleavage and ligation techniques to form a DNA molecular library, which is then subjected to high-throughput sequencing to obtain sequences of multiple DNA reads. A number of methods for structural variation detection based on Hi-C reads have emerged, such as HiNT (Wang et al., 2020), HiCNV (Chakraborty and Ay, 2018), HiSV (Li et al., 2023), EagleC (Wang et al., 2022), etc. HiNT is a method for detecting interchromosomal translocations using Hi-C read. It utilizes a 1 Mb bin chromosome contact matrix as input. HiNT first calculates the Gini coefficient and maximum contact frequency of the interchromosomal contact matrix to identify potential translocated chromosome pairs. Then, it employs the breakpoint function from the R package 'struchanger' to approximate the breakpoints of the translocation. Finally, it utilizes an algorithm based on soft-clipped read counts to achieve precise breakpoint detection at single base pair resolution. HiCNV is a method for detecting copy number variations (CNVs) based on Hi-C read. It first processes the contact counts at the level of individual restriction enzyme fragments to utilize Hi-C data with as high resolution as possible. HiCNV calculates one-dimensional read coverage for each restriction enzyme, normalizes for GC content, mappability, and fragment length, smoothes using kernel density estimation, and finally identifies potential CNV segments using a hidden Markov model. HiSV is a structural variation detection method based on a significance detection model, capable of identifying large-scale structural variations from Hi-C read. Firstly, HiSV calculates a distance-normalized Hi-C contact matrix to avoid interference from strong interactions on the diagonal. Then, HiSV computes the local spatially weighted dissimilarity for each pixel to measure significance, thus separating significant regions from complex backgrounds. Finally, HiSV uses a global variation segmentation approach to partition sparse significant subsets into segments, considering a segment as a structural variation event if the interaction frequency after segmentation exceeds a predefined threshold. EagleC transforms the problem of identifying structural variations from Hi-C maps into a multi-label image classification problem. It is an ensemble learning framework based on 50 different models and utilizes convolutional neural networks as individual models for prediction. Additionally, EagleC proposes a data augmentation algorithm to ensure a balanced distribution of samples across different types of structural variations and genomic regions. These methods compare the interaction levels between normal and variant regions; large variant regions exhibit clear interaction patterns, while small variant regions exhibit less distinct interaction patterns. Therefore, these methods perform well in detecting large variant regions but less effectively for small ones. Moreover, since cancer cell lines often contain a higher proportion of large variant regions, these methods are effective in identifying variant regions in cancer cell lines and can be used for disease prediction.

The recently emerged Pore-C technology (Zhong et al., 2023), on the other hand, refers to a new technology that combines chromatin conformation capture technology with Nanopore sequencing technology to capture the information of chromatin multiregional interactions, and is able to form a single long read from multiple sequence fragments that are in close proximity to each other in three-dimensional space. This technique is capable of generating long reads on a genome-wide scale; however, these techniques generate less information about interactions in regions larger than three. Because of the complexity and diversity of structural variants, methods for detecting structural variants based on Pore-C reads are not yet available.

Most of the deletion variant detection tools (Mahmoud et al., 2019) that currently exist are based on short or long read, and although short read have a low sequencing error rate, they are short in length and do not completely find the deletion site. Although the sequencing error rate of long reads is high, the length is relatively long and it can span the deletion breakpoints well, so most of the tools nowadays detect the deletion by long reads, mainly by utilizing two methods, which are traditional method or deep learning.

The traditional method mainly involves first extracting candidate loci by characterizing various variants, and then clustering the candidate loci to determine the exact variant loci and length. CuteSV (Jiang et al., 2022) analyzes the characteristics of each type of structural variation and uses them to find potential loci for each variation separately, and clusters and further refines the clustering of read from heterozygous ratios in localized regions to accurately distinguish between pure and heterozygous variants. Finally, a few specific rules are used for structural variant detection and genotyping. Svim (Heller and Vingron, 2019) also collects structural variant features from the alignment files of the input sequences, then clusters the detected features using a clustering method based on graph and structural variant feature distance metrics, and finally outputs the final result by merging multiple structural variant events. Sniffles (Sedlazeck et al., 2018) uses the results from the NGMLR alignment as input and utilizes features from the segmented reads alignment, high mismatch regions, and coverage to identify structural variants. To overcome the high error rate in the reads, sniffles also evaluates candidate structural variants based on features such as length, location, and consistency of breakpoints. SKSV(Liu et al., 2021) is a skeleton-based structural variation detection analysis toolkit that performs pseudo alignment from reads and generates a alignment skeleton through sparse dynamic programming. The generated alignment skeleton supports rapid read finding and non-collinear segments in the alignment skeleton indicate potential structural variant events. Compared to other methods, SKSV is extremely fast and achieves high sensitivity and accuracy in both structural variant detection and genotyping. Svsearcher (Zheng et al., 2023) differs from previous methods in that it first finds candidate structural variant regions by variant characterization, then clusters read within the candidate regions to find candidate structural variants, and sets a stricter criterion to filter out erroneous structural variants.

Structural variation detection based on deep learning is mainly based on first extracting various features according to the type of structural variation, and then through the continuous training of the neural network, and then through the neural network to make predictions. INSnet (Gao et al., 2023) is a deep learning-based insertion variant detection method that firstly divides the reference genome into contiguous sub-regions and acquires five features for each locus. INSnet uses a convolutional neural network to extract variant features and a gated recurrent unit to analyze connections between subregions. MAMnet (Ding and Luo, 2022), a structural variation detection method based on the combination of convolutional neural networks and long and short-term memory networks, achieved a better F1-score compared to other comparison tools. SVcnn is a deep learning method that can accurately detect deletion, insertion, duplication and inversion variants. SVcnn (Zheng and Shang, 2023) first identifies candidate structural variant regions from the BAM file, then converts the candidate structural variant regions into images and constructs a LetNet model, which filters out the erroneous structural variants and outputs the final structural variants. cnnLSV (Ma et al., 2023) is also a deep learning based structural variation detection method that utilizes alignment information of long reads and convolutional neural networks to achieve overall higher performance and utilizes principal component analysis and k-means clustering algorithms to efficiently eliminate mislabeled samples during the training model phase. The results show that cnnLSV outperforms existing methods in detecting insertions, deletions, inversions and duplicate variants.

The traditional method ignores the occurrence of two different lengths of deletion variants at the same locus and directly clusters the candidate deletion sites, which may affect the final results of the detection of deletion variants. And deep learning can take a lot of time when extracting features and training. Therefore, this paper proposes an effective deletion variant detection algorithm LcDel. LcDel firstly merges the deletion variants that are closer when finding candidate deletion sites, then uses two clustering algorithms to perform the first level of clustering according to the length of deletion and uses hierarchical clustering to perform the second level of clustering, respectively, and finally filters out the candidate clusters that do not match to identify the location and length of deletion.

## 2 Methods

LcDel is a long reads-based deletion variant detection method where the input is a sorted bam file including the alignments between long reads and genome reference. There are four main steps in LcDel: 1) Identification of candidate deletion sites by intra-read alignment and inter-read alignment; 2) Multiple large clusters are generated by performing the first layer of clustering based on deletion lengths using sliding window-based and coverage-based methods, respectively; 3) Generate candidate clusters based on the differences between deletion lengths for large clusters using hierarchical clustering; 4) Set the support read threshold to filter out non-compliant candidate clusters and determine the location and length of the deletion. LcDel workflow is shown in Figure 2 below.

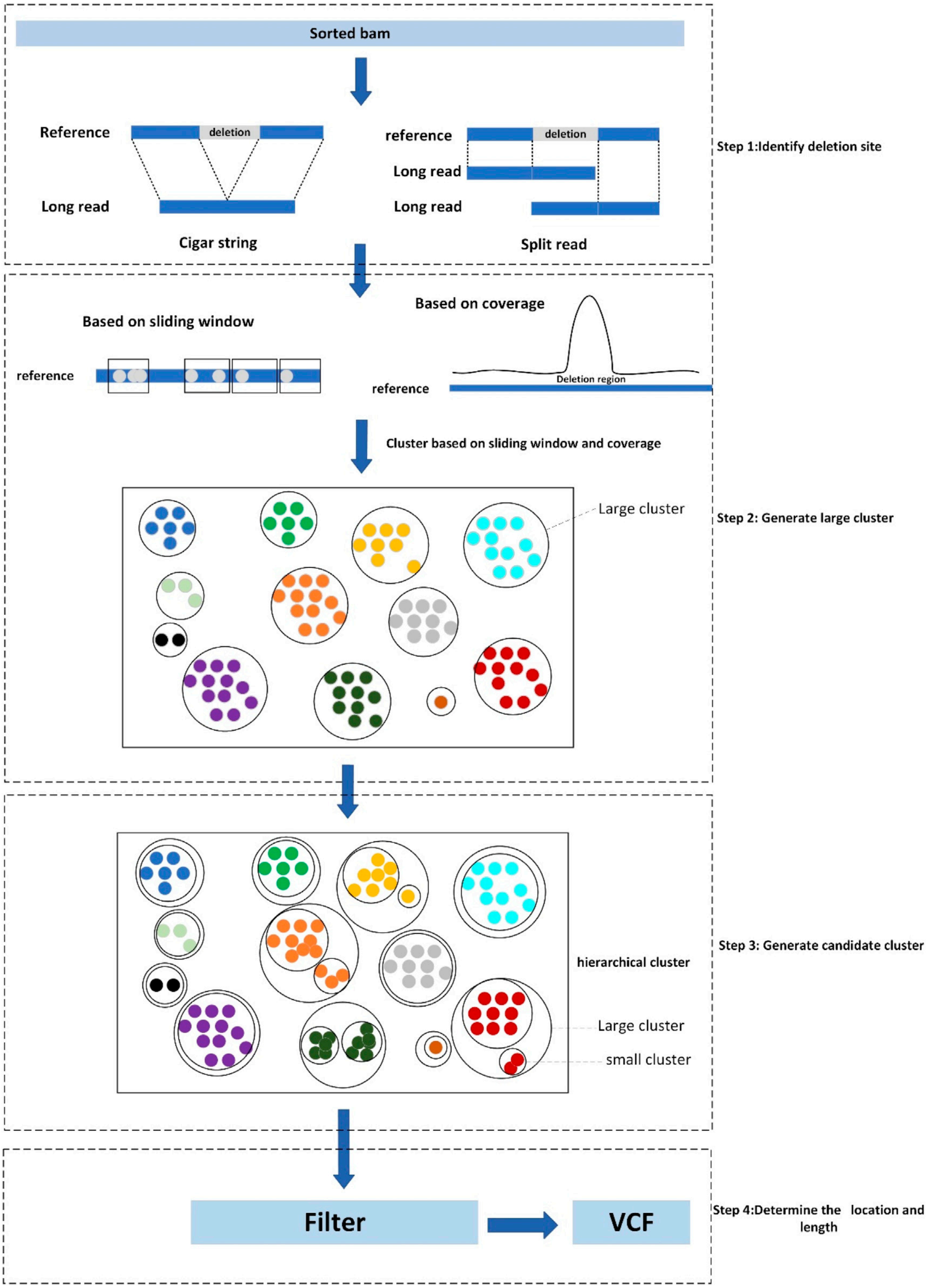


FIGURE 2
Workflow diagram of LcDel, step1 denotes identification of deletion sites, step2 denotes the first level of clustering to generate large clusters, step3 denotes the second level of clustering to generate candidate clusters, and step4 denotes the determination of the location and length of the deletion.

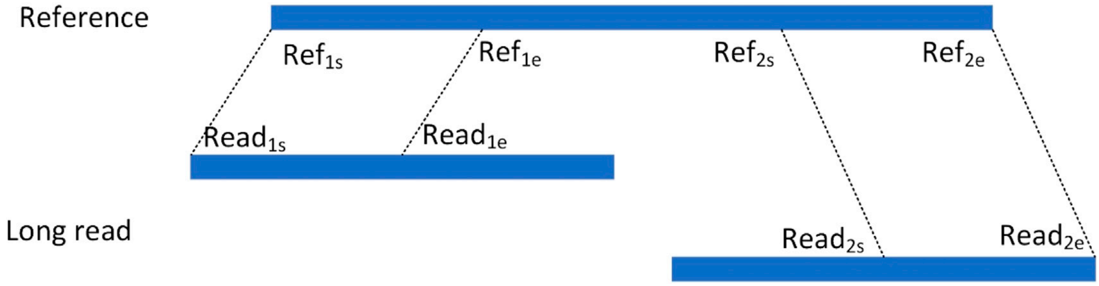


FIGURE 3
Split alignment of long read.

## 2.1 Identify candidate deletion site

Since small deletion variants are aligned to the reference genome, the alignment tool will directly display the deletion information in the cigar string, while large deletions are not directly displayed in the cigar string, but will be aligned to two non-contiguous regions of the same chromosome by splitting the reads, LcDel identifies candidate deletion loci by intra-read alignment and inter-read alignment, respectively.

LcDel first filters out alignments with mapping quality scores lower than 20 and unaligned ones, and then finds the ‘D’ identifiers with lengths greater than 30 in the cigar string, records the position and length of the deletion event on the reference genome, considers it as a deletion site and represents it as a quaternion Dt=(chr, start, svlen, end), where chr denotes the chromosome that the reads are aligned, start, and end denote the start and end positions of the deletion on the chromosome, respectively, and svlen denotes the length of the deletion on the chromosome. Due to the high sequencing error rate of long reads, which may result in a single deletion region being split into multiple smaller deletion regions during sequencing and alignment, it is necessary to determine whether merging can be performed if there are two deletion sites on the same read. For two quaternions $Dt_1$= ($chr_1$, $start_1$, $svlen_1$, $end_1$) and $Dt_2$= ($chr_2$, $start_2$, $svlen_2$, $end_2$) for the same read, where $Dt_1$ is assumed to be located in front of $Dt_2$, calculate the gap by using gap = $start_2$-$end_1$, and if $0<gap\leq 30$, then $Dt_1$ and $Dt_2$ are combined into a quaternion Dt= ($chr_1$, $start_1$, $svlen_1$+$svlen_2$, $end_2$), the new quaternion represents a large deletion variant.

For an alignment containing segmented reads, each matched read is represented as a hexadecimal Sig=(chr, $Ref_s$, $Ref_e$, $Read_s$, $Read_e$, orient), respectively, where chr denotes the chromosome to which the read is aligned, $Ref_s$ and $Ref_e$ denote the start and end points of the read alignment to the reference genome, respectively, $Read_s$ and $Read_e$ denote the start and end points of the segment of reads that are matched to the reference genome in the reads, respectively, and orient denotes the direction in which the read is aligned to the reference genome. As shown in Figure 3, the read is aligned to the reference genome due to the presence of variants resulting in splitting the read into two segments to be aligned to the reference genome separately, denoting the two alignments as the hexameric group $Sig_1$= ($chr_1$, $Ref_{1s}$, $Ref_{1e}$, $Read_{1s}$, $Read_{1e}$, $orient_1$) and $Sig_2$= ($chr_2$, $Ref_{2s}$, $Ref_{2e}$, $Read_{2s}$, $Read_{2e}$, $orient_2$), respectively. For two read segments of a split read comparison, which are aligned to the same chromosome in the same direction, i.e., $chr_1 = chr_2$ and $orient_1 = orient_2$, the spacing Distance_ref on the chromosome, the spacing Distance_read on the read, and the difference in their spacing Distance are computed, respectively. The setting of the upper limit of the distance threshold is described in detail in the results section 3.5.

$$\begin{cases} Distance\_ref = Ref_{2s} - Ref_{1e} \\ Distance\_read = Read_{2s} - Read_{1e} \\ Distance = Distance\_ref - Distance\_read \end{cases}$$

If Distance lies between the interval [50, 100,000], it indicates that this splitting read contains a deletion event, which is considered as a candidate deletion site and represented as a quaternion Deletion= (chr, $Ref_{1e}$, Distance, $Ref_{2s}$).

## 2.2 Generate large cluster

Clustering is commonly used for grouping data points in a dataset with similar characteristics into one category to help us better understand and utilize the information in the dataset, discover patterns and regularities in the data, and provide useful tasks for subsequent prediction and classification. The traditional methods for structural variant detection are generally to first find potential candidate variant sites through coverage, split reads and other features, and then filter the clusters with higher confidence as candidate clusters through clustering, and find the appropriate variant sites from the candidate clusters as the final result. In structural variation detection, a sliding window-based clustering method is usually used, which can effectively cluster candidate loci representing the same variant site together to facilitate structural variation detection. Clustering methods based on sliding windows need to set the window size in advance, and since the window size is fixed, when the window is set too large or too small it will result in the final finding of structural variants that are not complete. Since the coverage of the deletion region is lower than that of the normal region, the deletion region can be found by observing the coverage, so LcDel clusters in the first layer of clustering according to the length of deletion variation based on the two clustering methods of the sliding window and coverage, respectively, which can effectively cluster the deletion events that are mutated at the same locus together.

LcDel first sets a length threshold of 2000 and then refers to deletion lengths less than this threshold as small candidate deletion

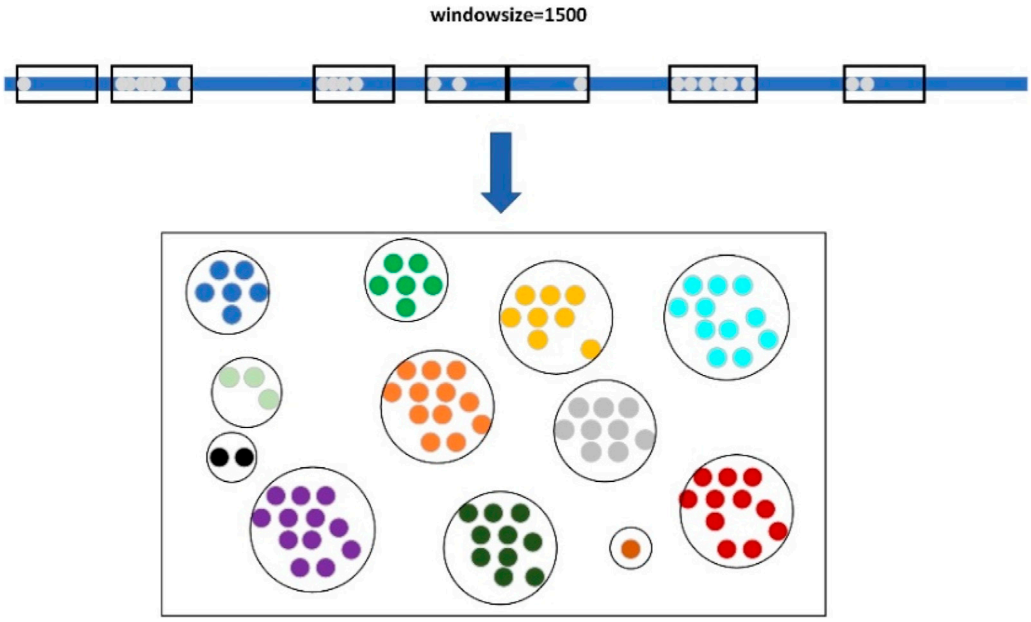


FIGURE 4
Clustering of large candidate deletion sites based on sliding windows.

sites and deletion lengths greater than this threshold as large candidate deletion sites. The small candidate deletion sites are then clustered using a coverage-based clustering method, while the large candidate deletion sites are clustered using a sliding window-based clustering method.

When clustering large candidate deletion sites, LcDel first sorts the large candidate deletion sites in ascending order according to their position on the reference genome, and then sets up a window of length 1,500, as shown in Figure 4 below. The starting position of the sliding window is the position of the first large candidate deletion site on the reference genome, and then keep sliding the window, if there is no large candidate deletion site in the sliding window at a certain moment, the next large candidate deletion site will be taken as the new starting point of the sliding window directly, and if the sliding window contains a large candidate deletion site, the next sliding window will need to take the end position of the window as the new starting point. In the process of continuously sliding the window, the large candidate deletion sites contained within the sliding window are clustered together to form a cluster, which is considered as a large cluster.

Since the coverage of the deletion region is significantly lower than that of the normal region, analyzing the feature of coverage can detect the deletion region, so the small candidate deletion sites are clustered using the coverage-based clustering method. LcDel sets up a list of chromosomes in the reference genome of length corresponding to the length of the corresponding chromosome, respectively, and the initial values of the list are all 0. Each position in the list corresponds to the corresponding base site on the reference genome. Then LcDel traverses each small candidate deletion site, looks at the region where each small candidate deletion site is located, takes out the list corresponding to the reference genome where the small candidate deletion site is located, and then adds 1 to the value of the list corresponding to the deletion region, and continually continues this process until all the small candidate deletion sites have been traversed. LcDel uses each locus of the reference genome as a horizontal coordinate and the list value corresponding to that locus as a vertical coordinate to build a planar graph.

As shown in Figure 1, step 2, the deletion region will form a shape similar to a mountain peak. Find the interval corresponding to that peak and cluster the small candidate deletion sites within that interval together to form a large cluster. In the first layer of clustering process, based on the sliding window and coverage clustering is carried out for different candidate deletion sites, the two clustering methods are independent of each other, so the two clustering methods are carried out at the same time, and the clustering for each chromosome is also processed in parallel by multi-threading, which makes LcDel extremely fast in the first layer of clustering. At the end of the first layer of clustering, all candidate deletion sites are clustered into multiple large clusters.

## 2.3 Generate candidate cluster

Since one human chromosome is composed of two homologous chromosomes, however, deletion variants of different lengths may have occurred on the two homologous chromosomes, such as deletion events of lengths 108 and 216 at locus 1,120,034 on chromosome 1, respectively. If the large clusters from the first level of clustering are used directly as final candidate clusters, the deletion length at that location may be determined incorrectly when determining the deletion length, affecting the final detection results. In order to separate these deletion variations, LcDel uses a hierarchical clustering method for a second clustering, which makes the clustering results more accurate and helps to determine the subsequent deletion length.

TABLE 1 Detection performance of LcDel in different situations.

| Coverage | | one_situation | two_situation |
|---|---|---|---|
| 69X | precision | 0.9535 | 0.9611 |
| | recall | 0.9725 | 0.9832 |
| | F1 | 0.962 | 0.9721 |
| 35X | precision | 0.9387 | 0.9485 |
| | recall | 0.9561 | 0.9764 |
| | F1 | 0.9473 | 0.9623 |
| 20X | precision | 0.9293 | 0.9369 |
| | recall | 0.9341 | 0.9446 |
| | F1 | 0.9316 | 0.9407 |
| 10X | precision | 0.9186 | 0.9273 |
| | recall | 0.8563 | 0.8671 |
| | F1 | 0.8861 | 0.8962 |
| 5X | precision | 0.9412 | 0.9573 |
| | recall | 0.6765 | 0.6866 |
| | F1 | 0.7871 | 0.7997 |

LcDel treats each candidate deletion site in a large cluster as a small cluster, and the average of the lengths of all the deletions in the small cluster is considered as the deletion length of the small cluster. LcDel calculates the length difference between any two small clusters contained in the large cluster each time, and then LcDel merges the two small clusters with the smallest length difference, and keeps iterating this merging process until the final large cluster contains only two small clusters. For two small clusters with deletion lengths of $len_1$ and $len_2$, respectively, the length difference rate between them can be calculated by the following formula. At this point, candidate deletion sites that support different deletion length variants can be clustered together to form small clusters. Finally, it is necessary to determine whether two small clusters represent the same deletion variant based on the difference in length between them, and to determine whether two small clusters in a large cluster can be merged into a single cluster. If the difference in the length of two clusters is less than 20%, the two clusters are merged into one candidate cluster, otherwise both clusters are considered as candidate clusters.

$$rate = \frac{abs(len_1, len_2)}{\max(len_1, len_2)}$$

## 2.4 Determine location and length of deletion

Candidate clusters have been identified through the previous two layers of clustering. LcDel then sets a support read threshold that filters out the following two types of candidate clusters: 1) The large cluster contains only one candidate cluster and the number of candidate deletion sites in the candidate cluster is less than the supported read threshold; 2) The large cluster contains two candidate clusters, and the number of candidate deletion sites in the candidate clusters is less than half of the threshold of supported reads. To better illustrate the benefits of splitting our filtering of candidate clusters into two cases, we also benchmarked the HG002 CLR dataset in one case (filtering out candidate clusters smaller than the threshold of supported reads), and the results are shown in Table 1 below. By analyzing LcDel on CLR datasets with different coverage, it can be seen that LcDel can effectively improve the detection of deletion variants when filtered in two cases.

TABLE 2 Description of the dataset.

| | HG002 CLR | HG002 CCS |
|---|---|---|
| Read Count | 2,915,733 | 6,596,012 |
| Average Length | 7,938 | 13,478 |
| Coverage | 69X | 28X |

For the candidate clusters that are left behind, the average of the deletion positions and lengths in that candidate cluster is calculated, and the candidate deletion site with the deletion position and length closest to the average in that candidate cluster is taken as the final result.

# 3 Results

In order to objectively evaluate the detection performance of LcDel for deletion variants, this paper compares LcDel with four of the more frequently used current structural variant detection tools. The four structural variant detection tools, all of which perform variant detection based on long reads, are cuteSV, sniffles, svim, and pbsv. High-confidence deletion variant regions collected by the Genome in a Bottle program were used as the reference standard dataset for this experiment, and Truvari was used to evaluate and record precision, recall, and F1-scores for all experimental results. In order to fully evaluate the deletion detection performance of LcDel, three human sample datasets that are currently more commonly used were selected: HG002 CLR (average length: 7938bp), HG002 CCS (average length: 13,478bp), and detailed information of the datasets is shown in Table 2 below. Additionally, this paper sets appropriate support read thresholds for each detection tool separately, with specific settings for detection performance on each dataset presented in the Detection Performance section.

In the following experiments, we detected deletion variation on chromosomes 1–22. The structural variation detection methods based on deep learning commonly selects a portion of chromosomes as the training set and a portion of chromosomes as the validation set. These models are continuously trained through the training and validation sets, and finally the remaining chromosomes are predicted in the test set. Therefore, it is not appropriate to compare LcDel with the methods using deep learning.

## 3.1 Detection performance of the structural variation detection tools on the CLR dataset

First, we benchmark LcDel, cuteSV, svim, sniffles and pbsv detection tools on the HG002 CLR dataset, and the experimental

**TABLE 3** Performance comparison of SV detection tools on CLR Dataset.

| Coverage | | LcDel | cuteSV | Sniffles | Svim | pbsv |
|---|---|---|---|---|---|---|
| 69X | Precision | 0.9611 | 0.9557 | 0.964 | 0.9595 | 0.9617 |
| | Recall | 0.9832 | 0.9436 | 0.9438 | 0.9461 | 0.9472 |
| | F1 | 0.9721 | 0.9496 | 0.9538 | 0.9527 | 0.9544 |
| 35X | Precision | 0.9485 | 0.9527 | 0.9641 | 0.957 | 0.9634 |
| | Recall | 0.9764 | 0.9361 | 0.9261 | 0.9368 | 0.9368 |
| | F1 | 0.9623 | 0.9443 | 0.9447 | 0.9468 | 0.9499 |
| 20X | Precision | 0.9369 | 0.9504 | 0.9622 | 0.958 | 0.9609 |
| | Recall | 0.9446 | 0.9091 | 0.8794 | 0.9052 | 0.8736 |
| | F1 | 0.9407 | 0.9293 | 0.919 | 0.9309 | 0.9152 |
| 10X | Precision | 0.9273 | 0.9434 | 0.9559 | 0.9396 | 0.967 |
| | Recall | 0.8671 | 0.8377 | 0.7918 | 0.8389 | 0.6496 |
| | F1 | 0.8962 | 0.8874 | 0.8662 | 0.8864 | 0.7772 |
| 5X | Precision | 0.9573 | 0.9656 | 0.9649 | 0.9647 | 0.973 |
| | Recall | 0.6866 | 0.6632 | 0.6285 | 0.6586 | 0.4042 |
| | F1 | 0.7997 | 0.7864 | 0.7612 | 0.7828 | 0.5712 |

results are shown in Table 3 below. In order to fully evaluate the deletion detection performance of LcDel on datasets with different coverage, we also randomly downsampled the HG002 CLR dataset to 35X, 20X, 10X and 5X and benchmarked it. For datasets with 69X, 35X, 20X, 10X and 5X coverage, the support read thresholds were set to 10, 5, 3, 2 and 2, respectively.

Compared to the other four popular structural variation detection tools, as shown in Table 3, LcDel achieves the highest recall and F1-score on all coverage CLR datasets. For the 69X dataset, the precision of LcDel detection is not the highest, but it is only 0.3% less than the first place, while the recall and F1-score are the highest, with the recall being 3.6% higher than the second place and the F1-score being 1.77% higher than the second place, which indicates that LcDel has a better performance of deletion detection on the high coverage dataset. For the CLR dataset with 35X coverage, LcDel detection had the lowest precision, 1.5% lower than the first place, but the recall was 0.9% higher than the second place and the F1-score was 1.2% higher than the second place. For the 20X CLR dataset, the precision of LcDel detection was also the lowest, 2.5% lower than the first place, but the recall was 3.5% higher than the second place, achieving the highest F1-score. The performance of structural variant detection tools for deletion variant detection decreases with decreasing sequencing depth. For the 10X dataset, the precision of LcDel detection was 3.9% lower than the first place, but the recall was 2.8% higher than the second place, achieving the highest F1-score. For the dataset of 5X, the precision of LcDel detection is 1.5% lower than that of pbsv, but the recall is 28.24% higher than that of pbsv, achieving the highest recall and F1-score, which indicates that LcDel has a better performance of deletion detection on CLR datasets of different coverage.

## 3.2 Performance of structural variation detection tools on different deletion lengths

In order to evaluate the performance of structural variant detection tools for different deletion lengths, in this paper, the variant lengths are categorized into five intervals of [50, 200], [200, 500], [500, 1,000], [1,000, 2000], and [2000+] for benchmarking respectively. The structural variation detection tool was benchmarked on the 69X、10X and 5X datasets of the HG002 CLR and the results are shown in Table 4 and 5 below, respectively.

Compared with the other four commonly used structural variation detection tools, as shown in Table 4, LcDel achieved the highest F1-scores at different deletion lengths, which indicates that LcDel has a better detection effect for different deletion lengths on the 69X dataset of HG002 CLR. On the [50, 200] interval, the precision of LcDel detection was 1.2% lower than the first place, but the recall was 3.5% higher than the second place, and the F1-score was 1.34% higher than the second place, which indicates that LcDel has a better detection effect on small deletion variants. On the [200, 500] interval, although the LcDel detection had the second highest recall, only 0.15% lower than the first place, it achieved the highest precision and F1-score. On the [500,1000] interval, although LcDel did not detect the highest precision, it achieved the highest recall and F1-score. On the [1,000,2000] interval, LcDel achieved the highest precision, recall, and F1-score, which were 1.06%, 1.05%, and 1.29% higher than the second place, respectively. By analyzing in the intervals [200, 500], [500, 1,000] and [1,000, 2000], it was found that LcDel has better performance for large deletion variant detection. On the [2000+] interval, LcDel also achieved the highest precision, recall, and F1-score, which were 2.7%, 1.8%, and 2.6% higher than the

TABLE 4 Comparison of detection performance for different deletion lengths.

| Coverage | Interval | | LcDel | cuteSV | Sniffles | Svim | pbsv |
|---|---|---|---|---|---|---|---|
| 69X | | Precision | 0.9394 | 0.936 | 0.9514 | 0.9446 | 0.947 |
| | 50–200 | Recall | 0.9768 | 0.9018 | 0.9208 | 0.9217 | 0.9416 |
| | | F1 | 0.9577 | 0.9186 | 0.9359 | 0.933 | 0.9443 |
| | | Precision | 0.9846 | 0.9718 | 0.9778 | 0.9754 | 0.9744 |
| | 200–500 | Recall | 0.9831 | 0.9831 | 0.9846 | 0.9792 | 0.9669 |
| | | F1 | 0.9838 | 0.9774 | 0.9812 | 0.9774 | 0.9706 |
| | | Precision | 0.9692 | 0.9447 | 0.9741 | 0.9489 | 0.9659 |
| | 500–1,000 | Recall | 0.9594 | 0.9543 | 0.9543 | 0.9441 | 0.8629 |
| | | F1 | 0.9643 | 0.9494 | 0.9641 | 0.9465 | 0.9115 |
| | | Precision | 0.9793 | 0.9687 | 0.9639 | 0.9585 | 0.9502 |
| | 1,000–2000 | Recall | 0.9844 | 0.9687 | 0.9739 | 0.9635 | 0.8958 |
| | | F1 | 0.9818 | 0.9687 | 0.9689 | 0.961 | 0.9222 |
| | | Precision | 0.9719 | 0.9386 | 0.9419 | 0.9444 | 0.9354 |
| | 2000+ | Recall | 0.9811 | 0.9622 | 0.8679 | 0.9088 | 0.9119 |
| | | F1 | 0.9765 | 0.9503 | 0.9034 | 0.9262 | 0.9235 |

second place, respectively, which demonstrated LcDel's better detection performance even for larger deletion variants.

As shown in Table 5, LcDel achieved the highest recall and F1-score on different intervals, which indicates that LcDel has a better detection effect for different deletion lengths on the 10X dataset of HG002 CLR. On the [50, 200] interval, although the pbsv detection had the highest precision, the LcDel detection had 20.58% higher recall and achieved the highest F1-score. On the [200, 500] interval, LcDel achieved the second highest precision, only 0.59% lower than the first place, but the recall and F1-score were 0.7% and 0.76% higher than the second place, respectively. On the [500, 1,000] interval, although the precision of LcDel detection was 1.97% lower than sniffles, the recall was 5.59% higher, achieving the highest F1-score. LcDel achieved the highest precision, recall, and F1-score on both the [1,000, 2000] and [2000+] intervals.

Although LcDel, cuteSV, sniffles, svim, and pbsv were all detected poorly on the 5X of the HG002 CLR dataset, LcDel still achieved the highest F1 scores on each interval, which suggests that LcDel has a better detection performance for deletion variants of different lengths even at low coverage.

## 3.3 Performance of LcDel on different support read parameters

Too large or too small support reads can affect the performance of detection of deletion variants, in order to evaluate the performance of LcDel in detecting deletion variants under different support reads thresholds, so this paper sets the support read support to 2, 3, 5 and 10 on 69X, 35X and 10X datasets of HG002 CLR to benchmark LcDel, respectively. The test results are shown in Table 6, which shows that on the 69X HG002 CLR dataset, as the number of supported reads continues to increase, the precision continues to increase and the recall continues to decrease, but the F1 score generally increases until a better detection result is achieved at a number of supported reads of 10. On the HG002 CLR dataset of 35X, it was found that LcDel achieves better deletion detection performance when the number of supported reads is set to 5. On the 20X dataset, better deletion detection results are achieved when the number of supported reads is set to 3. From the above analysis, it is found that the larger the support read are set, the greater the precision of LcDel detection and the smaller the recall.

## 3.4 Detection performance of the structural variation detection tool on the CCS dataset

In order to fully evaluate the performance of LcDel on different datasets for deletion variant detection, in addition to the CLR dataset of HG002, this paper also benchmarked LcDel, cuteSV, sniffles, svim, and pbsv on the CCS dataset of HG002, respectively. In addition to that, in this paper, the CCS dataset is randomly downsampled to 10X and 5X, and the support reads are set to 3, 2, and 1 for benchmarking, respectively, and the results are shown in Table 7. On CCS datasets with different coverage, although none of LcDel's precision is the highest, its recall is the highest and it achieves a good F1 score, which shows that LcDel can achieve similar deletion detection performance on CCS datasets as other detection tools.

## 3.5 Deletion detection performance of LcDel at different distance thresholds

Since reads spanning small deletion regions are compared with the reference genome, the deletion information is displayed directly

TABLE 5 Comparison of detection performance for different deletion lengths.

| | Interval | | LcDel | cuteSV | Sniffles | Svim | pbsv |
|---|---|---|---|---|---|---|---|
| 10X | 50–200 | Precision | 0.886 | 0.9221 | 0.9383 | 0.9169 | 0.9566 |
| | | Recall | 0.844 | 0.7975 | 0.7568 | 0.8113 | 0.6382 |
| | | F1 | 0.8645 | 0.8553 | 0.8378 | 0.8609 | 0.7656 |
| | 200–500 | Precision | 0.9709 | 0.9626 | 0.9709 | 0.9609 | 0.9768 |
| | | Recall | 0.8985 | 0.8915 | 0.8754 | 0.8907 | 0.7154 |
| | | F1 | 0.9333 | 0.9257 | 0.9207 | 0.9246 | 0.8259 |
| | 500–1,000 | Precision | 0.95 | 0.9389 | 0.9697 | 0.9389 | 0.952 |
| | | Recall | 0.868 | 0.8578 | 0.8121 | 0.8579 | 0.6041 |
| | | F1 | 0.9072 | 0.8965 | 0.8839 | 0.8966 | 0.7391 |
| | 1,000–2000 | Precision | 0.9583 | 0.9464 | 0.9491 | 0.9509 | 0.9357 |
| | | Recall | 0.8385 | 0.8281 | 0.776 | 0.8073 | 0.5313 |
| | | F1 | 0.8944 | 0.8833 | 0.8539 | 0.8732 | 0.6778 |
| | 2000+ | Precision | 0.9446 | 0.9283 | 0.9346 | 0.9387 | 0.9302 |
| | | Recall | 0.8585 | 0.8144 | 0.6289 | 0.7705 | 0.5031 |
| | | F1 | 0.8995 | 0.8677 | 0.7519 | 0.8463 | 0.6531 |
| 5X | 50–200 | Precision | 0.886 | 0.9221 | 0.9383 | 0.9169 | 0.9566 |
| | | Recall | 0.844 | 0.7975 | 0.7568 | 0.8113 | 0.6382 |
| | | F1 | 0.8645 | 0.8553 | 0.8378 | 0.8609 | 0.7656 |
| | 200–500 | Precision | 0.9709 | 0.9626 | 0.9709 | 0.9609 | 0.9768 |
| | | Recall | 0.8985 | 0.8915 | 0.8754 | 0.8907 | 0.7154 |
| | | F1 | 0.9333 | 0.9257 | 0.9207 | 0.9246 | 0.8259 |
| | 500–1,000 | Precision | 0.95 | 0.9389 | 0.9697 | 0.9389 | 0.952 |
| | | Recall | 0.868 | 0.8578 | 0.8121 | 0.8579 | 0.6041 |
| | | F1 | 0.9072 | 0.8965 | 0.8839 | 0.8966 | 0.7391 |
| | 1,000–2000 | Precision | 0.9583 | 0.9464 | 0.9491 | 0.9509 | 0.9357 |
| | | Recall | 0.8385 | 0.8281 | 0.776 | 0.8073 | 0.5313 |
| | | F1 | 0.8944 | 0.8833 | 0.8539 | 0.8732 | 0.6778 |
| | 2000+ | Precision | 0.9446 | 0.9283 | 0.9346 | 0.9387 | 0.9302 |
| | | Recall | 0.8585 | 0.8144 | 0.6289 | 0.7705 | 0.5031 |
| | | F1 | 0.8995 | 0.8677 | 0.7519 | 0.8463 | 0.6531 |

The above analysis reveals that LcDel has good detection performance for different deletion variant lengths on both high coverage (69X) and low coverage (10X and 5X) datasets of HG002 CLR.

in the cigar string, e.g., the presence of 100D in the cigar string indicates that the region to which the read is compared on the reference genome contains a deletion region of 100 bp in length. And for some reads containing large deletion regions do not display the deletion information directly in the cigar string when compared with the reference genome, they will be compared to two non-adjacent regions of the same chromosome by clipping the reads. Therefore, for the alignment with clipped reads, we need to set certain conditions to determine whether the alignment contains a deletion variant or not. If two segments of a read are aligned to the same chromosome in the same direction, the distance between the two segments on the read, Distance_read, is computed separately, and the distance between the two segments on the read, Distance_ref, is further computed for the two segments on the read aligned to the reference genome. Normally Distance_read should be 0, but due to sequencing errors and alignment tools, Distance_read may not be 0. In order to determine whether the alignment contains a deletion variant, it is necessary to determine whether the difference between them, Distance, is greater than 50; if Distance is greater than 50, then the alignment contains a deletion variant, but an upper threshold needs to be set for Distance. In order to explore a suitable upper threshold, we set different upper thresholds (Dt) to test the

**TABLE 6** LcDel deletion detection performance at different supported read.

| Coverage | Support>=2 | Support>=3 | Support>=5 | Support>=10 |
|---|---|---|---|---|
| 69X | 0.6094 | 0.7419 | 0.8755 | 0.9611 |
| | 0.9965 | 0.9942 | 0.9922 | 0.9832 |
| | 0.7563 | 0.8498 | 0.9302 | 0.9721 |
| 35X | 0.7602 | 0.8721 | 0.9488 | 0.9818 |
| | 0.9934 | 0.9908 | 0.9764 | 0.8503 |
| | 0.8612 | 0.9277 | 0.9624 | 0.9113 |
| 20X | 0.8708 | 0.9393 | 0.9774 | 0.9924 |
| | 0.9796 | 0.9444 | 0.8188 | 0.4103 |
| | 0.922 | 0.9418 | 0.8911 | 0.5806 |

**TABLE 7** Comparison of the performance of SV detection tools on the CCS dataset of HG002.

| Coverage | | LcDel | cuteSV | Sniffles | Svim | pbsv |
|---|---|---|---|---|---|---|
| 28X | Precision | 0.9378 | 0.9366 | 0.9487 | 0.9443 | 0.9459 |
| | Recall | 0.9504 | 0.9414 | 0.9399 | 0.9446 | 0.9346 |
| | F1 | 0.9441 | 0.939 | 0.9433 | 0.9444 | 0.9402 |
| 10X | Precision | 0.9346 | 0.9447 | 0.9525 | 0.9382 | 0.9567 |
| | Recall | 0.9113 | 0.9016 | 0.8787 | 0.9040 | 0.8214 |
| | F1 | 0.9228 | 0.9226 | 0.9141 | 0.9208 | 0.8839 |
| 5X | Precision | 0.9109 | 0.9217 | 0.9581 | 0.9164 | 0.9722 |
| | Recall | 0.8818 | 0.8731 | 0.6994 | 0.8743 | 0.5017 |
| | F1 | 0.8962 | 0.8968 | 0.8085 | 0.8949 | 0.6618 |

HG002 CLR dataset respectively, and the test results are shown in Table 8 below.

The above table shows that when the upper limit of the Distance threshold is set too small, LcDel has higher precision but lower recall in the detection results. As the upper limit of the Distance threshold continues to increase, the precision of the LcDel detection results gradually decreases, but the recall gradually increases. The reason for this analysis is that the upper limit of the Distance threshold is set too high causing many false positive events to be treated as missing events. If the upper Distance threshold is not set, LcDel will regard many false-positive events as deletion events, so it will ultimately lead to lower accuracy and higher recall of LcDel's detection results. Therefore, setting an appropriate upper Distance threshold is especially important for the detection of missing variants, and the upper Distance threshold is set to 100,000 by the above table.

## 3.6 Comparison of deletion performance of LcDel at the first level of clustering using different methods

Since the length of large candidate deletion variant sites is generally longer, the effect may not be so obvious if the clustering is done using the coverage-based method, while the use of the sliding window-based clustering of deletion variants with larger lengths can have good results. Due to the short length of small candidate deletion sites, the deletion detection performance of LcDel may be reduced if a sliding window-based approach is directly used to cluster all deletion sites at the first level. In order to assess the impact of using the sliding window clustering method on small candidate deletion sites, we performed benchmarking on the first layer of clustering on the HG002 CLR dataset using sliding window based one method alone (SW) and using sliding window, coverage based two methods (SW + CG), the benchmarking results are shown in Table 9 below.

From Table 9, it can be seen that if all candidate deletion sites are clustered using the sliding window-based method, it does not have a better detection performance than clustering using both sliding window-based and coverage-based methods. Consequently, clustering of small candidate deletion sites using the sliding window-based method decreases the detection performance of LcDel, which may be due to the relatively large window setting, resulting in some small candidate deletion sites that are relatively close to each other being clustered together, affecting the accuracy of LcDel.

TABLE 8 LcDel results for different upper Distance thresholds on the CLR dataset.

| Coverage | | Dt ≤ 1,000 | Dt ≤ 5,000 | Dt ≤ 10,000 | Dt ≤ 50,000 | Dt ≤ 100,000 | Dt ≤ 200,000 |
|---|---|---|---|---|---|---|---|
| 69X | precision | 0.9627 | 0.9638 | 0.9642 | 0.9635 | 0.9611 | 0.9605 |
| | recall | 0.9239 | 0.9521 | 0.9761 | 0.9791 | 0.9832 | 0.9832 |
| | F1 | 0.9429 | 0.9579 | 0.9701 | 0.9712 | 0.9721 | 0.9717 |
| 35X | precision | 0.9495 | 0.9488 | 0.9495 | 0.9487 | 0.9485 | 0.9487 |
| | recall | 0.9195 | 0.9455 | 0.9696 | 0.9764 | 0.9764 | 0.9764 |
| | F1 | 0.9343 | 0.9471 | 0.9594 | 0.9624 | 0.9623 | 0.9624 |
| 20X | precision | 0.9381 | 0.9393 | 0.9403 | 0.9393 | 0.9369 | 0.9314 |
| | recall | 0.8877 | 0.9144 | 0.9378 | 0.9443 | 0.9446 | 0.9446 |
| | F1 | 0.9122 | 0.9267 | 0.939 | 0.9406 | 0.9407 | 0.9379 |
| 10X | precision | 0.9281 | 0.9287 | 0.9289 | 0.9286 | 0.9273 | 0.9286 |
| | recall | 0.8092 | 0.8386 | 0.8605 | 0.8663 | 0.8671 | 0.8671 |
| | F1 | 0.8646 | 0.8813 | 0.8934 | 0.8961 | 0.8962 | 0.8964 |
| 5X | precision | 0.9611 | 0.9614 | 0.9613 | 0.9612 | 0.9573 | 0.9505 |
| | recall | 0.637 | 0.6666 | 0.6819 | 0.6845 | 0.6866 | 0.6865 |
| | F1 | 0.7662 | 0.7873 | 0.7978 | 0.7995 | 0.7997 | 0.7972 |

TABLE 9 Performance comparison of LcDel on different clustering methods.

| Coverage | | SW | SW + CG |
|---|---|---|---|
| 69X | precision | 0.9596 | 0.9611 |
| | recall | 0.9691 | 0.9832 |
| | F1 | 0.9643 | 0.9721 |
| 35X | precision | 0.9483 | 0.9485 |
| | recall | 0.9616 | 0.9764 |
| | F1 | 0.9549 | 0.9623 |
| 20X | precision | 0.9443 | 0.9369 |
| | recall | 0.9322 | 0.9446 |
| | F1 | 0.9382 | 0.9407 |
| 10X | precision | 0.8633 | 0.9273 |
| | recall | 0.8566 | 0.8671 |
| | F1 | 0.86 | 0.8962 |
| 5X | precision | 0.9062 | 0.9573 |
| | recall | 0.6766 | 0.6866 |
| | F1 | 0.7747 | 0.7997 |

## 3.7 Comparison of detection performance of LcDel on HG002 CLR dataset for different window sizes

The use of sliding window based method in clustering large candidate deletion sites will achieve good results, but the setting of the window size may affect the detection performance of LcDel on deletion variants to a certain extent, so this paper attempted to set the sliding window (ws) to 500, 1,500, 2,500, 4,000 and 5,000 on the CLR dataset of HG002 for benchmarking, and the results are as follows in Table 10.

Through Table 10, it can be found that when the window setting is small it will reduce the detection performance of LcDel, and when the window size is 1,500, no matter how to increase the window, it has little effect on the detection effect of LcDel, which is largely due to the hierarchical clustering in the second layer. If the setting of the window is very large, it will cluster many close deletion variants together to form a cluster, if the exact deletion site is determined directly at this time it will lead to misidentification or miss identification of the deletion variants, but the hierarchical clustering can separate them very well, so the effect of LcDel does not decrease with the increase of the window. Through analysis, LcDel sets the window size to 1,500.

# 4 Discussion

In this paper, we propose a long read based deletion variant detection method LcDel using two-layer clustering. LcDel first finds candidate deletion sites from the sorted bam file by intra-read alignment and inter-read alignment. A method of heuristics was used to merge relatively close deletion sites. Use sliding window and coverage methods based on deletion length to perform the first layer clustering and generate multiple large clusters. Then, hierarchical clustering is used to further cluster the large clusters and generate candidate clusters, in order to improve the accuracy of clustering and facilitate the determination of deletion positions and lengths in the future. Finally, the candidate clusters containing candidate

TABLE 10 Comparison of LcDel's detection performance on different window sizes.

| Coverage | | ws = 500 | ws = 1,500 | ws = 2,500 | ws = 4,000 | ws = 5,000 |
|---|---|---|---|---|---|---|
| 69X | precision | 0.9512 | 0.9611 | 0.9541 | 0.9635 | 0.9635 |
| | recall | 0.9627 | 0.9832 | 0.9811 | 0.9832 | 0.9832 |
| | F1 | 0.9569 | 0.9721 | 0.9674 | 0.9733 | 0.9733 |
| 35X | precision | 0.941 | 0.9485 | 0.9435 | 0.9485 | 0.9483 |
| | recall | 0.9657 | 0.9764 | 0.9744 | 0.9764 | 0.9764 |
| | F1 | 0.9531 | 0.9623 | 0.9587 | 0.9622 | 0.9621 |
| 20X | precision | 0.9315 | 0.9369 | 0.9393 | 0.9393 | 0.9393 |
| | recall | 0.9343 | 0.9446 | 0.9444 | 0.9443 | 0.9443 |
| | F1 | 0.9328 | 0.9407 | 0.9418 | 0.9418 | 0.9418 |
| 10X | precision | 0.9181 | 0.9273 | 0.9286 | 0.9286 | 0.9286 |
| | recall | 0.8556 | 0.8671 | 0.8663 | 0.8663 | 0.8663 |
| | F1 | 0.8857 | 0.8962 | 0.8964 | 0.8961 | 0.8964 |
| 5X | precision | 0.9518 | 0.9573 | 0.9608 | 0.9608 | 0.9601 |
| | recall | 0.6761 | 0.6866 | 0.6765 | 0.6763 | 0.6765 |
| | F1 | 0.7906 | 0.7997 | 0.7937 | 0.7938 | 0.7937 |

deletion sites are filtered out, and the position and length of the deletion are determined from the remaining candidate clusters. To evaluate the detection performance of LcDel for deletion variants, we compared it with four currently popular structural variant detection tools on multiple datasets. The experimental results show that LcDel has better detection performance for deletions.

However, LcDel still has some limitations in some aspects. First, LcDel only detects deletion variants but not other types of structural variants such as insertions, translocations and inversions. Second, LcDel does not genotype the detected deletions. We will gradually improve the above problems in our future work.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

YY: Conceptualization, Data curation, Methodology, Software, Validation, Writing–original draft, Writing–review and editing. RG: Methodology, Software, Writing–review and editing. JL: Conceptualization, Funding acquisition, Investigation, Methodology, Software, Supervision, Writing–original draft, Writing–review and editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported by the National Natural Science Foundation of China (Grant No. 62372156), Young Backbone Teachers of Henan Province (Grant No. 2020GGJS050), Doctoral Fund of Henan Polytechnic University (Grant No. B2018-36), Innovative Research Team of Henan Polytechnic University (Grant No. T2021-3), and Henan Provincial Department of Science and Technology Research Project (Grant No. 232102211046).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

# References

Aganezov, S., Goodwin, S., Sherman, R. M., Sedlazeck, F. J., Arun, G., Bhatia, S., et al. (2020). Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* 30 (9), 1258–1273. doi:10.1101/gr.260497.119

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* 53 (6), 779–786. doi:10.1038/s41588-021-00865-4

Chakraborty, A., and Ay, F. (2018). Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics* 34 (2), 338–345. doi:10.1093/bioinformatics/btx664

De Wit, E., and De Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* 26 (1), 11–24. doi:10.1101/gad.179804.111

Ding, H., and Luo, J. (2022). MAMnet: detecting and genotyping deletions and insertions based on long reads and a deep learning approach. *Brief. Bioinform* 23 (5), bbac195. doi:10.1093/bib/bbac195

Gao, R., Luo, J., Ding, H., and Zhai, H. (2023). INSnet: a method for detecting insertions based on deep learning network. *BMC Bioinforma.* 24 (1), 80. doi:10.1186/s12859-023-05216-0

He, Y.-S., Zhang, W., and Yang, Z.-Q. (2009). Structural variation in the human genome. *Yi Chuan* 31 (8), 771–778. doi:10.3724/sp.j.1005.2009.00771

Heller, D., and Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics* 35 (17), 2907–2915. doi:10.1093/bioinformatics/btz041

Jiang, T., Liu, S., Cao, S., and Wang, Y. (2022). Structural variant detection from long-read sequencing data with cuteSV. *Methods Mol. Biol.* 2493, 137–151. doi:10.1007/978-1-0716-2293-3_9

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., et al. (2010). Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.* 472, 431–455. doi:10.1016/s0076-6879(10)72001-2

Li, J., Gao, L., and Ye, Y. (2023). HiSV: a control-free method for structural variation detection from Hi-C data. *PLoS Comput. Biol.* 19 (1), e1010760. doi:10.1371/journal.pcbi.1010760

Liu, Y., Jiang, T., Su, J., Liu, B., Zang, T., and Wang, Y. (2021). SKSV: ultrafast structural variation detection from circular consensus sequencing reads. *Bioinformatics* 37 (20), 3647–3649. doi:10.1093/bioinformatics/btab341

Ma, H., Zhong, C., Chen, D., He, H., and Yang, F. (2023). cnnLSV: detecting structural variants by encoding long-read alignment information and convolutional neural network. *BMC Bioinforma.* 24 (1), 119. doi:10.1186/s12859-023-05243-x

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biol.* 20 (1), 246. doi:10.1186/s13059-019-1828-7

Maxam, A. M., and Gilbert, W. (1992). A new method for sequencing DNA. *Biotechnol. (Read. Mass)* 24 (2), 99–103. doi:10.1073/pnas.74.2.560

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74 (12), 5463–5467. doi:10.1073/pnas.74.12.5463

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., et al. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* 15 (6), 461–468. doi:10.1038/s41592-018-0001-7

Wang, S., Lee, S., Chu, C., Jain, D., Kerpedjiev, P., Nelson, G. M., et al. (2020). HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. *Genome Biol.* 21, 73. doi:10.1186/s13059-020-01986-5

Wang, X., Luan, Y., and Yue, F. (2022). EagleC: a deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. *Sci. Adv.* 8 (24), eabn9215. doi:10.1126/sciadv.abn9215

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37 (10), 1155–1162. doi:10.1038/s41587-019-0217-9

Zheng, Y., and Shang, X. (2023). SVcnn: an accurate deep learning-based method for detecting structural variation based on long-read data. *BMC Bioinforma.* 24 (1), 213. doi:10.1186/s12859-023-05324-x

Zheng, Y., Shang, X., and Sung, W.-K. (2023). SVsearcher: a more accurate structural variation detection method in long read data. *Comput. Biol. Med.* 158, 106843. doi:10.1016/j.compbiomed.2023.106843

Zhong, J. Y., Niu, L., Lin, Z. B., Bai, X., Chen, Y., Luo, F., et al. (2023). High-throughput Pore-C reveals the single-allele topology and cell type-specificity of 3D genome folding. *Nat. Commun.* 14 (1), 1250. doi:10.1038/s41467-023-36899-x

Frontiers in Genetics

TYPE Methods
PUBLISHED 23 May 2024
DOI 10.3389/fgene.2024.1407072

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Pu-Feng Du,
Tianjin University, China
Jiazhou Chen,
South China University of Technology, China
Hongdong Li,
Central South University, China

*CORRESPONDENCE
Bolin Chen,
blchen@nwpu.edu.cn
Jun Bian,
blandbird@126.com

RECEIVED 26 March 2024
ACCEPTED 01 May 2024
PUBLISHED 23 May 2024

CITATION
Chen B, Wang Y, Zhang J, Han Y, Benhammouda H, Bian J, Kang R and Shang X (2024), Specific feature recognition on group specific networks (SFR-GSN): a biomarker identification model for cancer stages.
*Front. Genet.* 15:1407072.
doi: 10.3389/fgene.2024.1407072

COPYRIGHT
© 2024 Chen, Wang, Zhang, Han, Benhammouda, Bian, Kang and Shang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Specific feature recognition on group specific networks (SFR-GSN): a biomarker identification model for cancer stages

Bolin Chen[1,2]*, Yuxin Wang[1], Jinlei Zhang[1], Yourui Han[1], Hamza Benhammouda[1], Jun Bian[3]*, Ruiming Kang[4] and Xuequn Shang[1,2]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, China, [2]Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an, Shaanxi, China, [3]Department of General Surgery, Xi'an Children's Hosptial, Xi'an Jiaotong University Affiliated Children's Hosptial, Xi'an, China, [4]Rewise (Hangzhou) Information Technology Co., Ltd, Hangzhou, China

**Background and Objective:** Accurate identification of cancer stages is challenging due to the complexity and heterogeneity of the disease. Current clinical diagnosis methods primarily rely on phenotypic observations, which may not capture early molecular-level changes accurately.

**Methods:** In this study, a novel biomarker recognition method was proposed tailored for cancer stages by considering the change of gene expression relationships. Utilizing the sample-specific information and protein-protein interaction networks, the group specific networks were constructed to address the limited specificity of potential biomarkers. Then, a specific feature recognition method was proposed based on these group specific networks, which employed the random forest algorithm for initial screening followed by a recursive feature elimination process to identify the optimal biomarker subset. During exploring optimal results, a strategy termed the Cost-Benefit Ratio, was devised to facilitate the identification of stage-specific biomarkers.

**Results:** Comparative experiments were conducted on lung adenocarcinoma and breast cancer datasets to validate the method's efficacy and generalizability. The results showed that the identified biomarkers were highly stage-specific, and the F1 scores for predicting cancer stages were significantly improved. For the lung adenocarcinoma dataset, the F1 score reached 97.68%, and for the breast cancer dataset, it achieved 96.87%. These results significantly surpassed those of three conventional methods in terms of F1 scores. Moreover, from the perspective of biological functions, the biomarkers were proved playing an important role in cancer stage-evolution.

**Conclusion:** The proposed method demonstrated its effectiveness in identifying stage-related biomarkers. By using these biomarkers as features, accurate prediction of cancer stages was achieved. Furthermore, the method exhibited potential for biomarker identification in subtype analyses, offering novel perspectives for cancer prognosis.

KEYWORDS

**biomarker, cancer stages, group specific network, multi classification tasks, edge feature**

## 1 Introduction

Cancer is a disease characterized by uncontrolled cell proliferation, posing a serious threat to human health. According to the World Health Organization, in 2020 alone, nearly 10 million people (about one-sixth of all deaths worldwide) died from cancer (Sung et al., 2021). Understanding cancer begins with an important dimension: its stages, which could describe the size and extent of tumor spread. Due to the high heterogeneity and complexity of cancer, it poses significant challenges for the identification of cancer stages (Burrell et al., 2013). Hence, investigating an intelligent model for the identification of stage-related biomarkers is very important. It helps in understanding the characteristics and changes during the development process of cancer. This research endeavor proves valuable in enhancing cancer treatment strategies and prognostic assessments.

As far as the biomarkers are concerned, encompass a range of molecules, cellular structures, or biological processes that can be objectively detected and quantified within or outside an organism (Moein et al., 2020). They play a crucial role in revealing an individual's health status, physiological functions, pathological conditions, and biological responses to treatment. This makes them integral players in the development of precision medicine and personalized treatment strategies (Holland, 2016). Specifically, stage-related biomarkers provide crucial information about tumor progression, metastasis, and treatment response (Amin et al., 2010; Van der Kloet et al., 2012). By analyzing the expression patterns and changes of stage-related biomarkers, healthcare professionals and researchers can gain a better understanding of the cancer's progression status, choose appropriate treatment strategies, and monitor treatment effectiveness.

However, molecular distinctions between different cancer stages are often subtle (Ye et al., 2020). For example, in early-stage cancer, molecular changes may be influenced by minor alterations in the activity of a few key genes or subtle modulation of signaling pathways. The boundaries between cancer stages, as defined clinically, are often indistinct at the molecular level. For instance, the molecular changes between stage I of a late-stage and stage II of an early-stage cancer could be very similar. Therefore, the identification of stage-related biomarkers at the molecular level has been a long-standing challenge.

Currently, two mainstream approaches primarily guide the identification of stage-related biomarkers. The first category is based on differential expression analysis. Deva Magendhra Rao et al. (2019) compared non-coding RNAs (lncRNAs) between invasive ductal carcinoma (IDC) breast cancer tissues and normal breast tissues. There were 375 differentially expressed lncRNAs identifying closely associated with the early-stage development of breast cancer. Shi et al. (2018) analyzed gene expression data from four stages of colorectal cancer, identifying stage-specific differentially expressed genes and exploring their shared biological functions. Wang et al. (2017) studied gene expression data in non-small cell lung cancer and found that differentially expressed genes at different stages significantly impacted biological functions and signaling pathways. However, these methods often overlook molecular interactions and typically validate their findings through functional or pathway enrichment analysis but few focus on the identification of stage-related biomarkers.

On the other hand, the second category, focuses on machine learning techeques. Patil and Bellary (2022) achieved good performance in stage identification of melanoma based on features from dermoscopic images and tumor thickness using machine learning. Ubaldi et al. (2021) performed a binary classification task to identify stage I and stage II non-small cell lung cancer using radiometric data and machine learning, achieving a high AUC value at 0.84. Jin et al. (2021) developed an interpretable machine learning model that could identify gene expression biomarkers for early-stage LUAD. However, these methods typically focus on building accurate prediction models similar to a "black box" with limited biological and clinical interpretability. Some researchers strive to construct interpretable machine learning models for identifying stage-related biomarkers, but this often leads to compromises in the predictive performance of the model to some extent for the samples are imbalanced, and there is minimal molecular-level difference between different stages. In summary, existent methods have weaker specificity in identifying stage-related molecular-level biomarkers.

In this paper, an efficient method was proposed to identify stage-related biomarkers through specific feature recognition on group specific networks (SFR-GSN), which could sensitively capture the differences between different stages and identify features that exhibit significant specificity between stages. Two mainly high-risk cancers, lung adenocarcinoma (LUAD) and breast carcinoma (BRCA), were used to evaluate the proposed method. Firstly, the clinical data, RNA-Seq data and protein-protein interactions (PPI) of LUAD and BRCA were first collected from public database. Then, based on the tumor samples and normal samples, the sample-specific networks (SSN) were constructed, which further intersected with PPI to construct the group-specific network (GSN). Through clinical data, GSNs were combined into one GSN corresponding to one cancer stage, which could address the weak specificity of existing biomarkers. Subsequently, a specific feature recognition (SFR) method based on these GSNs was proposed. SFR was designed in two-round, the first round was pre-screening by utilizing the random forest algorithm with Gini impurity quantifying the purity improvement. The second round was optimal subset screening of biomarkers by using the recursive feature elimination with cross-validation. Notably, during exploring the optimal results, the Cost-Benefit Ratio (CBR) was introduced as an important indicator for identifying the stage-related biomarkers. Eventually, comparative experiments among SFR-GSN and three

**TABLE 1** The number of samples of LUAD and BRCA in experiments.

| Cancer types | Normal | Stage I | Stage II | Stage III | Stage IV | Sum |
|---|---|---|---|---|---|---|
| LUAD | 59 | 273 | 122 | 83 | 26 | 563 |
| BRCA | 114 | 182 | 621 | 250 | 20 | 1,187 |

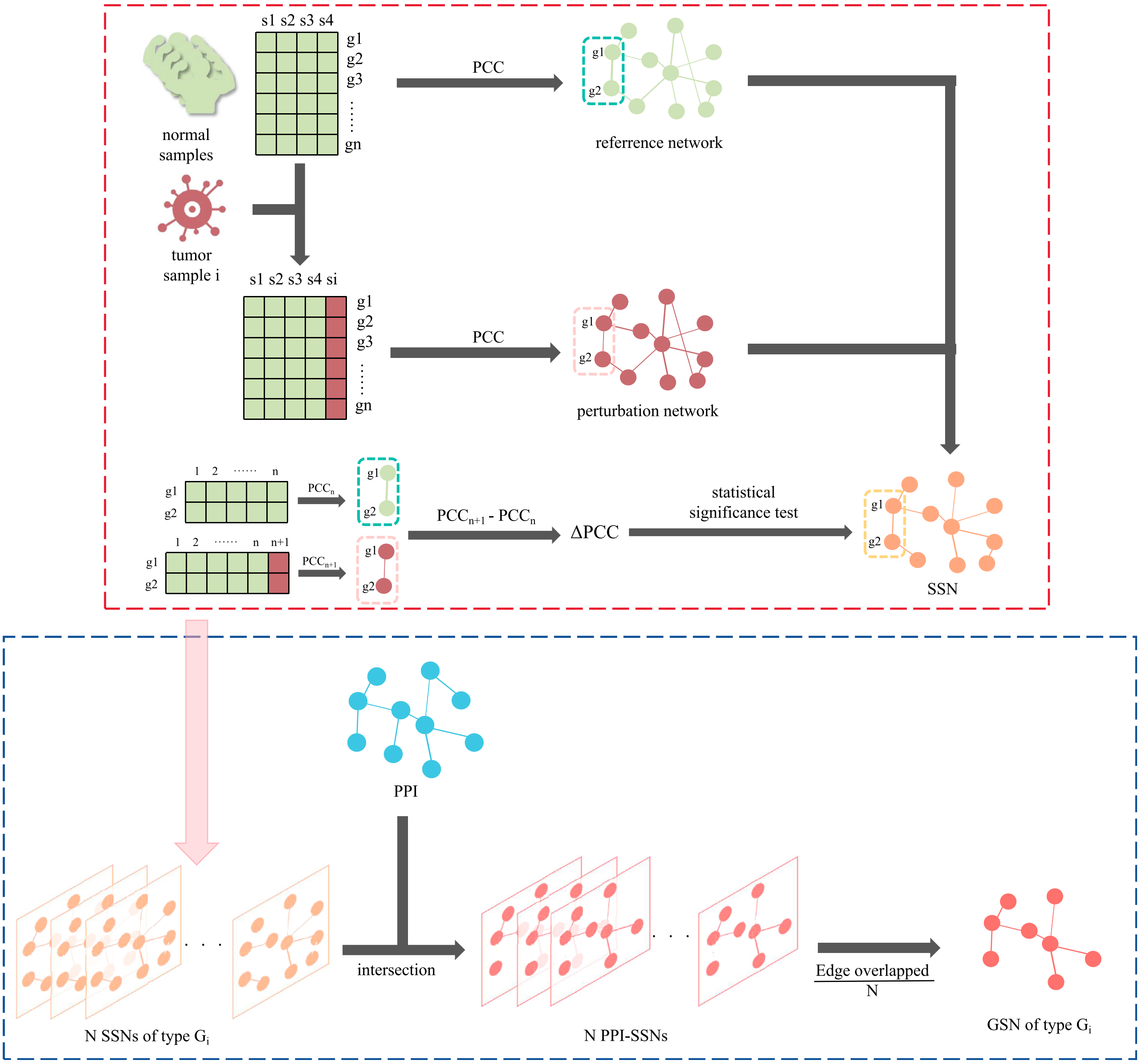


FIGURE 1
The flowchart of constructing GSN. The red dotted line section is the construction of SSN, while the blue dotted line section is the following part, using SSN and PPI to construct the GSN.

state-of-the-art methods were conducted on LUAD and BRCA datasets to validate the effectiveness and generalization ability of the proposed method. The results showed that the identified biomarkers significantly improved F1 scores for predicting cancer stages. Also from the perspective of biological functions, the biomarkers were proved playing an important role in cancer stage-evolution.

## 2 Methods

### 2.1 Data collection

In the study, we focused on two kinds of cancer, lung adenocarcinoma (LUAD) and breast cancer (BRCA). On one hand, LUAD and BRCA are both cancer types associated with high levels of

TABLE 2 F1 score and CBR for multi-class classification in stages of LUAD and BRCA at different feature quantity thresholds.

| LUAD | | | BRCA | | |
|---|---|---|---|---|---|
| Number of features | F1 score (%) | CBR | Number of features | F1 score(%) | CBR |
| 1 | 48.7420 | - | 1 | 71.9124 | - |
| 2 | 86.2510 | 37.5090 | 2 | 93.4443 | 21.5318 |
| 3 | 91.2557 | 5.0047 | 3 | 95.3618 | 1.9174 |
| 4 | 92.5050 | 1.2492 | 4 | 96.2117 | 0.8499 |
| 5 | 93.3089 | 0.8039 | 5 | 97.2808 | 1.0691 |
| 6 | 96.1537 | 2.8448 | 6 | 97.4655 | 0.1846 |
| 7 | 96.8517 | 0.6979 | 7 | 98.2629 | 0.7974 |
| 8 | 96.9067 | 0.0550 | 8 | **99.1047** | 0.8417 |
| 9 | 97.4935 | 0.5868 | 9 | 98.3727 | −0.7319 |
| 10 | **97.8804** | 0.3868 | 10 | 99.0001 | 0.6273 |

The bold values represent the best results among the column.

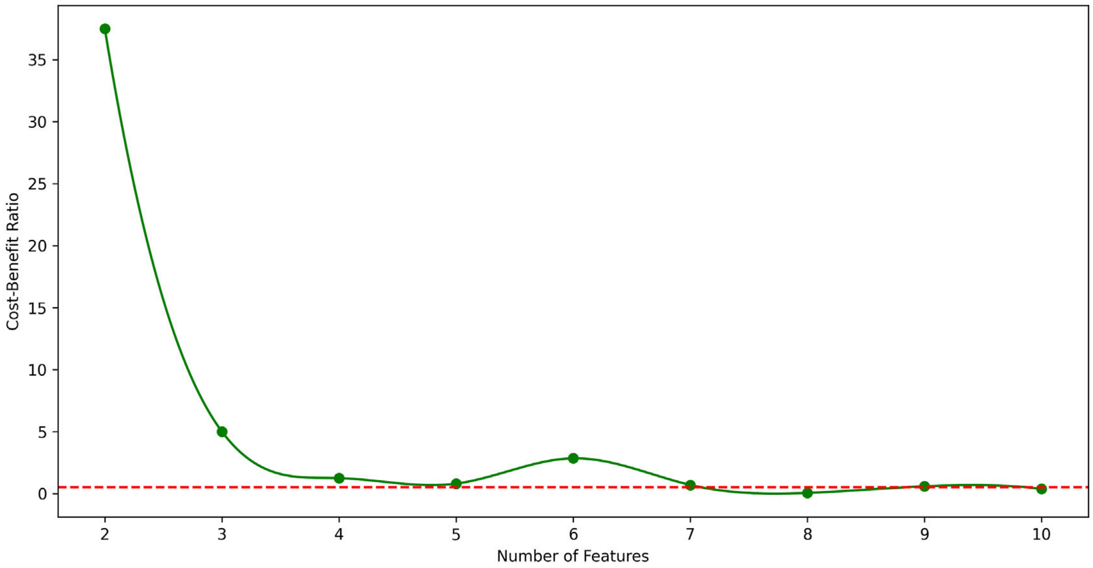


FIGURE 2
The relationship between *CBR* and number of features in LUAD datasets. The red dotted line represents *CBR* = 0.5 for parameter setting. From the figure, the best number of features in LUAD datasets is 7.

severity. LUAD is one of the most common subtypes of lung cancer, while BRCA is one of the most prevalent cancers among women. These two cancer types significantly impact patients' quality of life and survival rates. On the other hand, since LUAD and BRCA are two common types of cancer with relatively high incidence rates worldwide, as a result, these cancer types have ample sample data available. The richness of data helps improve the accuracy and reliability of the models. Therefore, studying and analyzing datasets related to LUAD and BRCA can enhance our understanding of the disease mechanisms, risk factors, and treatment strategies, providing valuable insights for cancer diagnosis and treatment.

We separately collected the clinical data and RNA-Seq data of LUAD and BRCA from Xena Tomczak et al. (2015); Wang et al. (2022) and separated the RNA-Seq data into different pathological stages. Then, the counts per million (CPM) (Law et al., 2016) were applied to filter the low-expression genes, and genes with a value higher than 2 CPM in at least half of the samples were retained. Additionally, the protein-protein interactions were compiled from STRING (Szklarczyk et al., 2023). PPI was widely used in identifying biomolecules, including biomarkers, and driver genes in many studies. The RNA-Seq datasets used in the experiments is shown in Table 1.

## 2.2 Construction of group specific networks

The group specific networks were constructed based on the two main kinds of networks: Sample-Specific Networks (SSN) and PPI

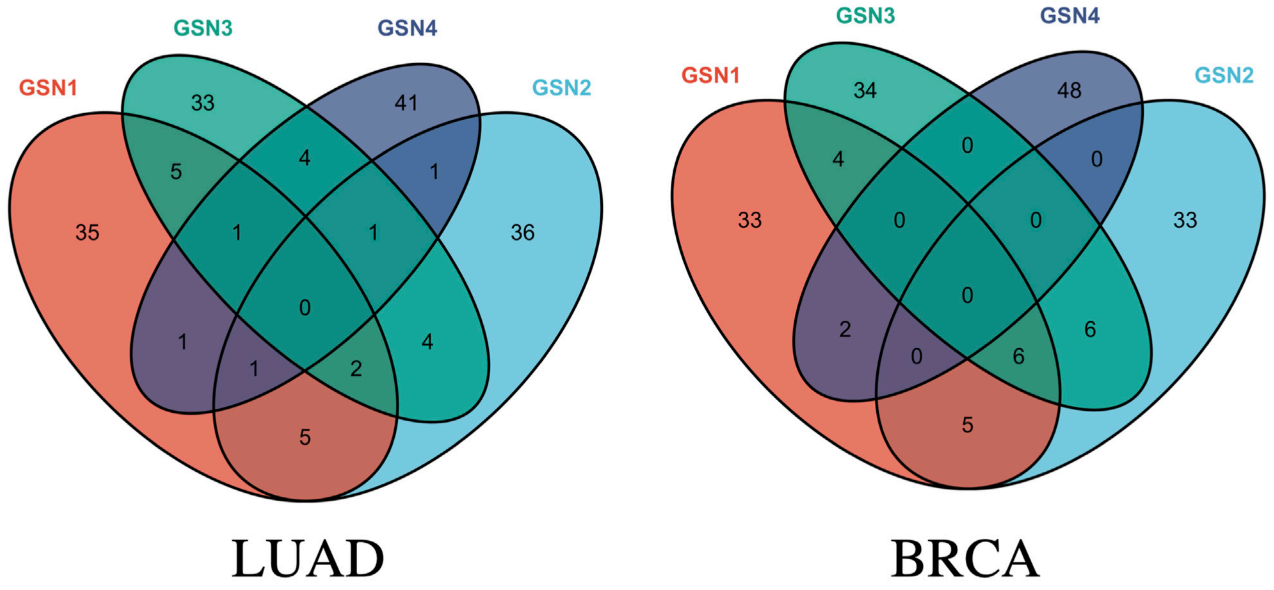


FIGURE 3
Venn graph of the obtained stage-related biomarkers for LUAD and BRCA.

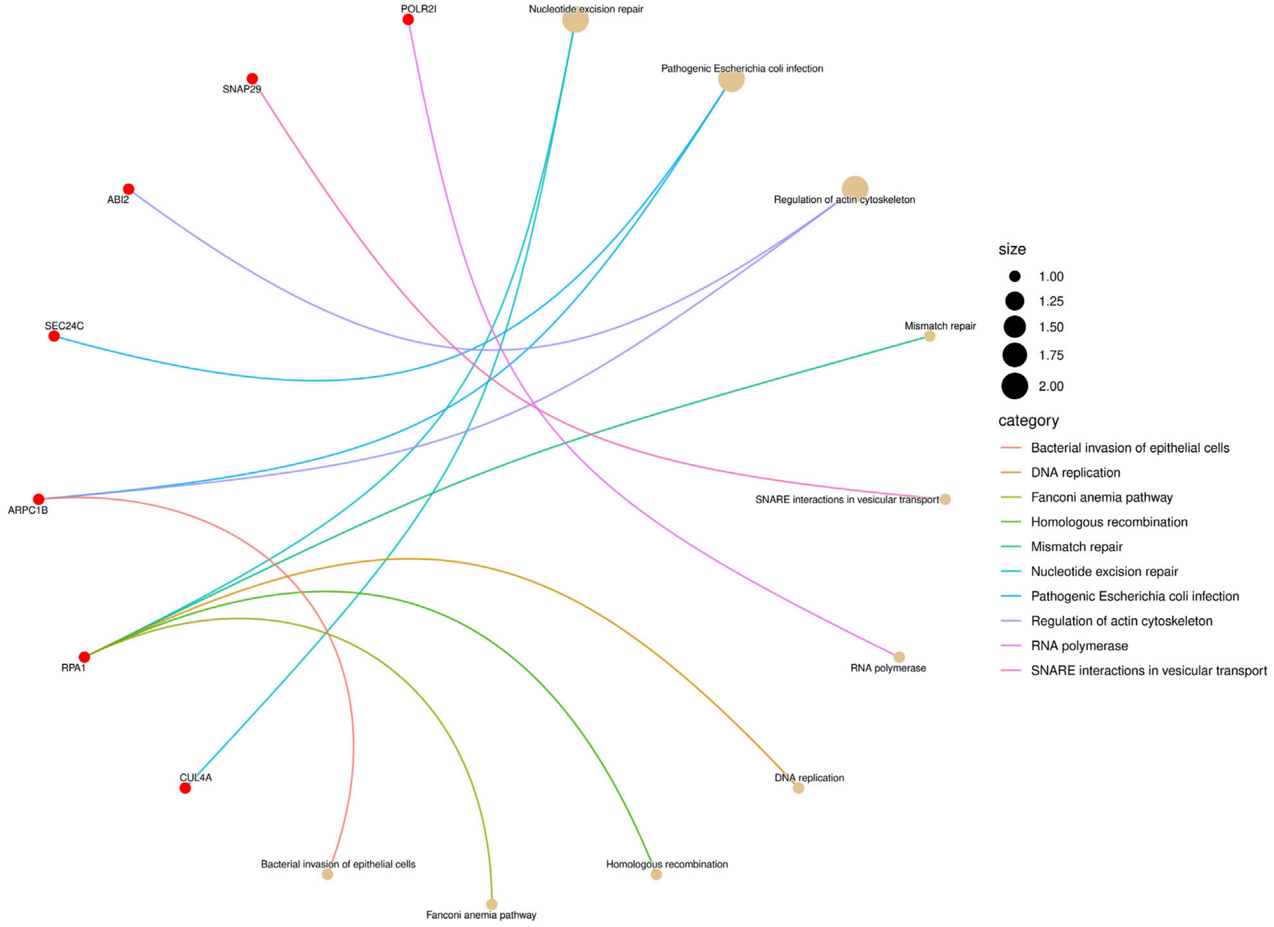

POLR2I
Nucleotide excision repair
SNAP29
Pathogenic Escherichia coli infection
ABI2
Regulation of actin cytoskeleton
SEC24C
Mismatch repair
ARPC1B
SNARE interactions in vesicular transport
RPA1
RNA polymerase
CUL4A
DNA replication
Bacterial invasion of epithelial cells
Homologous recombination
Fanconi anemia pathway
size
1.00
1.25
1.50
1.75
2.00
category
Bacterial invasion of epithelial cells
DNA replication
Fanconi anemia pathway
Homologous recombination
Mismatch repair
Nucleotide excision repair
Pathogenic Escherichia coli infection
Regulation of actin cytoskeleton
RNA polymerase
SNARE interactions in vesicular transport

FIGURE 4
KEGG pathway enrichment result of LUAD stage-related biomarker.

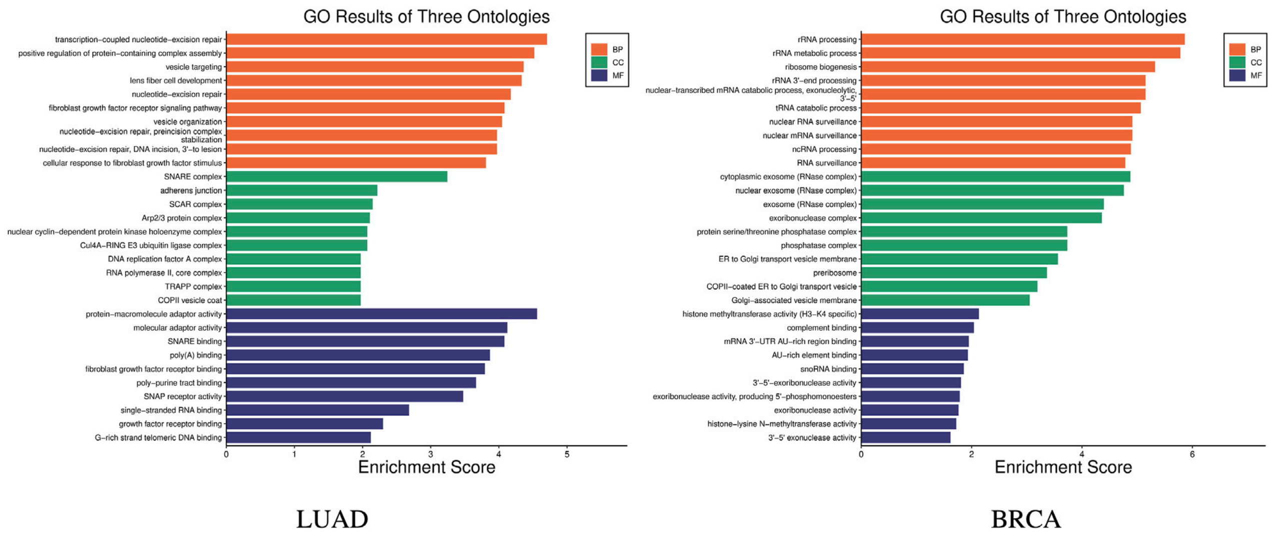


FIGURE 5
GO enrichment results of stage biomarkers for LUAD and BRCA.

**TABLE 3 The comparison of identification in stage-related biomarkers among SFR-GSN, three conventional methods, and all genes on LUAD and BRCA datasets.**

| Methods | LUAD | | BRCA | |
|---|---|---|---|---|
| | Number of features | F1 score(%) | Number of features | F1 score(%) |
| All Genes | 1,3326 | 38.90 | 1,3168 | 42.77 |
| DEGs | 225 | 42.42 | 318 | 42.51 |
| WGCNA | 151 | 40.35 | 396 | 43.89 |
| Relife | 100 | 42.29 | 100 | 43.49 |
| SFR-GSN | 7 | **96.85** | 5 | **97.28** |

The bold values represent the best results among the column.

**TABLE 4 F1 score and CBR for multi-class classification in stages of LUAD and BRCA at different feature quantity thresholds.**

| LUAD | | | BRCA | | |
|---|---|---|---|---|---|
| Number of features | F1 score(%) | CBR | Number of features | F1 score(%) | CBR |
| 1 | 73.2800 | - | 1 | 51.2668 | - |
| 2 | 91.3155 | 18.0354 | 2 | 81.9457 | 30.6788 |
| 3 | 95.9758 | 4.1303 | 3 | 89.6415 | 7.6958 |
| 4 | 96.8973 | 0.5300 | 4 | 93.2176 | 3.5760 |
| 5 | 97.3499 | 0.9214 | 5 | 94.4989 | 1.2813 |
| 6 | 97.7847 | 0.4526 | 6 | 96.4546 | 1.9557 |
| 7 | 97.7847 | 0.4347 | 7 | 96.9095 | 0.4549 |
| 8 | 97.7847 | 0 | 8 | 97.3826 | 0.4731 |
| 9 | 97.7847 | 0 | 9 | 97.3919 | 0.0093 |
| 10 | **98.2410** | 0.4563 | 10 | **97.7572** | 0.3652 |

The bold values represent the best results among the column.

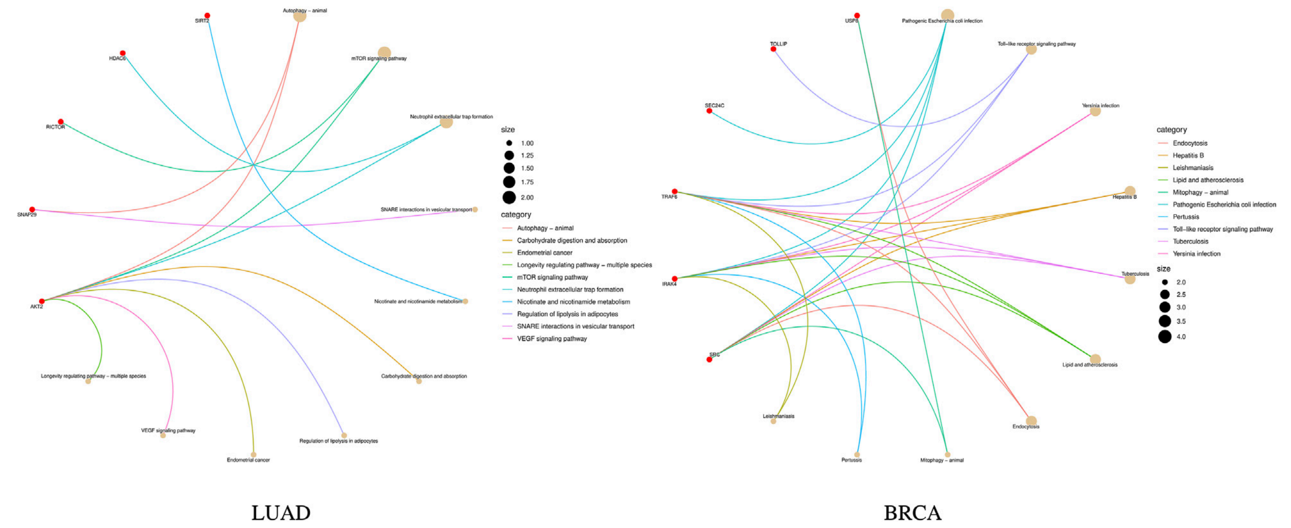


**FIGURE 6**
KEGG pathway enrichment results of subtype biomarkers for LUAD and BRCA.

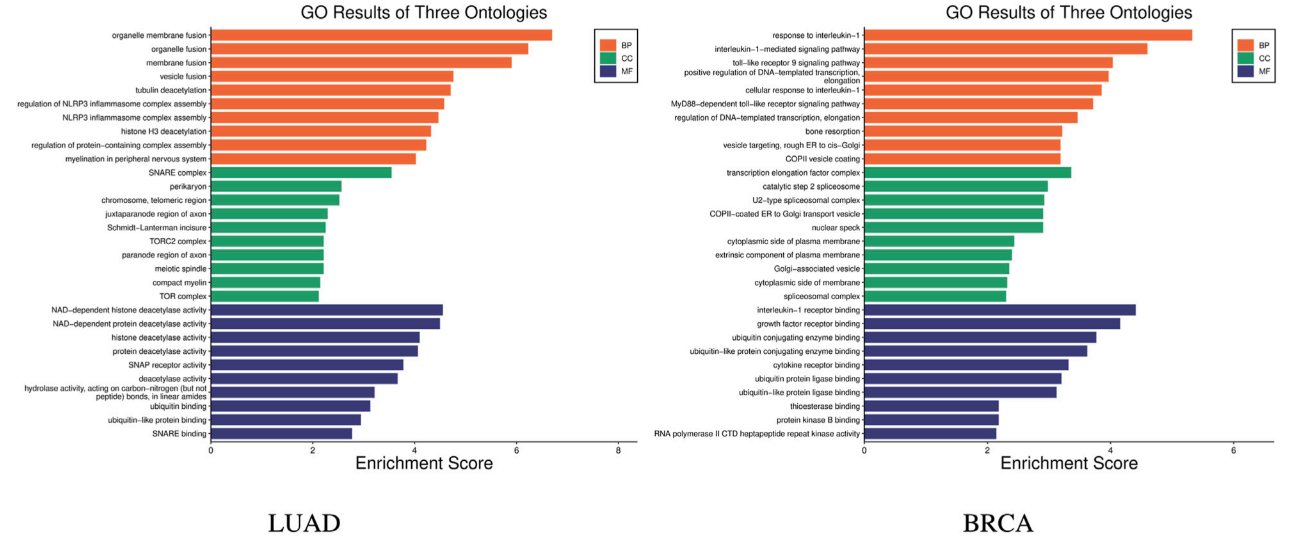


**FIGURE 7**
GO enrichment results of subtype biomarkers for LUAD and BRCA.

networks. Proposed by Liu et al. (2016), SSN could assist in identifying driver genes from the perspective of the personalized network. GSN, combined SSN, and the existing PPI could increase the robustness of the interactions. The flow of the construction of GSN was summarized in Figure 1.

SSN was initially constructed based on RNA-Seq data. For all normal samples, a reference network was constructed by calculating the pairwise gene-gene Pearson correlation coefficients (*PCC*, represented in the reference network as $PCC_n$). Meanwhile, for each disease sample, a perturbation network was generated by incorporating the normal sample set and reconstructing the network, resulting in $PCC_{n+1}$. Subsequently, the differential network was obtained by subtracting the perturbation network from the reference network, and the difference was derived as Formula (1).

$$\triangle PCC = PCC_{n+1} - PCC_n \tag{1}$$

Edges with a statistical *p-value* < 0.05 were considered significant and retained. In the constructed SSN, nodes represent genes, while the connections between nodes indicate significant differences in the correlation between the two genes in the disease sample compared to the normal sample set. This dissimilarity is quantified by $\triangle PCC$.

Then, on the basis of the SSN, intersections were combined with the PPI. We retained the experimentally validated edges presented in PPI, with the edge weight calculated from SSN. Due to the samples could be divided into different pathological stages, the PPI-SSN for all samples was classified according to different stages (groups) of cancer. For instance, within one specific cancer group $G_i$ responding to one GSN, consisting of $N$ cancer samples, the $N$ PPI-SSNs were integrated. Also, the edge weight was calculated by taking an average on the $\triangle PCC$ of same edge in $N$ different samples. As for the edges not appearing in the samples, their $\triangle PCC$ was set to 0. Finally, the edge weight of GSN was derived as Formula (2).

$$w = \frac{\sum_{i=1}^{N} \triangle PCC_i}{N} \tag{2}$$

Considering the generalization of GSN, a ten-fold cross-validation approach was employed during the experimental process. A GSN was constructed for each training fold, resulting in ten GSNs, and the edge weight from these ten GSNs was also averaged. Ultimately for every cancer group $G_i$, only one corresponding GSN was constructed, which is stage-specific.

## 2.3 Specific feature recognition

Based on the constructed GSN, we aimed to identify the most representative and minimal set of features as biomarkers. These features in the selected set contain a high degree of complementary information, resembling a minimal control network. Feature recognition consists of two main parts: pre-screening and optimal subset screening of biomarkers.

### 2.3.1 Pre-screening of biomarkers

The edge set of each GSN corresponding to each group is sorted in descending order based on the edge weights and subjected to pre-screening to obtain the top 50 edges. Among the top 50 edges, the features at both ends of these selected edges are obtained, and their union forms the candidate feature set. Then the candidate feature set is further filtered using the feature importance calculation algorithm embedded in random forest (Acharjee et al., 2020), narrowing it down to a new candidate feature set, which containing only the top 50 features based on their importance rankings. During the feature pre-screening, the Gini impurity was introduced to quantify the purity improvement achieved through branching. The Gini impurity, presented as *Gini*, could be derived as Formula (3).

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{3}$$

where $p_i$ represents the relative frequency of the $i$-th class in the dataset, which is the probability of that class occurring in the dataset, and $n$ is the total number of categories.

In random forest, the calculation of feature importance is based on the Gini impurity of each feature at each node in every tree. Specifically, for each feature, at each node of each tree, the algorithm splits the dataset into two subsets based on that feature. Then, the difference between the Gini impurity of the subsets after the split and the Gini impurity of the original node was calculated. Finally, by aggregating the feature importance scores from all nodes, the overall feature importance for each feature in the random forest was obtained. The built-in feature importance evaluation capability of the random forest makes it a powerful tool for understanding data and extracting key biomarkers in multi-class classification tasks. The whole pre-screening procession was described in Algorithm 1.

```
Require: Random forest model $RF$, training data set $D$;
Ensure: A list of feature importances $importance$;
for each $tree$ in $RF$ do
    for each $node$ in $tree$ do
        for each $feature\ f$ in $node$ do
            Split the dataset at $node$ into two subsets $D_{left}$ and $D_{right}$ based on feature $f$;
            Calculate the Gini impurity of the original node, denoted as Gini;
            Calculate the Gini impurity of $D_{left}$, denoted as $Gini_{left}$;
            Calculate the Gini impurity of $D_{right}$, denoted as $Gini_{right}$;
            Calculate the gain in impurity after splitting on feature $f$:
            $impurityGain = \frac{|D_{left}|}{|D|} \times (Gini - Gini_{left}) + |D \frac{right|}{|D| \times (Gini - Gini_{right})}$;
            Update the importance of $feature\ f$ based on the impurity gain:
            $importance\ [f] \leftarrow importance\ [f] + impurityGain$;
        end for
    end for
end for
Sort the features based on the values in $importance$ using a suitable sorting algorithm.
```

**Algorithm 1. Pre-screening of biomarkers by random forest feature importance calculation.**

### 2.3.2 Optimal subset screening of biomarkers

After the pre-screening, the top 50 candidate feature sets were further filtered by Recursive Feature Elimination with Cross-Validation (RFECV). The RFECV algorithm finds the optimal feature subset by iteratively removing features, involving model training and cross-validation for each reduced feature set. In each iteration, the algorithm removes the least important feature (the one contributing the least to the model's performance improvement), retrains the model on the remaining feature set, and performs cross-validation. This process continues until a specific number of features is reached or further removal of features significantly degrades model performance.

Notably, to select the minimum number of features that achieve the best predictive performance, the Cost-Benefit Ratio (CBR) was introduced to assist in screening the optimal feature set (De Picker and Haarman, 2021). The CBR could be defined as Formula (4).

$$CBR = \frac{100 \times PR}{INF \times UFC} \tag{4}$$

in this formula, the symbols represent the following:

- *PR*: Performance Gain, which refers to the improvement of the *F*1 score in the model.
- *INF*: Increased Number of Features.
- *UFC*: Unit Feature Cost.

Through CBR, we can quantitatively evaluate whether the performance improvement gained from adding specific features is worth the additional cost required. It is particularly important in situations where there is a need to balance decisions between performance improvement and cost.

During the model training, multiple thresholds (*thresh*) were set for the number of features and obtained their corresponding model performance evaluation metric, *F*1 score. Then, according to the CBR model, the optimal feature subset was screened in a recursive way. The optimal subset screening of biomarkers using RFECV was presented as Algorithm 2.

```
Require: candidate feature set, threshold for the number of features $thresh$;
Ensure: feature set $S$, model performance evaluation metric $F1$ score;
    Initialize the feature set $S$ and set it as the candidate feature set;
    Define the model performance evaluation metric $F1$ score;
    Define the threshold for the number of features $thresh$;
    while $S$ is not $\emptyset$ do
        Train the model using the feature set as the training set;
        Introduce cross-validation to evaluate the model performance;
        if the number of features = = $thresh$ then
            Save the current feature set as $S$;
            Save the current model performance metric as $F1$ score;
            break;
        end if
        Reove the least contributing features from $S$;
    end while
```

**Algorithm 2. Recursive Feature Elimination with Cross-Validation (RFECV) algorithm.**

# 3 Results

The experimental results were obtained using ten-fold cross-validation to ensure reliability. In each round, nine folds of the datasets were treated as a train set and the other one fold acted as a test set. The train set was used to construct the GSN and select the feature. The test set was utilized to evaluate the model performance. In addition, specific feature experiments and comparative analyses were conducted to validate the effectiveness of the model. Moreover, the proposed method was expanded to identify cancer subtypes related biomarkers as well.

## 3.1 Specific feature experiments

Specific feature experiments were conducted in the following two steps. Firstly, the important parameters were introduced including the CBR and number of features. Secondly, the stage-specific biomarkers in LUAD and BRCA datasets were identified. The effectiveness of the identified biomarker were performed through enrichment analysis.

### 3.1.1 Setting of the important parameter

CBR was designed as a key parameter to assist in screening the optimal feature set, which is directly related to the number of features. The proposed methods were conducted on LUAD and BRCA datasets to determine a series of feature counts, and the F1 scores and CBRs were calculated through the experiments which was summarized in Table 2.

From the table, it is shown that in LUAD datasets, as the number of features increases, the F1 score generally improves, but the *CBR* shows non-monotonic variations. Therefore, to further illustrate the relationship between *CBR* and the number of features, their relationship in LUAD datasets was plotted in Figure 2. In the figure, the *CBR* values were compared with 0.5, as this threshold is often used as a balancing point. When the *CBR* is greater than 0.5, it indicates a profitable decision, while a *CBR* lower than 0.5 suggests a cost-effective decision.

Therefore, the *CBR* metric was utilized to determine the optimal number of features.

Starting with a small number of features and gradually increasing, the point was identified where the first *CBR* value fell below 0.5.

The CBR indicates the overall benefit of adding a new feature to the model. Therefore, the feature count just before this point was identified as the optimal number of features.

### 3.1.2 Stage-specific biomarkers

Based on the parameter setting, features with CBR values greater than 0.5 were selected to maximize the F1 score. The obtained biomarkers were in the form of gene pairs or edges.

Compared with the node features, the edge biomarkers could better capture the interaction relationships between genes, aiding in understanding the structure and functionality of gene networks.

The edge features could reflect the interplay and coordinated regulation among genes, revealing more details about biological processes and disease development.

As for the LUAD dataset, seven features were eventually identified that meet this criterion, achieving an impressive F1 score at 96.8517% and a CBR at 0.6979. These features include: (ABI2, ARPC1B), (CDK12, POLR2I), (FRS2, FRS3), (PABPC4, ZC3H14), (SNAP29, TSNARE1), (SEC24C, TRAPPC6B), and (CUL4A, RPA1). Similarly, for the BRCA dataset, five features were selected that yielded a remarkable F1 score at 97.2808% and a CBR at 1.0691. These features are: (EXOSC3, SKIV2L2), (BYSL, UTP14C), (EXOSC8, UTP14C), (PPP3CB, WDR82), and (CD59, SEC24C). The Venn graph of the obtained biomarkers is shown in Figure 3, which demonstrates the biomarkers were highly stage-specific.

## 3.2 Enrichment analysis

Moreover, the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway analysis and Gene Ontology (GO) enrichment analysis were performed to validate the effectiveness of identified biomarkers.

KEGG pathway enrichment analysis is a frequently employed method in bioinformatics to interpret gene expression or protein expression data (Ogata et al., 1999). After performing a significance test on 14 genes in the biomarkers of LUAD stages, a total of seven

**TABLE 5** The comparison of identification in subtype-related biomarkers among SFR-GSN, three conventional methods, and all genes on LUAD and BRCA datasets.

| Methods | LUAD | | BRCA | |
|---|---|---|---|---|
| | Number of features | F1 score(%) | Number of features | F1 score(%) |
| All Genes | 1,3326 | 72.73 | 1,3168 | 85.9674 |
| DEGs | 2,478 | 82.38 | 3,922 | 87.77 |
| WGCNA | 426 | 78.00 | 632 | 86.32 |
| Relife | 100 | 81.16 | 100 | 83.45 |
| SFR-GSN | 5 | **96.89** | 6 | **96.45** |

The bold values represent the best results among the column.

genes were found to be enriched in 10 pathways. Among them, the gene RPA1 was found to be involved in five pathway processes, as shown in Figure 4. In the figure, the red dots represent genes, and the different colored curves represent different pathways. One end of the curve represents a gene, while the other end represents the hub of that pathway, and the size of the hub is proportional to the number of genes enriched in that pathway. As for stage-related biomarkers of BRCA, a total of three genes were found to be enriched in two pathways. Specifically, genes EXOSC8 and EXOSC3 were enriched in hsa03018: RNA degradation, while gene PPP3CB was enriched in hsa04370: VEGF signaling pathway. Due to the small number of genes, they were not visualized.

GO (Gene Ontology) enrichment analysis was carried out to help understand the roles of a set of genes in biological processes Harris et al. (2004). GO enrichment analysis was carried out on the 14 genes in the stage-related biomarkers of LUAD, and the results are shown in Figure 5 LUAD, which indicates these 14 genes are involved in a total of 240 biological processes. In the figure, the *x*-axis represents the enrichment score, which indicates the degree of influence of the target genes on the corresponding GO term, while *y*-axis represents the different GO terms. The different colors represent the three main categories of GO. Each category includes only the top 10 terms based on their enrichment score. Similarly, GO enrichment analysis was performed on the nine genes in the stage-related biomarkers of BRCA, and the results are shown in Figure 5 BRCA. These nine genes were found to participate in a total of 205 biological processes.

The enrichment results demonstrate significant specificity of the features constructed using our proposed method across different stages within the two major cancer types, LUAD and BRCA. The evidence further validates the effectiveness of the proposed method.

## 3.3 Comparative experiments

Comparative experiments were conducted mainly in view of stage-related biomarker prediction. The proposed method was compared with the three conventional methods on biomarker identification: differentially expression genes (DEGs) Love et al. (2014), WGCNA Horvath (2011) and RelifF Robnik-Šikonja and Kononenko (2003). DEGs were mainly obtained using R package DESeq2 to conduct differential expression analysis, and the DEGs were treated as biomarkers. Based on differential expression data, WGCNA (Weighted Gene Co-expression Network Analysis) is a method used to construct co-expression networks from gene expression data, which is currently widely applied in the identification of biomarkers for complex diseases and drug targets. RelifF is a machine learning method on binary classification, which could identify the biomarkers.

Moreover, to ensure an equivalent comparison, the four methods were compared based on their best performance. Also, the features in all genes were performed as a control group. The F1 score was employed for evaluation since it is not influenced by the varying number of features across different methods. The results of the comparative experiments on LUAD datasets and BRCA datasets are shown in Table 3.

From the table, it is shown that the proposed method significantly outperforms other methods in terms of F1 scores. Additionally, the proposed method provides fewer features than other methods, which indicates the proposed method could identify the biomarker more accurately.

## 3.4 SFR-GSN on cancer subtype-related biomarkers

Besides the evolutionary characteristics in different stages, cancer also exhibits various subtypes. As for LUAD, three types often occur in the evolution, which are Papillary Predominant (PP), Acinar Predominant (PI), and Trabecular (TRU). By studying subtype-related biomarkers, a better understanding of the differences in disease progression, treatment response, and prognosis among different subtypes could be obtained (Perou et al., 2000; Muller et al., 2022). Therefore, in order to enhance the generalization of our model, experiments on subtype-related data were conducted to identify the subtype-related biomarkers.

Firstly, the datasets were separated into the three subtypes and accordingly, three corresponding GSNs were constructed. Then, SFR was trained on the GSNs, features with CBR > 0.5 were obtained, and the F1 score and CBR were shown in Table 4. Finally, five features were identified as subtype-related biomarkers of LUAD, these are (HDAC6, SIRT2), (AKT2, RICTOR), (DHX33, PINX1), (SNAP29, TSNARE1) and (ASPSCR1, VCPIP1). Similarly, the BRCA datasets were divided into five groups due to the five subtypes of BRCA. Eventually, the results were shown in Table 4, where six features were screened as subtype-related biomarkers, these are (SRC, USP8), (IRAK4, TOLLIP), (SRC, TRAF6), (F8, SEC24C), (CDK12, SUPT5H) and (CDC40, SF3B2).

Further, the enrichment analysis was conducted on the identified features. In the subtype-related biomarkers of LUAD, five genes were enriched in 10 pathway pathways, with the gene AKT2 was found in eight pathway pathways. In that of BRCA, eight genes were enriched in 24 pathways, with the gene TRAF6 being enriched in 21 pathways and the gene IRAK4 was found in 20 pathways. KEGG pathway enrichment results are shown in Figure 6. Subsequently, the results of the GO enrichment analysis are shown in Figure 7. The 10 genes in the LUAD subtypes are involved in 360 biological processes, while 11 genes in the BRCA subtypes are involved in 407 biological processes.

After providing the results of SFR-GSN on the identification, the proposed method was also compared with three conventional methods and all genes. The results are shown in Table 5. SFR-GSN gains the best performance and the least features, which suggests SFR-GSN exhibits superior capability in identifying subtype-related biomarkers.

## 4 Conclusion

In this work, a novel method called SFR-GSN has been proposed to identify the stage-related biomarkers, which gained remarkable results on LUAD and BRCA datasets. First, the clinical data, RNA-Seq data, and PPI were collected. Second, according to the pathological stage, the GSNs were constructed by combining the SSN and PPI. Third, based on GSNs, a two-round SFR was conducted, which firstly used random forest to pre-screen and later used RFECV to obtain the optimal feature sets. The CBR was introduced to assist in identifying stage-related biomarkers.

Finally, the results of the proposed method showed that the identified biomarkers were highly stage-specific and significantly improved the F1 scores for cancer stage prediction. For the lung adenocarcinoma dataset, the F1 score reached 97.68%, and for the breast cancer dataset, it achieved 96.87%. The results outperform the other conventional methods on both accuracy and F1 scores. Moreover, the enrichment analysis of biomarkers was conducted to validate the effectiveness of the proposed method in view of biological functions. The proposed method exhibits superior performance in identifying subtype-related biomarkers. The proposed method could be applied to other cancers to offer new insight into cancer treatment prognosis.

## Data availability statement

The RNA-Seq data presented in the study are deposited in the UCSC Xena repository, accession number TCGA Lung Adenocarcinoma (LUAD) and TCGA Breast Cancer (BRCA), the url is https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA.BRCA.sampleMap%2FHiSeqV2.gz; the PPI data presented in the study are deposited in the STRING repository, accession number Homo sapiens, the url is https://stringdb-downloads.org/download/protein.physical.links.v12.0/9606.protein.physical.links.v12.0.txt.gz.

## Author contributions

BC: Data curation, Formal Analysis, Funding acquisition, Methodology, Supervision, Writing–review and editing, Writing–original draft. YW: Data curation, Formal Analysis, Methodology, Validation, Visualization, Writing–original draft. JZ: Methodology, Visualization, Writing–review and editing. YH: Methodology, Writing–review and editing. HB: Methodology, Writing–review and editing. Jun Bian: Supervision, Writing–review and editing. RK: Supervision, Writing–review and editing. XS: Supervision, Writing–review and editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Key R&D Program of China under Grant No. 2021YFA1000402, the National Natural Science Foundation of China under Grant No. 61972320, and Xi'an municipal bureau of science and technology under Grant No. 22YXYJ0057.

## Conflict of interest

RK was employed by Rewise (Hangzhou) Information Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Acharjee, A., Larkman, J., Xu, Y., Cardoso, V. R., and Gkoutos, G. V. (2020). A random forest based biomarker discovery and power analysis framework for diagnostics research. *BMC Med. genomics* 13, 178. doi:10.1186/s12920-020-00826-6

Amin, D. N., Ngoyi, D. M., Nhkwachi, G.-M., Palomba, M., Rottenberg, M., Büscher, P., et al. (2010). Identification of stage biomarkers for human african trypanosomiasis. *Am. J. Trop. Med. Hyg.* 82, 983–990. doi:10.4269/ajtmh.2010.09-0770

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi:10.1038/nature12625

De Picker, L. J., and Haarman, B. C. (2021). Applicability, potential and limitations of tspo pet imaging as a clinical immunopsychiatry biomarker. *Eur. J. Nucl. Med. Mol. imaging* 49, 164–173. doi:10.1007/s00259-021-05308-0

Deva Magendhra Rao, A. K., Patel, K., Korivi Jyothiraj, S., Meenakumari, B., Sundersingh, S., Sridevi, V., et al. (2019). Identification of lnc rna s associated with early-stage breast cancer and their prognostic implications. *Mol. Oncol.* 13, 1342–1355. doi:10.1002/1878-0261.12489

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids Res.* 32, D258–D261. doi:10.1093/nar/gkh036

Holland, R. L. (2016). What makes a good biomarker? *Adv. Precis. Med.* 1, 66. doi:10.18063/apm.2016.01.007

Horvath, S. (2011) *Weighted network analysis: applications in genomics and systems biology*. Springer Science and Business Media.

Jin, T., Nguyen, N. D., Talos, F., and Wang, D. (2021). Ecmarker: interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. *Bioinformatics* 37, 1115–1124. doi:10.1093/bioinformatics/btaa935

Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., et al. (2016). Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger. *F1000Research* 5, 1408. doi:10.12688/f1000research.9005.2

Liu, X., Wang, Y., Ji, H., Aihara, K., and Chen, L. (2016). Personalized characterization of diseases using sample-specific networks. *Nucleic acids Res.* 44, e164. doi:10.1093/nar/gkw772

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Moein, S. T., Hashemian, S. M., Mansourafshar, B., Khorram-Tousi, A., Tabarsi, P., and Doty, R. L. (2020). Smell dysfunction: a biomarker for covid-19. *Int. forum allergy and rhinology* 10, 944–950. doi:10.1002/alr.22587

Muller, K., Joms, J. M., and Tozbikian, G. (2022). What's new in breast pathology 2022: who 5th edition and biomarker updates. *J. pathology Transl. Med.* 56, 170–171. doi:10.4132/jptm.2022.04.25

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids Res.* 27, 29–34. doi:10.1093/nar/27.1.29

Patil, R., and Bellary, S. (2022). Machine learning approach in melanoma cancer stage detection. *J. King Saud University-Computer Inf. Sci.* 34, 3285–3293. doi:10.1016/j.jksuci.2020.09.002

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *nature* 406, 747–752. doi:10.1038/35021093

Robnik-Šikonja, M., and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Mach. Learn.* 53, 23–69. doi:10.1023/a:1025667309714

Shi, G., Wang, Y., Zhang, C., Zhao, Z., Sun, X., Zhang, S., et al. (2018). Identification of genes involved in the four stages of colorectal cancer: gene expression profiling. *Mol. Cell. probes* 37, 39–47. doi:10.1016/j.mcp.2017.11.004

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA a cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., et al. (2023). The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids Res.* 51, D638–D646. doi:10.1093/nar/gkac1000

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp. Oncology/Współczesna Onkol.* 2015, 68–77. doi:10.5114/wo.2014.47136

Ubaldi, L., Valenti, V., Borgese, R., Collura, G., Fantacci, M., Ferrera, G., et al. (2021). Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. *Phys. Medica* 90, 13–22. doi:10.1016/j.ejmp.2021.08.015

Van der Kloet, F., Tempels, F., Ismail, N., Van der Heijden, R., Kasper, P., Rojas-Cherto, M., et al. (2012). Discovery of early-stage biomarkers for diabetic kidney disease using ms-based metabolomics (finndiane study). *Metabolomics* 8, 109–119. doi:10.1007/s11306-011-0291-6

Wang, J., Song, J., Gao, Z., Huo, X., Zhang, Y., Wang, W., et al. (2017). Analysis of gene expression profiles of non-small cell lung cancer at different stages reveals significantly altered biological functions and candidate genes. *Oncol. Rep.* 37, 1736–1746. doi:10.3892/or.2017.5380

Wang, S., Xiong, Y., Zhao, L., Gu, K., Li, Y., Zhao, F., et al. (2022). Ucscxenashiny: an r/cran package for interactive analysis of ucsc xena data. *Bioinformatics* 38, 527–529. doi:10.1093/bioinformatics/btab561

Ye, Y., Jing, Y., Li, L., Mills, G. B., Diao, L., Liu, H., et al. (2020). Sex-associated molecular differences for cancer immunotherapy. *Nat. Commun.* 11, 1779. doi:10.1038/s41467-020-15679-x

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 27 May 2024
DOI 10.3389/fnins.2024.1351387

Check for updates

OPEN ACCESS

EDITED BY
Chunliang Wang,
Royal Institute of Technology, Sweden

REVIEWED BY
Edoardo De Rose,
University of Calabria, Italy
Moiz Khan Sherwani,
University of Calabria, Italy

*CORRESPONDENCE
Hui Yang
✉ yanghui2012@jlu.edu.cn
Chunjie Guo
✉ guocj@jlu.edu.cn

RECEIVED 28 December 2023
ACCEPTED 01 May 2024
PUBLISHED 27 May 2024

CITATION
Huang L, Shao Y, Yang H, Guo C, Wang Y, Zhao Z and Gong Y (2024) A joint model for lesion segmentation and classification of MS and NMOSD.
*Front. Neurosci.* 18:1351387.
doi: 10.3389/fnins.2024.1351387

COPYRIGHT
© 2024 Huang, Shao, Yang, Guo, Wang, Zhao and Gong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# A joint model for lesion segmentation and classification of MS and NMOSD

Lan Huang[1], Yangguang Shao[1], Hui Yang[2]*, Chunjie Guo[3]*, Yan Wang[1], Ziqi Zhao[1] and Yingchun Gong[1]

[1]College of Computer Science and Technology, Jilin University, Changchun, China, [2]Public Computer Education and Research Center, Jilin University, Changchun, China, [3]Department of Radiology, The First Hospital of Jilin University, Changchun, China

**Introduction:** Multiple sclerosis (MS) and neuromyelitis optic spectrum disorder (NMOSD) are mimic autoimmune diseases of the central nervous system with a very high disability rate. Their clinical symptoms and imaging findings are similar, making it difficult to diagnose and differentiate. Existing research typically employs the T2-weighted fluid-attenuated inversion recovery (T2-FLAIR) MRI imaging technique to focus on a single task in MS and NMOSD lesion segmentation or disease classification, while ignoring the collaboration between the tasks.

**Methods:** To make full use of the correlation between lesion segmentation and disease classification tasks of MS and NMOSD, so as to improve the accuracy and speed of the recognition and diagnosis of MS and NMOSD, a joint model is proposed in this study. The joint model primarily comprises three components: an information-sharing subnetwork, a lesion segmentation subnetwork, and a disease classification subnetwork. Among them, the information-sharing subnetwork adopts a dualbranch structure composed of a convolution module and a Swin Transformer module to extract local and global features, respectively. These features are then input into the lesion segmentation subnetwork and disease classification subnetwork to obtain results for both tasks simultaneously. In addition, to further enhance the mutual guidance between the tasks, this study proposes two information interaction methods: a lesion guidance module and a crosstask loss function. Furthermore, the lesion location maps provide interpretability for the diagnosis process of the deep learning model.

**Results:** The joint model achieved a Dice similarity coefficient (DSC) of 74.87% on the lesion segmentation task and accuracy (ACC) of 92.36% on the disease classification task, demonstrating its superior performance. By setting up ablation experiments, the effectiveness of information sharing and interaction between tasks is verified.

**Discussion:** The results show that the joint model can effectively improve the performance of the two tasks.

KEYWORDS

**MS, NMOSD, joint model, MRI, disease classification, lesion segmentation**

## 1 Introduction

The demyelinating disease of the central nervous system is an autoimmune disease characterized by multifocal and inflammatory demyelination of the central nervous system. Both Multiple sclerosis (MS) and Neuromyelitis optic spectrum disorder (NMOSD) are demyelinating diseases of the central nervous system (Bruscolini et al., 2018; McGinley et al., 2021). MS and

NMOSD may be easily confused clinically due to their overlapping features (Yokote and Mizusawa, 2016).

Magnetic Resonance Imaging (MRI) is a commonly used medical imaging technology in clinical practice (Bauer et al., 2013) for prognosis and treatment response evaluation of MS and NMOSD (Filippi et al., 2016; Rotstein and Montalban, 2019). Furthermore, the T2-weighted fluid-attenuated inversion recovery (T2-FLAIR) sequence can inhibit a certain range of fluid signals, thereby reducing the cerebrospinal fluid signal intensity and enhancing the visibility of small brain lesions and periventricular lesions. Therefore, T2-FLAIR sequence imaging plays a crucial role in the diagnosis of cerebral nervous system diseases (Wattjes et al., 2021).

In order to accurately diagnose MS and NMOSD clinically, it is usually necessary for radiologists to manually segment the white matter high signal presented on MRI images, and then diagnose MS according to the McDonald diagnostic criteria (Wattjes et al., 2021) and NMO diagnostic criteria (Griggs et al., 2016) according to the distribution and morphology of lesions and the clinical manifestations of patients. However, the entire diagnostic process is a time-consuming and onerous task for doctors.

Deep learning has achieved advanced performance in image processing due to a large amount of labeled data, enabling accurate diagnosis of MS and NMOSD (Lee et al., 2017; Lundervold and Lundervold, 2019). The MS and NMOSD auxiliary diagnosis based on deep learning mainly includes two tasks: lesion segmentation and disease classification. The task of lesion segmentation involves identifying and segmenting the lesions according to the high white matter signal observed by the patient's MRI, judging the severity of the patient, and monitoring the course of the disease through quantitative measurement. The task of disease classification aims to accurately diagnose patients, distinguishing between MS and NMOSD according to the shape and distribution characteristics of the lesions.

In most research, the segmentation and classification of MS and NMOSD are studied independently. In order to improve the efficiency and accuracy of the auxiliary diagnosis model, this study analyzes the correlation between lesion segmentation and disease classification tasks and combines the existing deep learning technology to carry out the following research work:

(1) We proposed a joint model of lesion segmentation and disease classification of MS and NMOSD, which is based on the intrinsic correlation between the two tasks and used to segment and classify MS and NMOSD simultaneously. The structure of the joint model (Figure 1) mainly includes three components: an information-sharing subnetwork, a lesion segmentation subnetwork, and a disease classification subnetwork.

(2) We proposed two information interaction methods to improve the performance of lesion segmentation and disease classification tasks in a mutually guided manner, one is a lesion guidance module and the other is a cross-task loss function. Moreover, the lesion location maps provide interpretability for the diagnosis of MS and NMOSD.

The subsequent work consists of four sections, each briefly described as follows: in Section 2 we provide an overview of the relevant research on lesion segmentation and disease classification in MS and NMOSD. In Section 3 we offer a detailed presentation of the joint model for MS and NMOSD in this study. In Section 4 we involve experimental validation of the performance of the joint models proposed in this study. In Section 5 we present the conclusions drawn from this study and provide prospects.

## 2 Related works

Deep learning-based auxiliary diagnosis of MS and NMOSD mainly includes two tasks: lesion segmentation and disease classification. The lesion segmentation task involves identifying and segmenting lesions based on high signal white matter in a patient's MRI, enabling quantitative measurement to assess the severity of the patient's condition and monitor disease progression. The disease classification task aims to diagnose the specific condition a patient has, distinguishing whether the patient has MS or NMOSD based on the morphology and distribution characteristics of the lesions.

Traditional segmentation methods include threshold segmentation algorithms, region-growing algorithms, edge detection algorithms, and watershed algorithms. Among these, threshold segmentation algorithms enhance images based on the differences in signal intensity between healthy brain tissue and lesion regions. These methods involve setting one or more thresholds manually or using algorithms after preprocessing the images to segment them into different parts based on intensity values. For

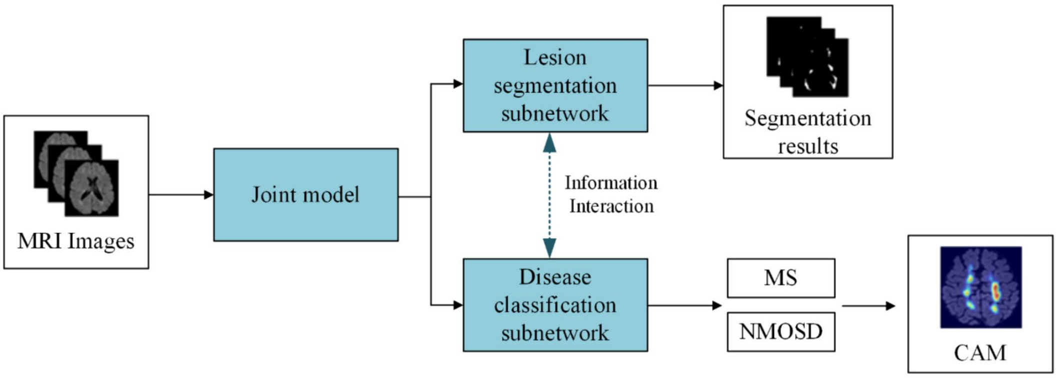


FIGURE 1
The main work content of this study.

instance, Wicks et al. (1992) proposed an intensity-based global threshold segmentation method for MS lesion segmentation. Hence, Wang et al. (1998) addressed the impact of scanner sensitivity on thresholds by proposing histogram matching algorithms. Their study demonstrated that histogram matching significantly reduces dependence on threshold selection for lesion segmentation.

In recent years, deep learning methods have exhibited superior performance in the field of image segmentation (Wang et al., 2022). Particularly, since the introduction of Fully Convolutional Networks (FCN) (Shelhamer et al., 2017), which can produce probability prediction maps of the same size as the original images without restricting input image size, there has been significant progress in the image segmentation task. One of the most classical models in medical image segmentation task is UNet (Ronneberger et al., 2015). UNet combines lower-level detailed information with higher-level semantic information through its encoder-decoder structure and skip connections. Building upon UNet, Bauer et al. (2013) replaced the convolutional blocks in UNet with dense blocks Huang et al. (2017), enabling features reuse across channel dimensions, and leading to a more accurate and easily trainable network. Zhou et al. (2020) proposed the UNet++, which enhances the skip connection structure to aggregate features from various scales in the decoder sub-network, thereby improving model flexibility.

As medical images like MRI are often three-dimensional, one approach involves slicing the images into 2D slices along specific dimensions, training 2D models, and then reassembling the segmentation results into a 3D format (Tseng et al., 2017). Aslani et al. (2019) designed an end-to-end encoder-decoder network. They divided MRI images of MS patients into 2D slices along three dimensions, inputting them into multiple 2D segmentation models, and reassembled the resulting 2D segmentations into a 3D format using a majority voting approach. Zhang et al. (2019) proposed a method using 3D stacked slices that combine information from adjacent slices in multiple channels, increasing inter-slice information. To better extract inter-slice information, Çiçek et al. (2016) replaced 2D convolutional operations with 3D convolutions, fully utilizing information within and between image slices. Building upon this, La Rosa et al. (2020) proposed 3D UNet-, segmenting cortical and white matter lesions based on FLAIR and MP2RAGE sequences of MS patients. Hu et al. (2020) introduced a three-dimensional context-guided module in the encoding and decoding stages of 3D UNet, expanding the perceptual field, guiding contextual information, and enriching feature representations of MS lesion segmentation using a three-dimensional spatial attention block. Gessert et al. (2020) proposed a dual-path 3D convolutional structure with attention-guided interaction, separately processing MS data from two-time points and effectively exchanging information.

Due to the similarities between MS and NMOSD, classifying MS and NMOSD is also a critical step in auxiliary diagnosis.

Imaging-based classification methods using handcrafted features involve constructing a feature set from digital medical images and subsequently employing machine learning models for analysis. These methods typically require experienced radiologists to manually extract high-dimensional image data into low-dimensional handcrafted features. These features, along with relevant clinical variables, are used to create a feature set. Feature selection is then performed, and an optimal subset of features is utilized to build a predictive model. Huang (2019) extracted 273 radiomic features from the lesion area of patients' brain T2-weighted images, including semantic, intensity, and texture features. They incorporated 11 radiomic features using the LASSO method, combined with 5 clinical features, to construct a diagnostic radiomic signature, achieving an AUC result of 0.93 on the test set. Kister et al. (2013) conducted quantitative analysis on the shape and distribution of localized T2 white matter lesions based on clinical brain MRI sequences of 44 AQP4-IgG antibody-positive NMOSD patients and 50 MS patients, creating a diagnostic procedure for classifying MS and NMOSD. Liu et al. (2019) extracted 9 features, including lesion heterogeneity and lesion volume, from patient imaging data, combined with clinical information, to build a logistic regression model to differentiate MS and NMOSD. However, these imaging-based methods rely heavily on radiologists manually extracting imaging features, limiting the repeatability and generalizability of these methods.

Deep learning-based classification methods efficiently capture classification features automatically without requiring manual extraction and selection. For instance, Hagiwara et al. (2021) developed an automatic classification model based on Convolutional Neural Networks (CNN). Due to limited available data, they primarily utilized SqueezeNet to prevent overfitting, achieving an accuracy of 0.81, sensitivity of 0.80, and specificity of 0.83 using common features to classify MS and NMOSD. Wang et al. (2020) compressed 3D MS and NMOSD MRI images into multi-channel 2D images and used a 2D ResNet model for classification. By leveraging transfer learning subsequent to pre-training the model on ImageNet, they attained an accuracy of 0.75. Kim et al. (2020) developed a 3D CNN model based on the ResNeXt concept to differentiate MS and NMOSD, which effectively utilized MRI spatial features and achieved improved performance with an accuracy of 0.71, sensitivity of 0.87, and specificity of 0.61 when integrating clinical information.

From existing research, it is evident that deep learning-based methods generally exhibit promising results in the task of classifying MS and NMOSD, often without extensive involvement from radiologists, thus possessing considerable practical value.

# 3 Materials and methods

## 3.1 Datasets and evaluation metrics

### 3.1.1 Datasets

The datasets used in this study are MS and NMOSD MRI datasets. The MS datasets come from the Multiple Sclerosis Lesion Segmentation Challenge organized by the 2015 IEEE International Symposium on Biomedical Imaging (ISBI) (referred to as the ISBI dataset) and The First Hospital of Jilin University. The NMOSD dataset comes from the First Hospital of Jilin University.

Table 1 outlines the composition of the MS and NMOSD datasets utilized in this study. The ISBI dataset contains brain MRI images of 5

TABLE 1 MS and NMOSD datasets.

| Disease classification | Data sources | Number of samples | Sequence type |
|---|---|---|---|
| MS | ISBI 2015 | 21 | T2-Flair |
| | The First Hospital of Jilin University | 48 | T2-Flair |
| NMOSD | The First Hospital of Jilin University | 62 | T2-Flair |

TABLE 2 MRI parameters were collected from the First Hospital of Jilin University.

| Parameter type | Parameter values |
|---|---|
| Slice number | 160–192 |
| Slice thickness (mm) | 1 |
| Repeat time (ms) | 4,800 |
| Echo time (ms) | 279–324 |
| Reverse time (ms) | 1,650 |
| Flip angle (°) | 40 |

MS patients scanned at different time points, among which 4 patients scanned 4 groups of images and 1 patient scanned 5 groups of images, a total of 21 groups of image data, each group of images can utilize as a separate sample. Each group of images includes MRI images of four modalities: T1WI, T2WI, PD, and T2-FLAIR. This study only utilized the data of the T2-FLAIR modality. All the MRI images were scanned using a 3.0 Tesla MRI scanner and were registered to standard space, with the image size normalized to $181 \times 217 \times 181$. At the same time, the ISBI dataset provides a lesion segmentation map manually annotated by experts corresponding to each image.

The brain MRI images of T2-FLAIR sequences obtained from the First Hospital of Jilin University comprised 48 MS samples and 62 NMOSD samples. Among them, all MS samples met the McDonald diagnostic criteria revised in 2017, and all NMOSD patients met the NMO diagnostic criteria revised in 2015. Due to the long collection time of the datasets, there are some differences in the collection parameters among images, as shown in Table 2.

At the same time, the First Hospital of Jilin University provided lesion annotation maps for all images that were manually annotated by two radiologists with 5 and 10 years of experience in diagnosing brain diseases respectively, and were finally combined with the annotations of the two doctors. The intersection of the results is utilized as the annotation map of this study. Therefore, the datasets utilized in this study are a total of 69 MS samples and 62 NMOSD samples, all of which are T2-FLAIR modality.

### 3.1.2 Data preprocessing

The ISBI dataset provides preprocessed images, which have been skull-stripped and registered to the MNI template. Therefore, we preprocessed the datasets provided by the First Hospital of Jilin University. The specific preprocessing steps are divided into the following four stages: Firstly, skull stripping was performed on the brain MRI images using the Brain Extraction Tool (BET) (Smith, 2002). Secondly, the images were corrected for bias field using the N4 bias field correction method (Tustison et al., 2010). Thirdly, the black background area of the images was removed. Finally, the size of the images is normalized to (160, 160, 160), while the voxel values of the images are normalized to a standard data distribution with a mean of 0 and a standard deviation of 1 using the Z-Score method.

### 3.1.3 Data augmentation

Due to the challenges in collecting and annotating MS and NMOSD datasets, there is a limited amount of data available, making it difficult to train models. Therefore, this study utilizes data augmentation to expand the datasets and improve the generalization ability of the models. When training the deep learning models, the preprocessing images are flipped randomly (randomly selecting one of the three axes to flip) and rotated randomly with a fixed probability $p = 0.5$ each time (Figure 2).

## 3.2 Joint model

### 3.2.1 Model overall architecture

The joint model comprises three primary components: an information-sharing subnetwork, a lesion segmentation subnetwork, and a disease classification subnetwork. It leverages two types of information interaction methods to enhance the performance of both tasks. The overall architecture of this joint model is illustrated in Figure 3.

Initially, the MRI datasets are fed into the model and processed through the information-sharing subnetwork. Subsequently, the obtained segmentation feature maps $f_{seg}$ and classification feature maps $f_{cls}$ are inputted into the lesion segmentation subnetwork and the disease classification subnetwork, respectively. The outputs of the lesion segmentation subnetwork comprise binarized lesion segmentation maps, while the disease classification subnetwork provides classified predictions for MS or NMOSD.

Moreover, to bolster the interaction between these two tasks during the model training, we have introduced a Lesion Guidance Module (LGM) and a cross-task loss function. These additions aim to mutually guide and enhance the model's performance.

### 3.2.2 Information-sharing subnetwork

MS and NMOSD manifest characteristics of dispersion and multifocality, with lesions varying in shape, size, and discrete distribution. In the tasks involving segmentation and disease classification, it becomes imperative to consider not only the local details such as lesion shape and contour but also their global distribution within the brain. While convolutional operations primarily capture local information, they fail to establish long-distance dependencies across the entire image (He et al., 2016).

To concurrently capture both local and global information, the information-sharing subnetwork in the joint model is structured as a two-branch architecture. The local branch employs convolutional operations to extract detailed information about specific lesions, while the global branch utilizes Swin Transformer (Vaswani et al., 2017; Liu et al., 2021) coding modules to model long-range dependencies among image contexts. This specific architectural design is depicted in Figure 4.

To achieve the integration of local and global information, the segmentation feature maps for input into the segmentation subnetwork and classification feature maps for input into the classification subnetwork are obtained by combining them with cross-elements in a linear weighted manner.

In the global branch, to reduce computational complexity, image blocks of size $16 \times 16 \times 16$ within the input images are treated as computational units for self-attention in the Transformer. A convolutional layer with a kernel size of $8 \times 8 \times 8$ and a stride of 8 is utilized to extract features from the image $I \in \mathbb{R}^{H \times W \times D}$. This extraction process is followed by an average pooling operation, reducing the scale of the extracted features by half. The resulting feature maps $E \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}}$ serve as the input to the Transformer.

The output of the global branch can be computed as follows (Equations 1,2):

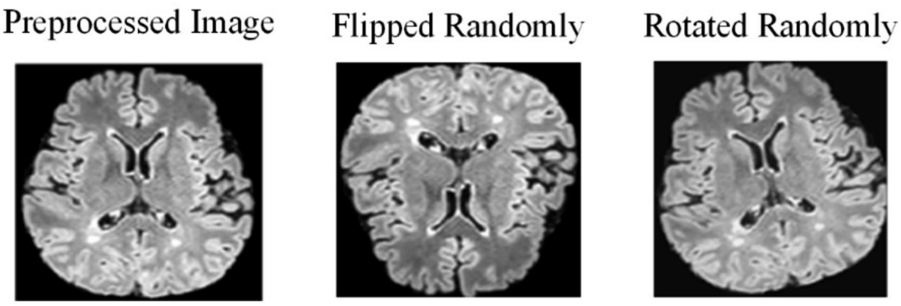


FIGURE 2
Effect of data augmentation.

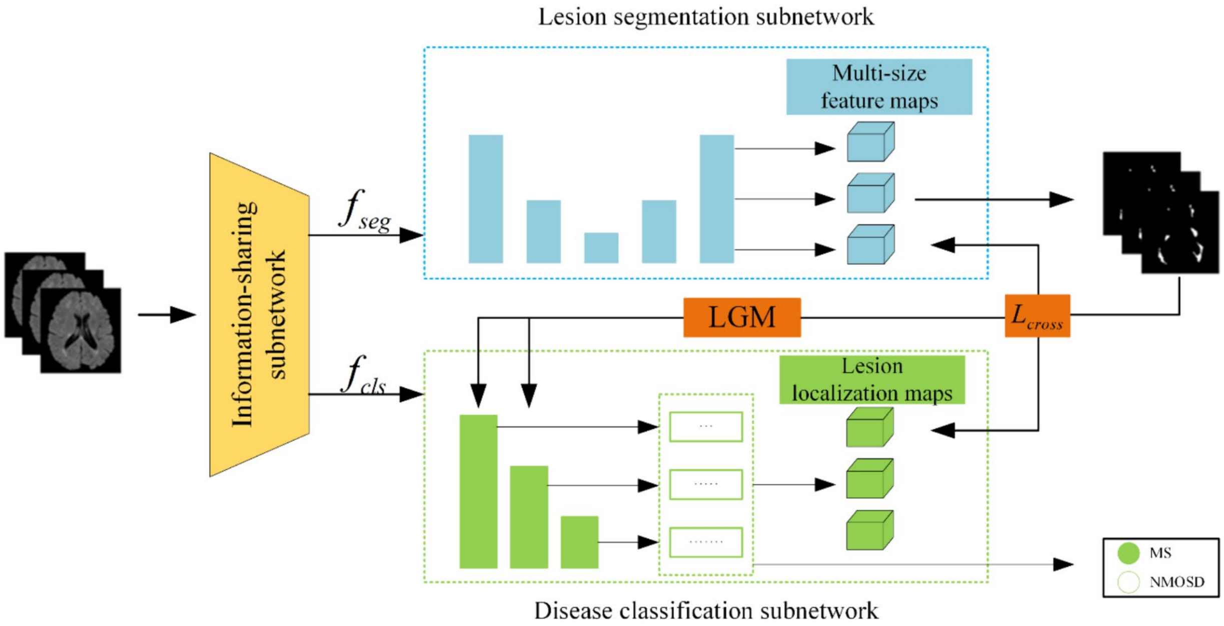

Lesion segmentation subnetwork
Multi-size feature maps
Information-sharing subnetwork
$f_{seg}$
$f_{cls}$
LGM
$L_{cross}$
Lesion localization maps
MS
NMOSD
Disease classification subnetwork

FIGURE 3
Structure of the joint model.

$$GTX_{global} = UP\left(ST\left(LN\left(E\right)\right)\right) \tag{1}$$

$$E = PE\left(I\right) \tag{2}$$

Where $PE(\cdot)$ denotes patch embedding, which involves two operations: a convolution operation with a kernel size of $8\times8\times8$ and a stride of 8, followed by an average pooling operation with a kernel size of $2\times2\times2$ and a stride of 2. $LN(\cdot)$ represents layer normalization. $ST(\cdot)$ refers to two layers of Swin Transformer, with a window size of $5\times5\times5$ and a switch between different window modes across these two layers (shift window mechanism). $UP(\cdot)$ denotes the output of the global branch achieved through interpolation. Through these operations, applied to the original images of size $160\times160\times160$, we capture sufficient global information within this context.

In the local branch, three convolutional layers with $3\times3$ kernels are utilized as context extractors to capture local features. The first convolutional layer has a stride of 2, while the other two layers have a stride of 1. The output of the local branch can be computed as follows (Equation 3):

$$CTX_{local} = Conv\left(LN\left(I\right)\right) \tag{3}$$

Where $CTX_{local}$ denotes the output of the local branch.

To enhance the information interaction between the classification and segmentation tasks, the local feature $CTX_{local}$ and global features $CTX_{global}$ are linearly combined by the crossover unit to obtain the segmentation feature maps $f_{seg}$ for the classification subnetwork and the classification feature maps $f_{cls}$ for the classification subnetwork respectively, which are calculated as Equation 4:

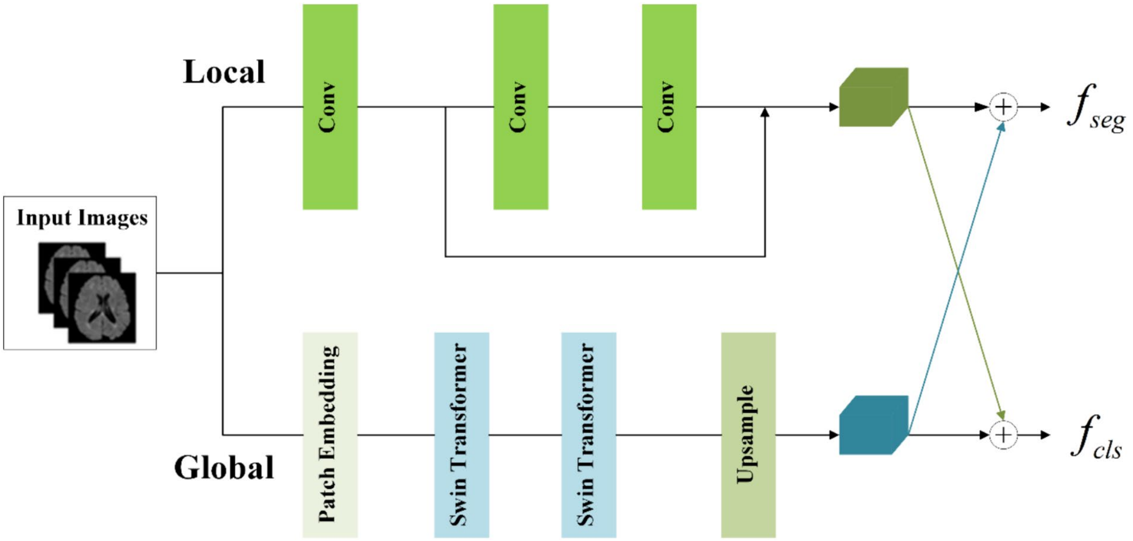


FIGURE 4
Structure of information-sharing subnetwork.

$$\begin{bmatrix} f_{seg} \\ f_{cls} \end{bmatrix} = \begin{bmatrix} w_{seg,1} & w_{cls,2} \\ w_{seg,2} & w_{cls,1} \end{bmatrix} \begin{bmatrix} CTX_{local} \\ CTX_{global} \end{bmatrix} \tag{4}$$

Where $w_{seg,1}$, $w_{seg,2}$, $w_{cls,1}$, and $w_{cls,2}$ are learnable parameters. The size of $f_{seg}$ and $f_{cls}$ is both (32, 80, 80, 80).

### 3.2.3 Lesion segmentation subnetwork

The lesion segmentation subnetwork is utilized to segment brain white matter lesions in MS and NMOSD. Figure 5 shows the schematic structure of the lesion segmentation subnetwork, which mainly consists of three parts: the contraction path, the expansion path, and the multiscale binding module.

In the lesion segmentation subnetwork, the contraction path comprises three encoding modules and two downsampling modules aimed at capturing contextual information within the segmented feature maps $f_{seg} \in \mathbb{R}^{\frac{D}{2} \times \frac{H}{2} \times \frac{W}{2}}$ and extracting lesion-related features. Meanwhile, the expansion path involves three decoding blocks and three upsampling modules for restoring the feature maps to the dimensions of the input images. Between the contraction path and the expansion path, the low-level feature maps obtained from the contraction path and the high-level feature maps obtained from the expansion path are merged in the channel dimension through skip connections. This process aids in integrating detailed image information into the high-level semantic features, thereby enhancing segmentation performance. Additionally, the channel established by the skip connections between high and low levels facilitates gradient backpropagation (Zhang et al., 2018).

Given the substantial disparity in lesion sizes between MS and NMOSD, a strategy is employed to combine richer multi-scale features. Feature maps obtained at multiple levels are weighted and fused to generate the final lesion probability maps $M_{lesion}$ as follows (Equation 5):

$$M_{lession} = sigmoid\left(\sum_{i=1}^{4}\left(\alpha_i \cdot Up\left(f_{1\times1}\left(f_{seg}^{i}\right)\right)\right)\right) \tag{5}$$

Where $f_{1\times1}(\cdot)$ denotes the convolution operation with a kernel size of $1\times1\times1$ acting on the feature maps of size $20\times20\times20$, $40\times40\times40$, $80\times80\times80$, $160\times160\times160$ for adjusting the number of output channels to 1. $Up(\cdot)$ denotes the upsampling operation using nearest neighbor interpolation to resize the multi-scale feature maps to match the original image dimensions. Subsequently, the four-layer feature maps undergo a weighted combination with respective weights of 0.25, 0.25, 0.5, and 1. Finally, the feature maps are mapped to the (0, 1) interval using the sigmoid activation function, resulting in the final lesion probability maps.

### 3.2.4 Disease classification subnetwork

The disease classification subnetwork is dedicated to the classification and diagnosis of two diseases, MS and NMOSD. The outputs $y^{(i)} = \left\{y^{(i)} \in \{0,1\}\right\}_{i=1}^{N_L}$ provide classified predictions for MS or NMOSD. For the ith sample, $y^{(i)} = 0$ denotes classification as MS, while $y^{(i)} = 1$ denotes classification as NMOSD. Considering the relatively limited number of samples, an excessively complex or deeply layered model can lead to overfitting. In this study, the classification model is composed of three 3D coding blocks with residual connections, as depicted in the structure outlined in Figure 6.

The traditional approach in image classification involves unwinding the feature maps generated by the convolutional layer to form feature vectors for inputting into the fully connected layer. However, this method escalates the model's parameter count, raising the risk of overfitting (Szegedy et al., 2015). In this study, we adopt a global average pooling of feature maps as an alternative to feature maps unwinding. There are several reasons for this approach: Firstly, global average pooling integrates spatial information from the feature maps, thereby enhancing the model's generalization while preserving

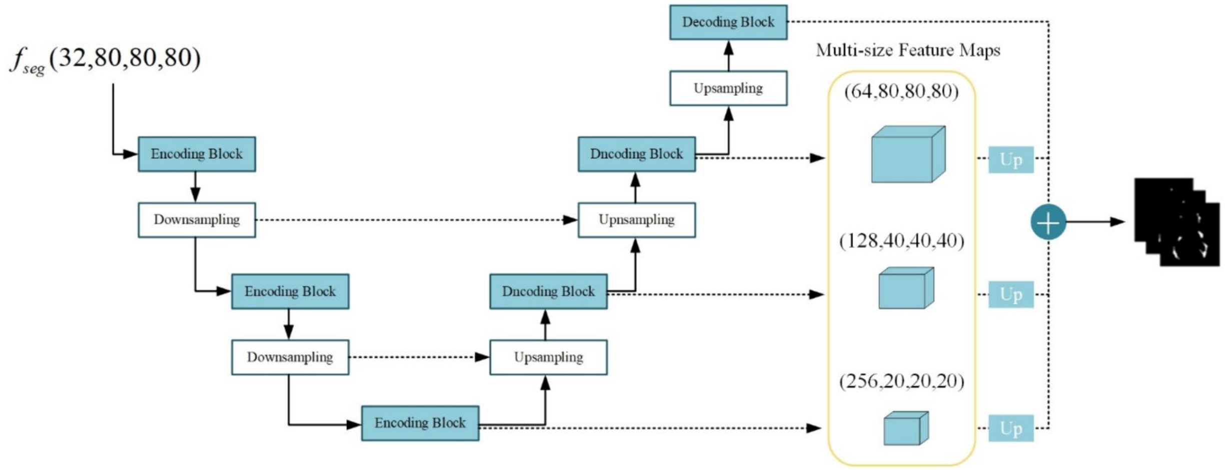


FIGURE 5
Structure of the lesion segmentation subnetwork.

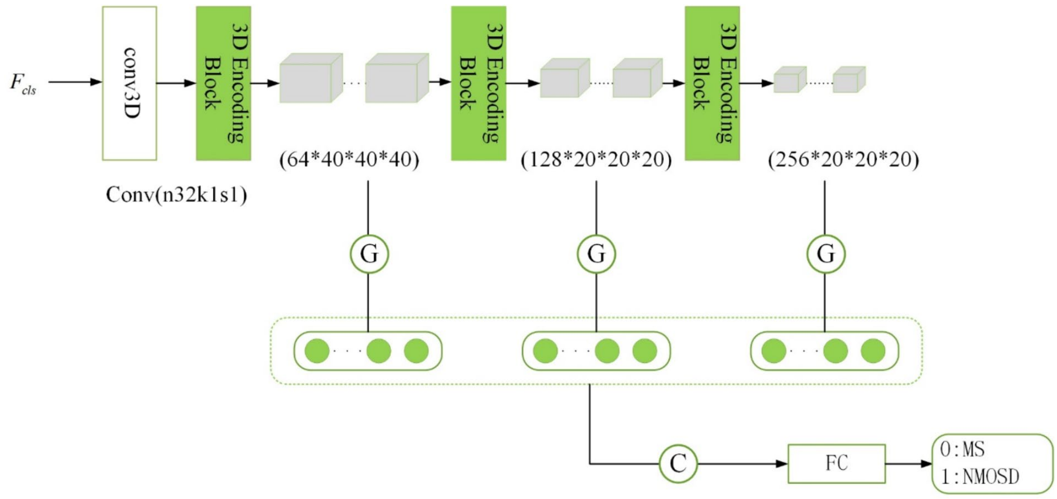


FIGURE 6
Structure of disease classification subnetwork.

classification performance. Secondly, global average pooling requires fewer training parameters, mitigating the risk of overfitting associated with a fully connected layer.

Inspired by the concept of skip connections in the segmentation subnetwork, our model dynamically extracts classification features. The initial layers in the network capture low-level features like texture and color, while deeper layers extract high-level semantic features. To enhance the model's performance, we combine the low-level feature maps with the high-level ones. Consequently, the final category prediction probability is computed as follows (Equation 6):

$$p_{cls} = soft\max\left( f_{FC}\left( \sum_{k=1}^{64} f_{1,k\cdot w_1^{k,n}} + \sum_{k=2}^{128} f_{2,k} \cdot w_2^{k,n} + \sum_{k=3}^{256} f_{3,k\cdot w_3^{k,n}} \right)\right) \quad (6)$$

Where $\{f_{1,k}\}_{k=1}^{64}$, $\{f_{2,k}\}_{k=2}^{128}$, and $\{f_{3,k}\}_{k=3}^{256}$ denotes the feature vectors obtained after global average pooling of the feature maps of the three layers with different scales. These feature vectors were connected as the input of the fully connected layer $f_{FC}(\cdot)$ and then the softmax activation function was applied to obtain the classification prediction probabilities of MS and NMOSD.

## 3.3 Information interaction module

### 3.3.1 Lesion guidance module

Since the lesion regions only occupy a small portion of the brain MRI, the majority of the image consists of normal brain tissues or blood vessels. This can present challenges for the model when attempting to accurately classify based solely on the presence of lesions. To address this, the lesion segmentation feature maps are utilized as prior information for lesion location and morphology by sharing the lesion probability maps obtained from the segmentation subnetwork into the classification subnetwork. By doing so, the influence of other parts of the MRI image on classification is mitigated, thereby facilitating disease classification and diagnosis.

To enhance the effectiveness of the segmented prior information, a Lesion Guidance Module (LGM) is proposed. The structure of the LGM, depicted in Figure 7, consists of two main components: computation of attention maps and fusion of raw features, corresponding to (a) and (b) in Figure 7, respectively.

Firstly, the lesion probability map, which is the output of the last layer of the segmentation subnetwork, undergoes feature transformation and normalization to generate the lesion attention map. Next, the lesion attention map is element-wise multiplied by the classification feature map, emphasizing the presence of lesions. The specific calculation process is as follows:

First, we utilize a $3\times3$ convolution with a stride of 2 to make the lesion segmentation probability map $\{P\}\in\mathbb{R}^{D\times W\times H}$ output from the last layer of the segmentation subnetwork to downsample to the same size as the categorized feature map $\{D,P'\}\in\mathbb{R}^{\frac{D}{2}\times\frac{H}{2}\times\frac{W}{2}}$. Then the features are decoupled by a linear transformation of the Sigmoid function to generate the lesion attention map $\alpha_i$ (Equations 7,8):

$$\alpha_i = Soft\max\left(f_P\left(P_i\right)\right) \tag{7}$$

$$D' = \alpha_i \cdot D \tag{8}$$

Where $\{P\}\in\mathbb{R}^{D\times W\times H}$ denotes the lesion segmentation probability map, $f_P\left(\cdot\right)$ denotes a convolution operation with a kernel size of $3\times3\times3$ and a stride of 2, $\alpha_i\in\mathbb{R}^{\frac{D}{2}\times\frac{H}{2}\times\frac{W}{2}}$ denotes the lesion attention map, $D\in\mathbb{R}^{\frac{D}{2}\times\frac{H}{2}\times\frac{W}{2}}$ denotes classification feature map output from the information-sharing subnetwork, $D'$ denotes segmentation feature map after emphasizing the lesions.

In order to further utilize the prior knowledge of lesion segmentation, the segmentation probability map is directly fused with the classification feature map after emphasizing the lesion in the channel dimension as a form of auxiliary information in the original feature fusion part. The joint features $\{J\}\in\mathbb{R}^{C\times\frac{D}{2}\times\frac{H}{2}\times\frac{W}{2}}$ are obtained by downscaling (Equation 9):

$$J_i = f_{1\times1}\left(concat\left(f_{3\times3}\left(P_i^{'}\right), f_{3\times3}\left(D_i^{'}\right)\right)\right) \tag{9}$$

Where $f_{1\times1}\left(\cdot\right)$ denotes a convolution operation with a kernel size of $1\times1\times1$, $f_{3\times3}\left(\cdot\right)$ denotes a convolution operation with a kernel size of $3\times3\times3$ and a stride of 2, $P_i^{'}$ denotes lesion segmentation probability map after downsampling, and $concat\left(\cdot\right)$ denotes channel dimension splicing.

### 3.3.2 Cross-task loss function

For the classification task of MS and NMOSD, the basis for classification lies in the morphological and positional features of the lesions. Thus, the lesion point of the classification subnetwork during the classification process should be the lesion regions. Based on this analysis, this study utilizes the idea of CAM (Zhou et al., 2016) to achieve lesion localization in MRI images and generate lesion localization maps. To effectively assist the segmentation task, a cross-task loss function is proposed, which supervises the lesion localization map created in the classification subnetwork and the corresponding lesion segmentation maps in the segmentation subnetwork through a loss function. This helps enhance the lesion localization ability in each task. According to the theory of CAM, the lesion localization maps are obtained by global average pooling the feature maps, multiplying the resulting scalar with the class weights corresponding to the output layer, and accumulating it with the feature maps. To incorporate multi-scale features, this study generates lesion localization maps at three different layers. The specific calculation formula is as follows (Equation 10):

$$V_i = \sum_{k=1}^{N} F_{i,k} \cdot w_{i,k}^{c} \tag{10}$$

Where $V_i$ represents the lesion localization map formed in the ith layer (i = 1,2,3), $N$ represents the total number of feature maps in the ith layer, $F_{j,k}$ represents the kth feature map in the ith layer, and $w_{i,k}^{c}$ represents the weight corresponding to class c. Then, the lesion localization maps are weighted and computed with the lesion

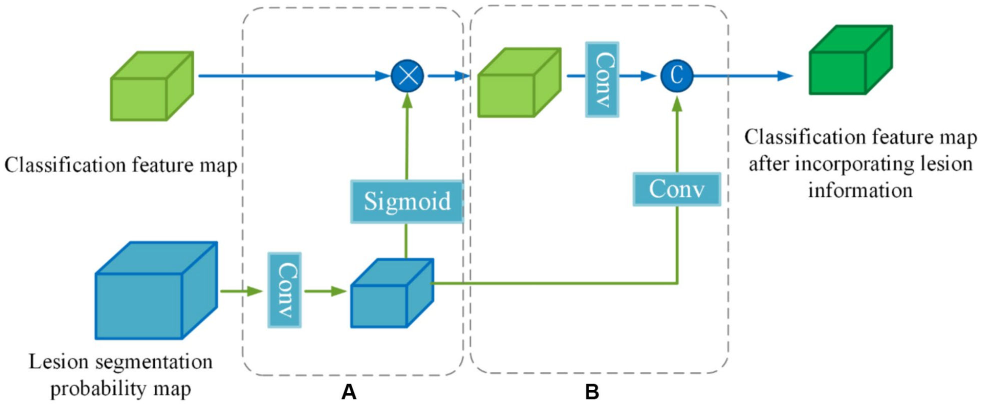

Classification feature map
Sigmoid
Conv
Conv
C
Conv
Classification feature map after incorporating lesion information
Lesion segmentation probability map
A
B

FIGURE 7
Structure of lesion guidance module.

segmentation maps of the same size from the segmentation subnetwork using the Mean Square Error (MSE) loss. This results in a cross-task loss function, with the specific formula as follows (Equation 11):

$$L_{cross} = \sum_{i=1}^{3} \eta_i \cdot \left\| S_i - V_{i2}^{2} \right\| \tag{11}$$

Where $S_i$ represents the intermediate multi-size segmentation maps obtained from the lesion segmentation subnetwork. $\eta_i$ are set to 0.25, 0.25, and 0.5, respectively.

Furthermore, the generation of lesion localization maps achieves the interpretability of deep learning black-box models in the process of classification diagnosis, which is of significant importance for research related to medically auxiliary diagnosis.

## 3.4 Overall loss function

For the disease classification task, the difference between the predicted categories and the true labels was evaluated using the binary cross entropy loss with the formula Equation 12:

$$L_{cls} = -\sum y \log \hat{y} + (1-y)\log(1-\hat{y}) \tag{12}$$

Where $\hat{y}$ represents the categories of model predictions, $y$ represents the real categories.

For the lesion segmentation task, the Dice Similarity Coefficient (DSC) is utilized to measure the degree of similarity between the segmentation results and the real segmentation maps, which can measure the accuracy of the segmentation results and take the value in the range of [0,1]. Therefore, the segmentation model can be supervised by using 1-Dice as a loss function, called Dice loss, the specific formula is Equation 13:

$$L_{seg} = 1 - Dice = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{13}$$

Where $X$ represents the segmentation result of the lesion segmentation task, $Y$ represents the real segmentation result, and $|\cdot|$ represents the number of voxels that satisfy the condition.

Therefore, to optimize the learnable parameters $w_t = (\theta_t)$ of the joint model, where $\theta_t$ is the model parameter, an overall loss function for the joint model is designed in conjunction with the single-task loss described above. Since single tasks may have different levels of contribution in optimizing the parameters of the model, the single task loss function is weighted by setting a weighting factor $\beta_t$, and the joint loss function constituted is Equation 14:

$$L = \beta_1 L_{seg} + \beta_2 L_{cls} + \beta_3 L_{cross} \tag{14}$$

The variable $\beta_i, i = 1,2,3$ represents a hyperparameter. Considering that the segmentation task involves pixel-level classification of images, while the classification task involves categorizing individual samples, segmentation tasks are comparatively more complex and challenging to learn. Therefore, in the training process, the contribution of parameter optimization for the segmentation task should be relatively higher. Hence, in this study, $\beta_1$, $\beta_2$, and $\beta_3$ are set to 1, 0.8, and 1, respectively.

# 4 Experiment and results

## 4.1 Experiment settings

The hardware platform for the experiments in this study is the NAVIDA GTX 3090 graphics card and the network models are all built by the PyTorch framework. During model training, the batch size was set to 2, the number of iterations was set to 200, the optimization algorithm utilized Adam optimizer with default parameters (Kingma and Ba, 2014), and the learning rate was initially set to 0.0001. If the loss of the model did not decrease after surpassing 10 training iterations, the learning rate was then reduced to half of its original value.

To reduce experimental variability and provide a more accurate and objective reflection of model performance, we utilized a five-fold cross-validation strategy during the experiment. Initially, the entire dataset is randomly divided into five subsets. During each training iteration, four of these subsets are used as the training set, while the remaining subset serves as the test set. Upon completion of each training iteration, evaluation results are obtained on the corresponding test set. The final experimental outcome is determined by averaging the evaluation results obtained from the five training iterations.

The images used in the experiments are obtained from the dataset after undergoing the data preprocessing and data augmentation described in Section 3.1. The original size of each input is (160, 160, 160). After being processed by the information-sharing subnetwork, the inputs for the lesion segmentation subnetwork and disease classification subnetwork are resized to (80, 80, 80).

## 4.2 Evaluation metrics

The joint model proposed in this study mainly consists of two tasks, the lesion segmentation task and the disease classification task. Multiple evaluation metrics were employed to assess the performance of each model.

The evaluation metrics utilized for the segmentation task are Dice Similarity Coefficient (DSC), Positive Predict Value (PPV), True Positive Rate (TPR), and Volume Difference (VD) (Equations 15–18).

$$DSC = \frac{2 \times \left| Y \cap \hat{Y} \right|}{|Y| + \left| \hat{Y} \right|} \tag{15}$$

Where $Y$ represents the ground truth, $\hat{Y}$ represents the output of the model, $\cap$ represents the intersection operation of two matrices, and $|.|$ represents the number of elements in the matrix. The higher the DSC, the closer the prediction to the manually segmented label.

$$PPV = \frac{\left|Y \cap \hat{Y}\right|}{\left|\hat{Y}\right|} \tag{16}$$

PPV represents the proportion of true positive voxels among all voxels predicted as positive. In this context, positive refers to the lesion voxels. A higher PPV indicates that the impact of the noise caused by them on the model is smaller.

$$TPR = \frac{\left|Y \cap \hat{Y}\right|}{\left|Y\right|} \tag{17}$$

TPR represents the proportion of true positive voxels among all actual positive voxels. The higher the TPR, the stronger the model's ability to identify lesions.

$$VD = \frac{TP_P - TP_{gt}}{TP_{gt}} \tag{18}$$

Where $TP_p$ represents the number of predicted TP voxels and $TP_{gt}$ represents the number of lesion voxels in the ground truth. A lower VD indicates a better agreement between the predicted and true lesion volumes.

The metrics utilized in the classification task are Accuracy (ACC), Sensitivity (SN), Specificity (SP) (Equations 19–21), and Area Under the ROC Curve (AUC).

$$ACC = \frac{TP + TN}{TP + FP + FN + FN} \tag{19}$$

$$SN = \frac{TP}{TP + FN} \tag{20}$$

$$SP = \frac{TN}{TN + FP} \tag{21}$$

Where TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative. These metrics can be used to assess the performance of classification models.

## 4.3 Comparison and analysis of experimental results

To validate the superiority of the proposed joint model for lesion segmentation and disease classification in MS and NMOSD, this section conducts comparative analyses between the joint model and three advanced segmentation methods: 3D UNet, VNet, and AttentionUNet, as well as three advanced classification methods: 3D ResNet34, 3D ResNet50, and 3D DenseNet. Throughout the experiments, efforts are made to ensure that the primary parameters of all methods remain consistent with those introduced in Section 4.1.

### 4.3.1 Comparison of the segmentation methods

Table 3 shows the comparison of the results of the joint model on lesion segmentation. It can be seen that the joint model performs the best on the lesion segmentation task and achieves the highest DSC, TPR, and VD, which are 74.87, 72.21, and 22.34%, respectively, and PPV achieves the sub-optimal results, which is second only to the 3D UNet. Where AttentionUNet achieves sub-optimal results on DSC, TPR, and VD which differed from the joint model by 2.8, 0.89, and 4.33%, respectively. The results prove that the lesion segmentation results obtained by the joint model have higher similarity with the real lesion segmentation results and higher check-accuracy for lesion pixel points. The result of PPV is lower than that of 3D UNet by 0.17%, proving that checking accuracy for lesion pixel points of the segmentation model is slightly lower than 3D UNet. Overall, the joint model has the optimal segmentation effect, and AttentionUNet is the second best.

Figure 8 shows the lesion segmentation visualization results of the four segmentation models acting on a MS case and a NMOSD case, respectively. In order to clearly demonstrate the segmentation effect, the lesion segmentation results are superimposed on the original MRI image in red. From left to right, it shows the ground truth segmentation annotated manually by the doctors, the joint model segmentation results, the UNet segmentation results, the VNet segmentation results, and the AttentionUNet segmentation results. As can be seen from the visualization results, the models can achieve localization for lesion regions with relatively large volumes. However, the discrete point lesions in the brains of MS and NMOSD patients are difficult to recognize, as well as some regions in the brain MRI with similar imaging features to the white matter high signals are very easy to confuse during segmentation. The blue boxed part in the figure shows the under-segmentation or over-segmentation problem of the model during segmentation. In contrast, the joint model greatly avoids the above problems due to the combination of the lesion location information provided by the classification network, and the segmentation results are closer to the ground truth segmentation.

Overall, based on evaluations of various metrics and visualized results, it can be observed that the segmentation performance of the joint model is superior. Particularly for some challenging pinpoint lesions and locations with similar imaging characteristics, the segmentation results achieved by the joint model is more refined. Given the highly irregular morphology of MS and NMOSD lesions, even expert radiologists find it challenging to completely delineate their contours during annotation. However, the model proposed in our study ensures the ability to accurately locate the majority of lesions.

TABLE 3 Comparative experimental results of the lesion segmentation method.

| Method | DSC (%) | PPV (%) | TPR (%) | VD (%) |
|---|---|---|---|---|
| 3D UNet | 70.64 | **74.11** | 65.25 | 36.38 |
| VNet | 71.93 | 72.24 | 69.71 | 43.43 |
| AttentionUNet | 72.07 | 70.22 | 71.32 | 26.67 |
| Joint Model | **74.87** | 73.94 | **72.21** | **22.34** |

Bold values indicate the best results in evaluation metrics. Underlined values indicate the second best results in evaluation metrics.

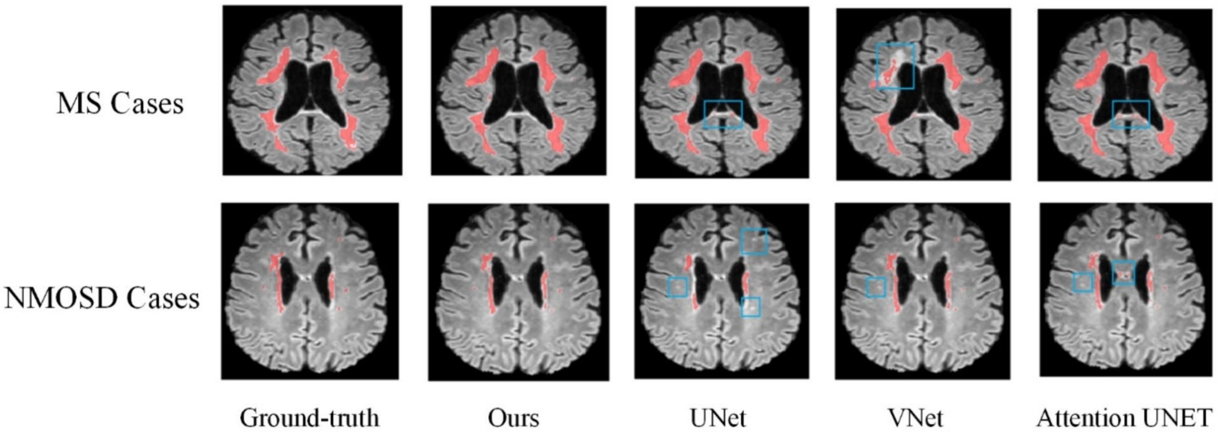


FIGURE 8
Comparison of experimental visualization results.

**TABLE 4 Comparative experimental results of lesion classification method.**

| Method | ACC (%) | SN (%) | SP (%) | AUC (%) |
|---|---|---|---|---|
| ResNet50 | 87.24 | 87.14 | 88.33 | 94.84 |
| ResNet101 | 86.63 | 80.28 | 92.38 | 87.66 |
| ResNet152 | 86.40 | 85.52 | 85.47 | 85.43 |
| DenseNet121 | 88.36 | 90.85 | 83.57 | 92.47 |
| SENet50 | 89.80 | 86.00 | 93.57 | 93.33 |
| Joint Model | **92.16** | **95.60** | **92.60** | **96.33** |

Bold values indicate the best results in evaluation metrics.

### 4.3.2 Comparison of the classification methods

Table 4 demonstrates the disease classification results of the joint model and the other five advanced classification models for MS and NMOSD. From the results in the table, it can be seen that the joint model achieved a classification accuracy of 92.16%, which is the best performance among several methods and 2.36% higher than SENet50 with the second-best accuracy. In addition, the ROC curves for the classification results of the five methods are shown in Figure 9. Compared with the other models, the joint model has the largest area under the line of the ROC curve, which is 96.33%, indicating that the joint model has a better effect.

## 4.4 Results and analysis of ablation experiments

The joint model proposed in this study focuses on the mutual guidance of the lesion segmentation and disease classification tasks through an information-sharing subnetwork and two information interaction methods which include the lesion guidance module and the cross-task loss function. In this section, we will first validate the effectiveness of the joint learning strategy, and then conduct ablation experiments on the information sharing subnetwork and the two information interaction methods, respectively, to verify the influence of each component in improving the performance of the two tasks.

### 4.4.1 Influence of the joint learning strategy

To demonstrate the impact of joint learning strategy on model performance, this section will compare the results of the joint model and two fundamental models: the lesion segmentation model (corresponding to the lesion segmentation subnetwork in Section 3.2.3) and the disease classification model (corresponding to the disease classification subnetwork in section 3.2.4). This comparison aims to establish the effectiveness of the joint learning strategy. The joint model will be abbreviated as the Joint Model, the lesion segmentation model will be abbreviated as Only Seg Model and the disease classification model will be abbreviated as Only Cls Model.

(1) Influence on the lesion segmentation task

Table 5 compares the performance of the Only Seg Model and Joint Model in lesion segmentation tasks. The Joint Model demonstrated improvement across all metrics compared to the Only Seg Model, with an increase of 3.63% in DSC, 3.80% in PPV, 2.34% in TPR, and 2.63% in VD. This demonstrates that incorporating information from the classification task effectively enhances the performance of lesion segmentation.

Figure 10 displays the lesion segmentation results using the Only Seg Model and Joint Model for three different cases. From left to right: original T2-FLAIR images, manually segmented images, visualizations of results from the Only Seg Model, and visualizations of results from the Joint Model. In the visualizations, red indicates true positives—pixels classified correctly as lesions; green represents false positives—pixels classified as lesions but normal tissues; yellow denotes false negatives—pixels classified as normal tissues but lesions. In the segmentation results of Case One, the Only Seg Model misclassifies some normal tissue as lesions (shown in green) due to its similarity to high-intensity white matter, leading to misjudgments. However, the Joint Model correctly identifies this portion. For Case Two, the Only Seg Model struggles to delineate the contours of patchy lesions (depicted in yellow), indicating segmentation inadequacies. Similar issues of segmentation insufficiency are observed for Case Three with the Only Seg Model (yellow portion). Because lesions in MS and NMOSD often exhibit highly irregular shapes, segmenting lesion edges presents a challenge. The visual results show that both the Only Seg Model and Joint Model have some false positives and negatives

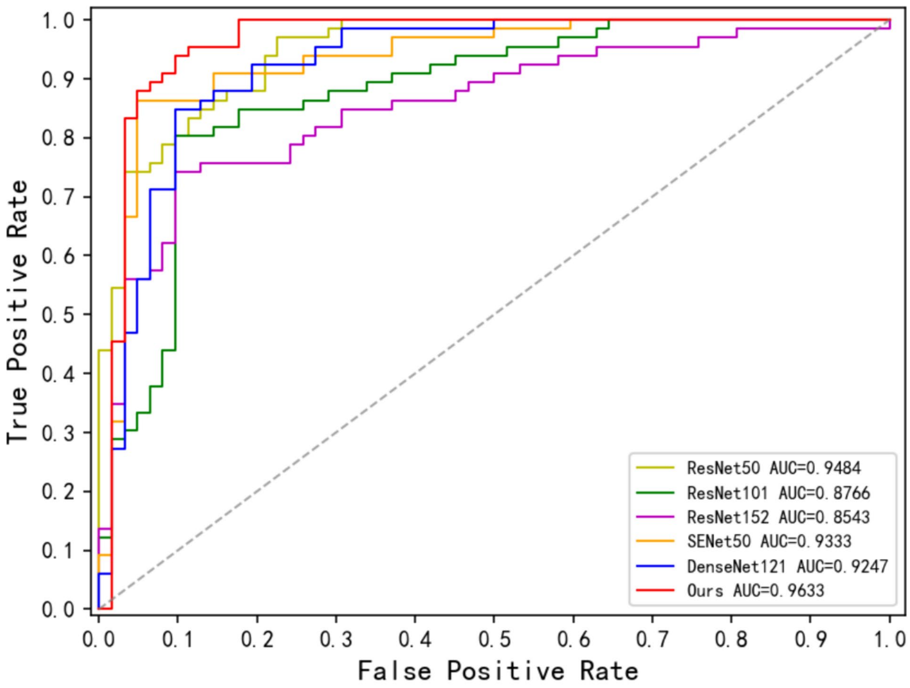


FIGURE 9
ROC curve of the classification methods.

TABLE 5 The influence of the joint learning strategy on segmentation task.

| Method | DSC (%) | PPV (%) | TPR (%) | VD (%) |
|---|---|---|---|---|
| Only Seg Model | 71.24 | 70.14 | 69.87 | 24.97 |
| Joint Model | **74.87** | **73.94** | **72.21** | **22.34** |

Bold values indicate the best results in evaluation metrics.

along the lesion edges. Nevertheless, overall, the Joint Model demonstrates significantly better segmentation performance compared to the Only Seg Model.

The visual results demonstrate that the segmentation performance of the Joint Model surpasses that of the Only Seg Model across all four samples. This outcome suggests that the lesion features extracted from the classification task, especially positional characteristics, effectively assist the segmentation task in capturing the varied locations and sizes of MS and NMOSD lesions.

(2) Influence on the Disease Classification task

Table 6 is the comparison between the Only Cls Model and the Joint model on the results of the MS and NMOSD classification task. Compared to the Only Cls Model, the joint model demonstrated improvement across all metrics, with ACC increasing by 4.36%, SN by 8.36%, and SP by 3.14%.

Figure 11 displays the ROC curves for the Only Cls Model and Joint Model in the classification task. The AUC for the Joint Model is 96.33%, while the AUC for the Only Cls Model is 87.59%. When combined with the table, this demonstrates that leveraging the information extracted from the segmentation task and utilizing the segmented lesion results effectively guides the classification task, thereby enhancing the overall performance of the classification task.

Based on the comprehensive analysis, the Joint Model based on the joint learning strategy demonstrated performance improvements in both lesion segmentation and disease classification tasks. It exhibited superior results across various metrics and visualization outputs compared to single-task models. This validates that the joint learning strategy effectively leverages the features, harnesses hidden information learned from shared classification and segmentation tasks, and enhances the model's fitting capability. Consequently, it elevates the performance in both tasks.

### 4.4.2 Influence of the information sharing module

To investigate the effectiveness of the dual-branch structure within the information-sharing subnetwork, the following experiments were conducted in this section: Removing the information-sharing subnetwork, denoted as 'w/o share'; Using only the local branch to share underlying information through hard parameter sharing, denoted as 'only local'; Using only the global branch to share underlying information through hard parameter sharing, denoted as 'only global.'

Table 7 presents the ablation experiment results for the information-sharing subnetwork. The experimental findings indicate that the information-sharing subnet significantly enhances the performance of both lesion segmentation and disease classification tasks. Performance is notably poorest when the information-sharing subnet is entirely removed, while the Joint Model demonstrates the optimal performance. When utilizing only the local branch for hard parameter sharing of the underlying information, the performance in the lesion segmentation task ranks second, following closely behind the Joint Model. This demonstrates that local lesion information, such as morphology and edges, holds greater significance for the lesion segmentation task. When employing local sharing, the model

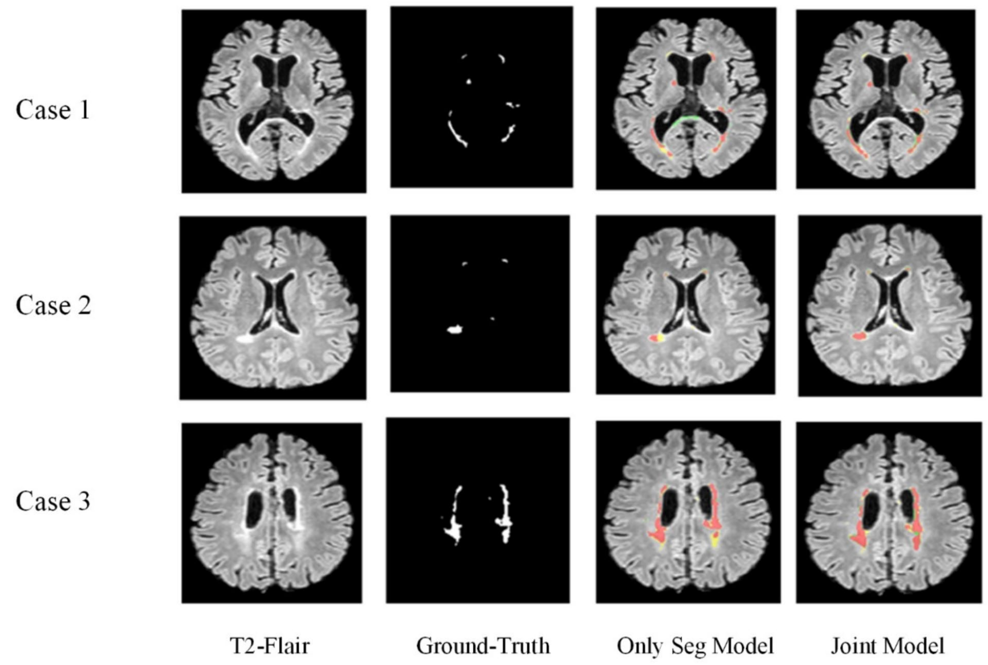


FIGURE 10
Visual display of lesion segmentation results.

TABLE 6 The influence of the joint learning strategy on classification task.

| Method | ACC (%) | SN (%) | SP (%) | AUC (%) |
|---|---|---|---|---|
| Only Cls Model | 87.80 | 87.24 | 89.46 | 87.59 |
| Joint Model | **92.16** | **95.60** | **92.60** | **96.33** |

Bold values indicate the best results in evaluation metrics.

prioritizes aspects related to the segmentation task. In the case of using only the global branch for hard parameter sharing of the underlying information, the performance in the disease classification task ranks second, closely following the Joint Model. This highlights that global image information, such as lesion distribution, plays a more advantageous role in the disease classification task.

### 4.4.3 Influence of the lesion guidance module

As mentioned in section 3.3, the information interaction consists of two parts, one is the LGM for combining the results of lesion segmentation, and the other is the cross-task loss function. In this section, the effectiveness of the lesion guidance module is verified through ablation experiments.

The LGM serves to utilize the segmentation probability maps as prior information about lesion distribution and morphology to guide the classification subnet. Its structure involves emphasizing lesions by first applying an attention mechanism through the dot product operation between the segmentation probability maps and the classification feature maps. Subsequently, the segmentation probability maps are concatenated with the classification feature maps along the channel dimension to further integrate lesion information. In this section, we investigate the effectiveness of LGM for the classification task, as well as the efficacy of the LGM structure. We conduct ablation experiments as follows: (1) Removing the LGM, denoted as 'w/o LGM'. (2) Using only the dot product operation to combine the segmentation probability maps and the classification feature maps, denoted as 'dot product'. (3) Using only the concatenation along the channel dimension to combine the segmentation probability maps and the classification feature maps, denoted as 'concat.' (4) Given the common approach of fusing information by pixel-wise addition, such as in, we compare using the addition operation to combine the segmentation probability maps and the classification feature maps, denoted as 'dot add.' Table 8 presents the performance of different forms of LGM on the classification task.

Table 8 highlights that removing the LGM notably worsens classification performance, underlining the importance of merging lesion segmentation into disease classification. Using attention on lesion segmentation probability maps, like with dot product, notably boosts classification, although it's not the best method.

Combining the segmentation maps with classification features through channel dimension concatenation and dot product both improve the classification. However, channel concatenation works better, likely because the segmentation maps may be inaccurate, causing problems with direct addition or summation.

The proposed LGM, using both dot product and channel concatenation, significantly enhanced classification performance, offered the most improvement.

### 4.4.4 Influence of cross-task loss function

We set up ablation experiments for the cross-task loss function of the information interaction approach in this section. According to Equation (14), the overall loss function of the model consists of three parts: segmentation loss function, categorization loss function, and cross-task loss function, which is verified in this section by removing the cross-task loss function, called w/o cross loss.

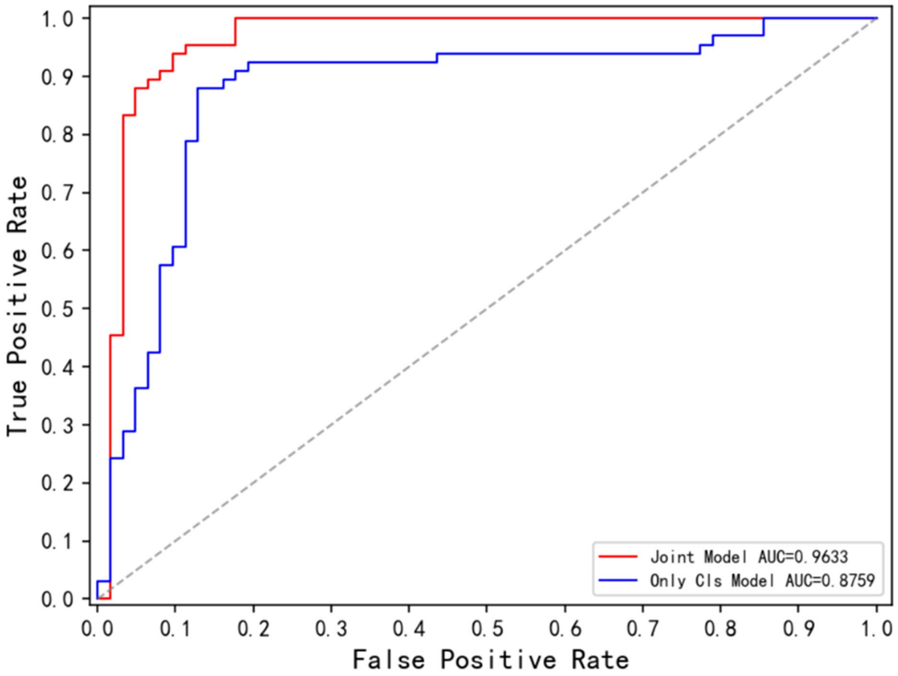


FIGURE 11
ROC curve corresponding to Joint Model and Only Cls Model.

TABLE 7 The influence of the information-sharing module.

| Method | DSC (%) | PPV (%) | TPR (%) | VD (%) | ACC (%) | SN (%) | SP (%) | AUC (%) |
|---|---|---|---|---|---|---|---|---|
| w/o share | 71.98 | 72.11 | 67.47 | 26.35 | 88.20 | 86.24 | 89.24 | 88.56 |
| Only local | 73.80 | 73.95 | 71.34 | 24.12 | 86.14 | 86.00 | 88.26 | 89.32 |
| Only global | 72.04 | 70.18 | 70.44 | 24.92 | 90.22 | 91.46 | 89.60 | 92.68 |
| Joint Model | **74.87** | 73.94 | **72.21** | **22.34** | 92.16 | **95.60** | **92.60** | **96.33** |

Bold values indicate the best results in evaluation metrics. Underlined values indicate the second best results in evaluation metrics.

TABLE 8 The effect of the LGM on classification performance.

| Interactive mode | ACC (%) | SN (%) | SP (%) | AUC (%) |
|---|---|---|---|---|
| w/o LGM | 89.44 | 89.30 | 86.24 | 89.38 |
| dot product | 91.20 | 93.42 | 89.28 | 94.36 |
| concat | 90.66 | 93.34 | 91.60 | 92.24 |
| dot add | 89.48 | 89.36 | 85.24 | 89.90 |
| Joint Model | **92.16** | **95.60** | **92.60** | **96.33** |

Bold values indicate the best results in evaluation metrics. Underlined values indicate the second best results in evaluation metrics.

Table 9 demonstrates the impact of cross-task loss functions on lesion segmentation and disease classification tasks. The results indicate that cross-task loss functions effectively enhance the performance of both tasks, particularly in the case of lesion segmentation. The average metrics show an improvement of 1.23%, validating that the lesion localization maps constructed by the classification subnetwork effectively aid the segmentation subnetwork in localizing lesions, thereby enhancing segmentation performance. For the disease classification task, there are improvements across metrics such as ACC, SP, and AUC. This demonstrates that the feature maps obtained by the segmentation subnetwork effectively guide the classification subnet in capturing lesions.

In addition, Figure 12 illustrates the comparison between lesion localization maps generated by the joint model and the annotated lesion gold standard for four test samples. The lesion localization maps visualize the model's focus areas during classification using Grad-CAM (Wagner et al., 2019) technology, where deeper colors indicate higher model attention. Comparing the visualized results with manually annotated segmentation gold standards reveals a substantial alignment between the areas the model emphasizes during classification and the actual lesion locations. Particularly noteworthy is the accurate localization of minute dot-like lesions present in Sample 3, which represent lesions challenging for the segmentation model to distinguish. However, the lesion localization maps manage to accurately pinpoint these lesions. This further validates the reliability of guiding the segmentation model through cross-task losses.

Simultaneously, these lesion localization maps offer interpretability for the joint model in diagnosing MS and NMOSD. They can serve as a basis for deriving diagnostic conclusions for MS and NMOSD in clinical practice.

## 5 Conclusion

This study proposes a joint model for lesion segmentation and disease classification of MS and NMOSD. Leveraging the correlation

TABLE 9 The effect of cross-task loss function.

| Method | DSC (%) | PPV (%) | TPR (%) | VD (%) | ACC (%) | SN (%) | SP (%) | AUC (%) |
|---|---|---|---|---|---|---|---|---|
| w/o cross loss | 73.96 | 72.06 | 70.64 | 22.90 | 92.08 | **95.62** | 92.54 | 95.54 |
| Joint Model | **74.87** | **73.94** | **72.21** | **22.34** | **92.16** | <u>95.60</u> | **92.60** | **96.33** |

Bold values indicate the best results in evaluation metrics. Underlined values indicate the second best results in evaluation metrics.

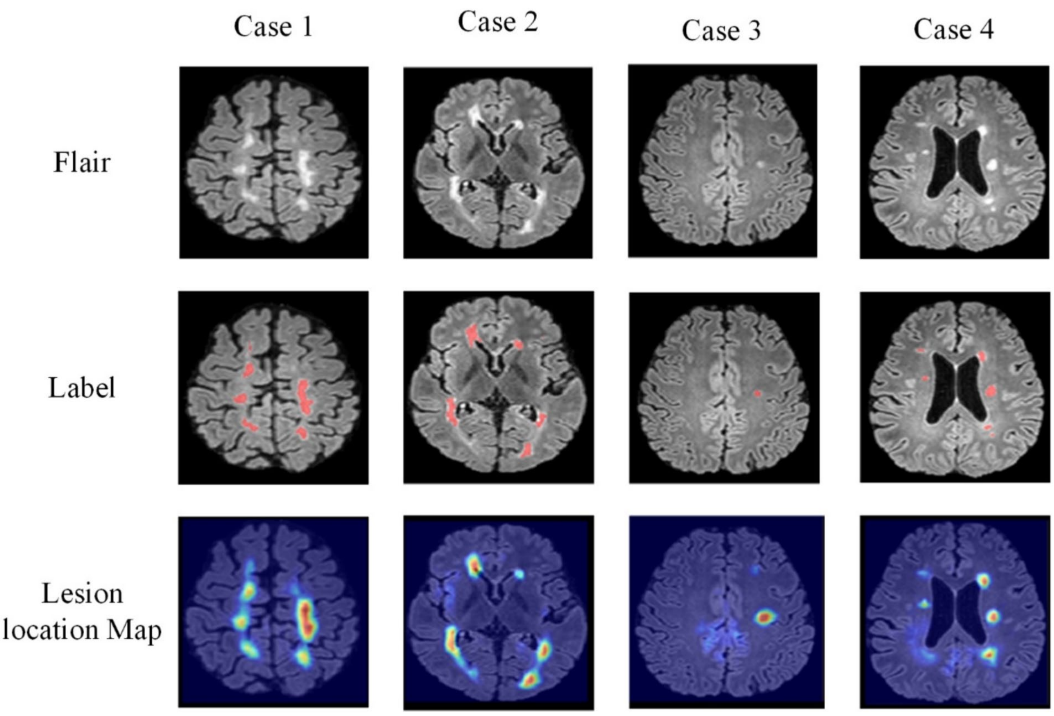


FIGURE 12
Visual display of lesion localization maps.

between lesion segmentation and disease classification tasks, the model facilitates mutual guidance between the two tasks through information sharing and interaction. This approach allows for the effective utilization of the information from limited datasets. Furthermore, comparative experiments confirm the joint model's ability to significantly enhance the performance of both tasks. Ablation experiments validate the effectiveness of information sharing and interaction mechanisms within the joint model. While the joint model exhibits strong performance in lesion segmentation and disease classification tasks for MS and NMOSD, its generalization capability to other diseases remains limited and somewhat unstable. Therefore, we plan to gather more extensive datasets to enhance the model's generalizability. Additionally, utilizing multimodal data as input to the model aims to augment its practical applicability.

## Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the ISBI dataset used in this study is publicly available: https://smart-stats-tools.org/lesion-challenge-2015. However, the dataset from the First Hospital of Jilin University cannot be made public due to ethical considerations. For access to this dataset, please contact the corresponding author. Requests to access these datasets should be directed to guocj@jlu.edu.cn.

## Author contributions

LH: Supervision, Writing – review & editing. YS: Methodology, Writing – original draft. HY: Supervision, Writing – review & editing. CG: Data curation, Writing – review & editing. YW: Supervision, Writing – review & editing. ZZ: Data curation, Writing – review & editing. YG: Methodology, Writing – original draft.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (No. 62072212), the Development Project of Jilin Province of China (Nos. 20220508125RC and 20230201065GX), National

Key R&D Program (No. 2018YFC2001302), the Jilin Provincial Key Laboratory of Big Data Intelligent Cognition (No. 20210504003GH), the Natural Science Foundation of Jilin Province (No. 20210101273JC), and the Science and Technology Achievement Transformation Fund of the First Hospital of Jilin University (No. JDYY2021-A0010).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M. A., et al. (2019). Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage* 196, 1–15. doi: 10.1016/j.neuroimage.2019.03.068

Bauer, S., Wiest, R., Nolte, L. P., and Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Phys. Med. Biol.* 58, R97–R129. doi: 10.1088/0031-9155/58/13/R97

Bruscolini, A., Sacchetti, M., La Cava, M., Gharbiya, M., Ralli, M., Lambiase, A., et al. (2018). Diagnosis and management of neuromyelitis optica spectrum disorders - an update. *Autoimmun. Rev.* 17, 195–200. doi: 10.1016/j.autrev.2018.01.001

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation", in *MICCAI 2016. Lecture Notes in Computer Science()*, (eds.) S., Ourselin, L., Joskowicz, M., Sabuncu, G., Unal, and W. Wells. Berlin, Germany: Springer. 9901, 424–432.

Filippi, M., Rocca, M. A., Ciccarelli, O., De Stefano, N., Evangelou, N., Kappos, L., et al. (2016). MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *Lancet Neurol.* 15, 292–303. doi: 10.1016/S1474-4422(15)00393-2

Gessert, N., Krüger, J., Opfer, R., Ostwaldt, A. C., Manogaran, P., Kitzler, H. H., et al. (2020). Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs. *Comput. Med. Imaging Graph.* 84:101772. doi: 10.1016/j.compmedimag.2020.101772

Griggs, R. C., Mao, Z. F., Qiu, W., Hu, X. Q., Wingerchuk, D. M., and Weinshenker, B. G. (2016). International consensus diagnostic criteria for neuromyelitis optica spectrum disorders. *Neurology* 86:491. doi: 10.1212/wnl.0000000000002366

Hagiwara, A., Otsuka, Y., Andica, C., Kato, S., Yokoyama, K., Hori, M., et al. (2021). Differentiation between multiple sclerosis and neuromyelitis optica spectrum disorders by multiparametric quantitative MRI using convolutional neural network. *J. Clin. Neurosci.* 87, 55–58. doi: 10.1016/j.jocn.2021.02.018

He, K. M., Zhang, X. Y., Ren, S. Q., and Sun, J.IEEE (2016). “Deep residual learning for image recognition” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016*, Washington, DC, USA: IEEE Computer Society. 770–778.

Hu, C., Kang, G. X., Hou, B. B., Ma, Y. Y., Labeau, F., and Su, Z. C.IEEE (2020) "ACU-net: a 3D attention context u-net for multiple sclerosis lesion segmentation", in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, Washington, DC, USA: IEEE Computer Society. 1384–1388.

Huang, D. (2019). Quantitative radiomic biomarkers for discrimination between neuromyelitis optica spectrum disorder and multiple sclerosis. *Mult. Scler. J.* 25:923. doi: 10.1002/jmri.26287

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q.IEEE (2017). "Densely connected convolutional networks", in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, US, 2017, Washington, DC, USA: IEEE Computer Society. 2261–2269.

Kim, H., Lee, Y., Kim, Y. H., Lim, Y. M., Lee, J. S., Woo, J., et al. (2020). Deep learning-based method to differentiate neuromyelitis optica spectrum disorder from multiple sclerosis. *Front. Neurol.* 11:599042. doi: 10.3389/fneur.2020.599042

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]*.

Kister, I., Ge, Y. L., Herbert, J., Sinnecker, T., Wuerfel, J., and Paul, F. (2013). Distinction of seropositive nmo spectrum disorder and MS brain lesion distribution. *Neurology* 81:1966. doi: 10.1212/01.wnl.0000436079.95856.1f

La Rosa, F., Abdulkadir, A., Fartaria, M. J., Rahmanzadeh, R., Lu, P. J., Galbusera, R., et al. (2020). Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE. *Neuroimage Clin.* 27:102335. doi: 10.1016/j.nicl.2020.102335

Lee, J. G., Jun, S., Cho, Y. W., Lee, H., Kim, G. B., Seo, J. B., et al. (2017). Deep learning in medical imaging: general overview. *Korean J. Radiol.* 18, 570–584. doi: 10.3348/kjr.2017.18.4.570

Liu, Y. O., Dong, D., Zhang, L. W., Zang, Y. L., Duan, Y. Y., Qiu, X. L., et al. (2019). Radiomics in multiple sclerosis and neuromyelitis optica spectrum disorder. *Eur. Radiol.* 29, 4670–4677. doi: 10.1007/s00330-019-06026-w

Liu, Z., Lin, Y. T., Cao, Y., Hu, H., Wei, Y. X., Zhang, Z., et al. (2021). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 9992–10002.

Lundervold, A. S., and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift Fur Medizinische Physik* 29, 102–127. doi: 10.1016/j.zemedi.2018.11.002

Mcginley, M. P., Goldschmidt, C. H., and Rae-Grant, A. D. (2021). Diagnosis and treatment of multiple sclerosis a review. *JAMA* 325, 765–779. doi: 10.1001/jama.2020.26858

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation", in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science)*, (eds.) N., Navab, J., Hornegger, W., Wells, and A., Frangi. Berlin, Germany: Springer. vol 9351, 234–241.

Rotstein, D., and Montalban, X. (2019). Reaching an evidence-based prognosis for personalized treatment of multiple sclerosis. *Nat. Rev. Neurol.* 15, 287–300. doi: 10.1038/s41582-019-0170-8

Shelhamer, E., Long, J., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 640–651. doi: 10.1109/TPAMI.2016.2572683

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062

Szegedy, C., Liu, W., Jia, Y. Q., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015.*Washington, DC, USA: IEEE Computer Society. 1–9.

Tseng, K. L., Lin, Y. L., Hsu, W., and Huang, C. Y.IEEE (2017). "Joint sequence learning and cross-modality convolution for 3D biomedical segmentation", in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, Washington, DC, America: IEEE Computer Society. 3739–3746.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y. J., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: Improved N3 Bias Correction. *Ieee Transactions on Medical Imaging*, 29, 1310–1320.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). “Attention is all you need” in *Advances in neural information processing systems, 30*. Cambridge, Massachusetts, USA: MIT Press.

Wagner, J., Köhler, J. M., Gindele, T., Hetzel, L., Wiedemer, J. T., Behnke, S., et al. (2019). "Interpretable and fine-grained visual explanations for convolutional neural networks", in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019*, Washington, DC, USA: IEEE Computer Society. 9089–9099.

Wang, L., Lai, H. M., Barker, G. J., Miller, D. H., and Tofts, P. S. (1998). Correction for variations in MRI scanner sensitivity in brain studies with histogram matching. *Magn. Reson. Med.* 39, 322–327. doi: 10.1002/mrm.1910390222

Wang, R. S., Lei, T., Cui, R. X., Zhang, B. T., Meng, H. Y., and Nandi, A. K. (2022). Medical image segmentation using deep learning: a survey. *IET Image Process.* 16, 1243–1267. doi: 10.1049/ipr2.12419

Wang, Z., Yu, Z. Z., Wang, Y., Zhang, H. M., Luo, Y. S., Shi, L., et al. (2020). 3D compressed convolutional neural network differentiates neuromyelitis optical spectrum disorders from multiple sclerosis using automated white matter hyperintensities segmentations. *Front. Physiol.* 11:612928. doi: 10.3389/fphys.2020.612928

Wattjes, M. P., Ciccarelli, O., Reich, D. S., Banwell, B., de Stefano, N., Enzinger, C., et al. (2021). 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *Lancet Neurol.* 20, 653–670. doi: 10.1016/S1474-4422(21)00095-8

Wicks, D. A., Tofts, P. S., Miller, D. H., Du Boulay, G. H., Feinstein, A., Sacares, R. P., et al. (1992). Volume measurement of multiple sclerosis lesions with magnetic resonance images. A preliminary study. *Neuroradiology* 34, 475–479. doi: 10.1007/BF00598953

Yokote, H., and Mizusawa, H. (2016). Multiple sclerosis and neuromyelitis optica spectrum disorders: some similarities in two distinct diseases. *Neural Regen. Res.* 11, 410–411. doi: 10.4103/1673-5374.179048

Zhang, Z. X., Liu, Q. J., and Wang, Y. H. (2018). Road extraction by deep residual U-net. *IEEE Geosci. Remote Sens. Lett.* 15, 749–753. doi: 10.1109/LGRS.2018.2802944

Zhang, H. H., Valcarcel, A. M., Bakshi, R., Chu, R. X., Bagnato, F., Shinohara, R. T., et al. (2019). "Multiple sclerosis lesion segmentation with tiramisu and 2.5D stacked slices", in Medical Image Computing and Computer-assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-assisted Intervention, Berlin, Germany: Springer. 11766, 338–346.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). “Learning deep features for discriminative localization” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016*, Washington, DC, USA: IEEE Computer Society. 2921–2929.

Zhou, Z. W., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. M. (2020). UNet plus plus: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* 39, 1856–1867. doi: 10.1109/TMI.2019.2959609

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 17 June 2024
DOI 10.3389/fgene.2024.1424085

# Identification of DNA motif pairs on paired sequences based on composite heterogeneous graph

Qiuqin Wu[1], Yang Li[2], Qi Wang[1], Xiaoyu Zhao[1], Duanchen Sun[1]* and Bingqiang Liu[1]*

[1]School of Mathematics, Shandong University, Jinan, China, [2]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, United States

OPEN ACCESS

EDITED BY
Fa Zhang,
Beijing Institute of Technology, China

REVIEWED BY
Shiwei Sun,
Chinese Academy of Sciences (CAS), China
Bingbo Wang,
Xidian University, China
Han Wang,
Northeast Normal University, China

*CORRESPONDENCE
Bingqiang Liu,
✉ bingqiang@sdu.edu.cn
Duanchen Sun,
✉ dcsun@sdu.edu.cn

RECEIVED 27 April 2024
ACCEPTED 22 May 2024
PUBLISHED 17 June 2024

CITATION
Wu Q, Li Y, Wang Q, Zhao X, Sun D and Liu B (2024), Identification of DNA motif pairs on paired sequences based on composite heterogeneous graph.
*Front. Genet.* 15:1424085.
doi: 10.3389/fgene.2024.1424085

COPYRIGHT
© 2024 Wu, Li, Wang, Zhao, Sun and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

**Motivation:** The interaction between DNA motifs (DNA motif pairs) influences gene expression through partnership or competition in the process of gene regulation. Potential chromatin interactions between different DNA motifs have been implicated in various diseases. However, current methods for identifying DNA motif pairs rely on the recognition of single DNA motifs or probabilities, which may result in local optimal solutions and can be sensitive to the choice of initial values. A method for precisely identifying DNA motif pairs is still lacking.

**Results:** Here, we propose a novel computational method for predicting DNA Motif Pairs based on Composite Heterogeneous Graph (MPCHG). This approach leverages a composite heterogeneous graph model to identify DNA motif pairs on paired sequences. Compared with the existing methods, MPCHG has greatly improved the accuracy of motifs prediction. Furthermore, the predicted DNA motifs demonstrate heightened DNase accessibility than the background sequences. Notably, the two DNA motifs forming a pair exhibit functional consistency. Importantly, the interacting TF pairs obtained by predicted DNA motif pairs were significantly enriched with known interacting TF pairs, suggesting their potential contribution to chromatin interactions. Collectively, we believe that these identified DNA motif pairs held substantial implications for revealing gene transcriptional regulation under long-range chromatin interactions.

KEYWORDS

**DNA motifs, DNA motif pairs, chromatin interactions, TF pairs, gene transcriptional regulation**

## 1 Introduction

The identification and recognition of DNA motifs binding to transcription factors (TFs) are pivotal for comprehending the regulatory mechanisms governing gene expression and cellular processes (Wong et al., 2013). A DNA motif denotes to a short, similarly repeated pattern of nucleotides that holds biological significance (Hashim et al., 2019). Deciphering these binding DNA motifs provides researchers with insights into the regulation and control of genes, fostering a deeper understanding of diverse biological phenomena (Liu et al., 2018; Yang et al., 2019; Li et al., 2024; Wang et al., 2024). With the development of high-throughput technology, several experimental techniques are available for determining TF binding DNA motifs, such as Chromatin Immunoprecipitation (ChIP) (Park, 2009), Electrophoretic Mobility Shift Assay (EMSA) (Hellman and Fried, 2007), DNA Affinity Purification Sequencing (DAP-seq) (Bartlett et al., 2017), and Systematic Evolution of

Ligands by Exponential Enrichment (SELEX) (Gold, 2015). Moreover, researchers can access relevant databases to query for associated DNA motifs. For instance, JASPAR (Castro-Mondragon et al., 2022) is a widely utilized DNA motif database for storing and analyzing transcription factor binding site. TRANSFAC (Wingender et al., 2000) is a classic database containing DNA motifs of transcription factors and regulatory elements, offering a wealth of DNA motif data and associated biological information. Other databases include UniProbe, Cis-BP, motifMap, ScerTF, TFcat, and FlyTF (Fulton et al., 2009; Pfreundt et al., 2010; Daily et al., 2011; Robasky and Bulyk, 2011; Spivak and Stormo, 2012; Weirauch et al., 2014). However, the action of a single DNA motif is limited, and actual gene regulation often involves intricate interactions among multiple DNA motifs, giving rise to DNA motif pairs (Pilpel et al., 2001). These pairs of DNA motifs play a pivotal role in maintaining the accuracy and flexibility of gene expression (Clauss and Lu, 2023).

When two DNA motifs coexist and interact in a specific manner during gene regulation, they can either cooperate or compete to influence gene expression. This is pivotal for unraveling the intricate mechanisms of gene regulation networks, cell signaling, and biological processes (Kim and Wysocka, 2023). Moreover, these predictions of the interaction between DNA motifs find broad applications in bioinformatics, facilitating genome annotation and the anticipation of protein-nucleic acid interactions, thereby equipping researchers with potent tools to decipher biological data (Khodabandelou et al., 2020; Wang et al., 2022). Lastly, the underlying chromatin interactions between different DNA motifs are associated with various diseases (Bhatia and Kleinjan, 2014). Consequently, predictions based on DNA motif pairs hold promise for discovering new drug targets and innovations in the field of biotechnology, deepening our understanding of gene regulation networks (Makolo and Suberu, 2016).

The essence of DNA motif pairs lies in discerning pattern pairs, specifically identifying statistically significant pattern pairs within two correlated sequences, derived from different sequences. Current methods for identifying DNA motif pairs can be broadly classified into two types. The first approach is direct, involving the independent identification of statistically significant DNA motifs from two correlated sequences. Subsequently, the threshold is calculated to combine the DNA motifs on both sides of sequences to select statistically significant DNA motif pairs. This method may result in the exclusion of DNA motifs capable of forming pairs but are underrepresented. The second approach is based on statistical significance and involves predicting DNA motif pairs through a global optimization model. This method requires constructing a well-designed model for predicting DNA motif pairs. The algorithm developed by Ka-Chun Wong's research group in 2016, referred to as Wong's 2016 (Wong et al., 2016), and EPmotifPair (Wang et al., 2022) both belong to the first category of methods in existing approaches on HI-C (van Berkum et al., 2010) data for predicting DNA motif pairs. Wong's 2016 is presently the first method for identifying DNA motif pairs on HI-C data. It can more flexibly learn sequence features in different directions, such that disturbances in predictions on one side may not affect predictions on the other side. EPmotifPair (Wang et al., 2022) predicts DNA motif pairs in a set of sequences integrated from enhancer sequences and promoter sequences. By comprehensively considering multiple co-occurring sequence patterns, it reduces the error rate compared to the separate prediction of DNA motifs. MotifHyades (Wong, 2017) belongs to the second category of methods for predicting DNA motif pairs. It adopts the probability model and utilizes two derived optimization algorithms to find DNA motif pairs with linear complexities. However, Wong's 2016 (Wong et al., 2016) not only overlooks underrepresented DNA motifs that could have formed pairs but is also time-consuming. EPmotifPair (Wang et al., 2022) not only fails to account for potential interactions between DNA motifs but also requires the specification of numerous parameters, such as the predetermined number of DNA motifs. MotifHyades (Wong, 2017) improves the computational speed and accuracy compared with Wong's 2016 (Wong et al., 2016), but it is sensitive to the choice of the initial value. Additionally, the probability model adopted by MotifHyades (Wong, 2017) assumes conditional independence within each sequence pair, disregarding potential interactions among DNA motifs.

To address the aforementioned challenges, we propose a graph theory-based approach named MPCHG. The methodology is elucidated in Figure 1 (This paper takes Enhancer-Promoter as an example). It helps capture multiple relationships between different *k*-mers, including both within-sequence and between-sequence relationships. Subsequently, a community detection algorithm is employed to obtain a dense subgraph, considering not only the topology of the network but also the practical significance of node connections. Importantly, we refrain from predefining the length of DNA motifs and the number of DNA motif pairs, avoiding the loss of some important DNA motifs or the presence of high noise. We apply MPCHG to analyze seven sets of HI-C data. The results reveal a higher proportion of DNA motifs matching existing databases for predicted DNA motif pairs. The identified paired DNA motifs demonstrate higher DNase accessibility than the background sequences, and the functional consistency of DNA motifs within pairs is evident. Particularly noteworthy is the acquisition of predicted TF pairs from the predicted DNA motif pairs, and we discover that the predicted TF pairs are enriched with the interacting TFs in the STRING database. It can be seen that predicting DNA motif pairs on HI-C data can help us understand the regulatory mechanisms of genes.

## 2 Methods

### 2.1 Data collection

The input data comprises of Hi-C data from seven sets derived from six distinct cell lines, namely,: K562, GM12878, HeLa-S3, HUVEC, IMR90, and NHEK. Two sets of Hi-C data (referred to as K562_1 and K562_2, respectively) are obtained from the K562 cell line, featuring variations in data preprocessing and annotation approaches. A set of protein-protein interaction data retrieved from the STRING database (Mering et al., 2003) serves as benchmark data to assess the performance of predicted DNA motif pairs. The first set of processed Hi-C data from the K562 cell line (K562_1) is acquired from the article published by

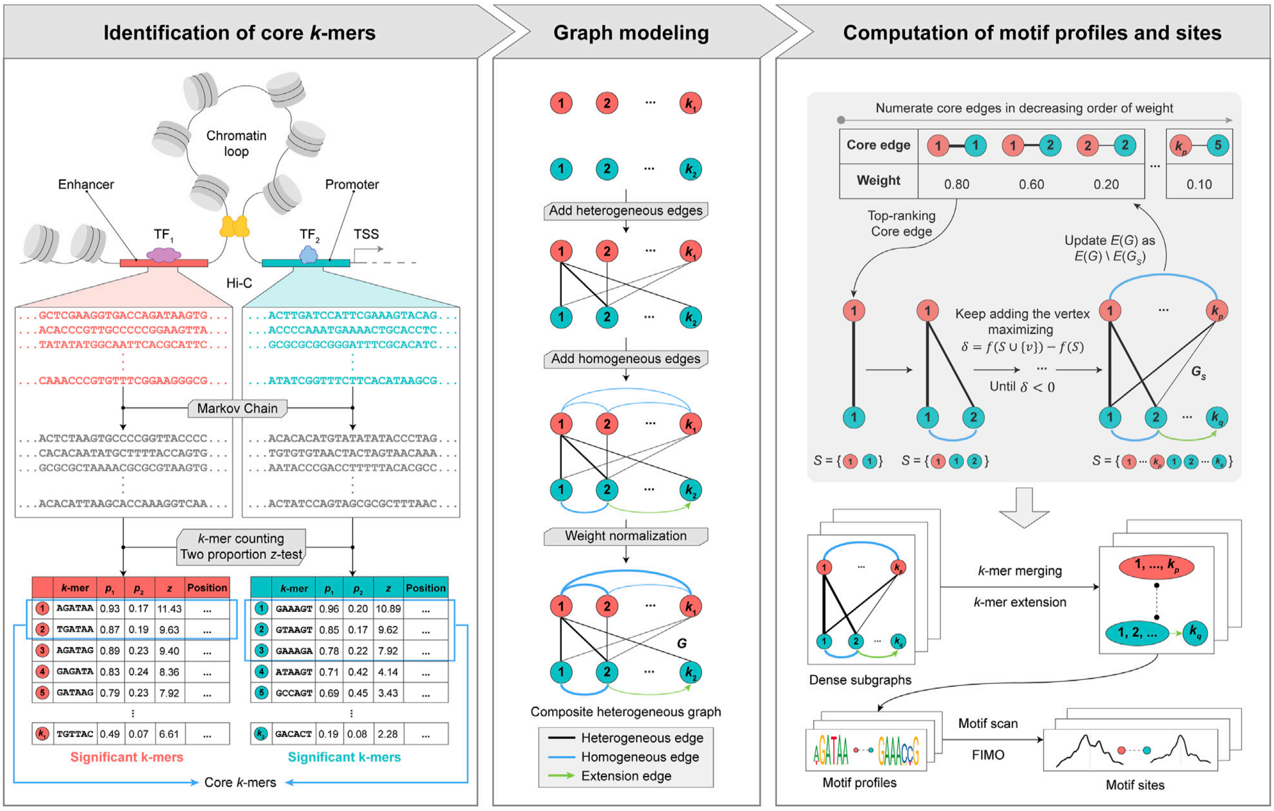


FIGURE 1
Overview of MPCHG. Enhancer (red)—Promoter (blue) interaction is used as an example for DNA motif pairs identification. $p_1$ denotes the frequencies of *k*-mers in the real sequence sets and $p_2$ denotes the frequencies of *k*-mers in the background sequence sets. *z* denotes *z*-score, which is used to measure the significance of *k*-mers. The *k*-mers are arranged in descending order by the size of the *z*-score. The *k*-mers framed by the blue square indicates core *k*-mers. In section Graph modeling, the black lines represent the heterogeneous edges, which connect different types of *k*-mers, and the thickness of the lines indicates the weight of the connected edges, the greater the weight, the thicker the lines. The blue line represents the homogeneous edge, and they will connect the overlapping *k*-mers, and the thickness of the lines indicates the weight of the connected edges, the thicker the line, the greater the weight of the connected edge. The green arrow is called extension edge, indicating the overlap between the two *k*-mers, which can be used to merge and extend the two *k*-mers in subsequent steps.

Ka-Chun Wong in 2016 (Wong et al., 2016). In this study, chromatin fragments are classified into four categories: E (Enhancer), TSS (Promoter), WE (Weak Enhancer), and PF (Promoter-Flanking Region). These categories collectively form 10 interacting pairs, resulting in a total of 74,552 long-range regulatory region pairs. The number of each interaction type is detailed in Supplementary Figure S1A. The remaining six sets of processed HI-C data are sourced from the article published by Wang in 2022 (Wang et al., 2022). Which are normalized using the Knight and Ruiz normalization vectors (Lyu et al., 2020) by Rao *et al.* (Wang et al., 2022). Notably, their chromatin interaction type is exclusively Promoter-Enhancer, in contrast to the first set of data. The long-range regulatory region pairs are summarized in Supplementary Figure S1B. In pursuit of elucidating the mechanism of DNA motif interactions, protein-protein interaction data are obtained from the STRING database, resulting in the extraction of 4,950,896 pairs of experimentally validated data. By comparing the protein names in the STRING database with transcription factors (TFs) in the JASPAR database, experimentally verified TF-TF interactions are identified, encompassing a total of 65,290 TF-TF interactions, involving 583 TFs.

## 2.2 Generation of background sequences

We utilize a third-order Markov model (Eddy, 2004) to create background sequences corresponding to each sequence (referred to as the real sequence) within the input sequence pairs. The generated background sequences are designed to align with the number and length of the given chromatin sequences, and their composition is determined by the nucleotide frequencies observed in the dataset.

## 2.3 Identification of significant **k**-mers

We enumerate all possible *k*-mers (with $k = 6$ by default) employing a sliding window approach in both the real and background sequence sets concurrently. Let $n_F(k_i)$ and $n_B(k_i)$ represent the counts of occurrences of a *k*-mer $k_i$ in the real and background sequence sets, respectively. Similarly, let $p_F(k_i)$ and $p_B(k_i)$ denote the frequency of each *k*-mer $k_i$ in the real and background sequence sets, respectively. Recognizing the reverse complementary nature of DNA, we define the frequency of a *k*-mer as the sum of the frequencies of the *k*-mer and its reverse complementary counterpart. Additionally, we exclude *k*-mers such as AAAAAA due to insufficient variation and discriminative power. Including them in the statistics could introduce noise and compromise the performance of the model. Assuming that the frequency distribution of *k*-mers follows a normal distribution, we retain *k*-mers with frequencies exceeding one standard deviation in the real sequences, deeming these *k*-mers as significant. Subsequently, we maintain the same selection *k*-mers in the background sequences. Following this, we use a two-proportion z-test with the null hypothesis that the frequencies $p_F(k_i)$ in the real sequence sets and $p_B(k_i)$ in the background sequence sets are the same to evaluate the significance of *k*-mers occurrences (Eqs 1-4):

$$H_0\colon\ p_F(k_i) = p_B(k_i), \tag{1}$$

$$H_1\colon\ p_F(k_i) > p_B(k_i), \tag{2}$$

$$z_i = \frac{p_F(k_i) - p_B(k_i)}{\sqrt{p_i(1-p_i)\left(\frac{1}{\sum n_F(k_i)} + \frac{1}{\sum n_B(k_i)}\right)}}, \tag{3}$$

where,

$$p_F(k_i) = \frac{n_F(k_i)}{\sum_j n_F(k_i)},\ p_B(k_i) = \frac{n_B(k_i)}{\sum_j n_B(k_i)},\ p_i = \frac{n_F(k_i) + n_B(k_i)}{\sum n_F(k_i) + \sum n_B(k_i)}. \tag{4}$$

Where the *k*-mer $k_i$ is considered a core *k*-mer if it corresponds to a z-score greater than 1.96.

## 2.4 Construction of composite heterogeneous graph

We treat each *k*-mer as a node and construct a composite heterogeneous graph by establishing edges between them. Based on the positional information of each type of *k*-mer in the real sequence pairs, if two distinct types of *k*-mers are situated in different sequences within a sequence pair, we establish a connection between these two *k*-mers, referring to this connection as pair edges. The weights for pair edges are computed using Eq. 5. The first term in Eq. 5 assesses the practical significance of the edge connection between nodes $v_i$ and $u_j$ based on the number of sequence pairs they co-occur in. If they appear frequently together, the edge weight will be higher. The second and third terms in Eq. 5 consider the topological structure of the graph. They incorporate the number of neighborhoods for nodes $v_i$ and $u_j$, respectively, relative to the total number of k-mers belonging to enhancers and promoters. This helps balance the importance of the nodes in the graph. Next, we introduce the concept of a neighborhood: for *k*-mers of the same type (promoter or enhancer), if one *k*-mer differs from another *k*-mer by only one mismatched base or has at least four consecutive identical bases, we consider the two *k*-mers as neighbors and establish a connection between them, denotes as neighborhood edges. The weights for neighborhood edges are determined using Eq. 6. It considers the proportion of common k-mers between nodes $v_p$ and $v_q$ relative to the total number of k-mers in each node. Higher weights indicate a higher similarity or overlap between the k-mers, which signifies a stronger relationship in the graph. Finally, we normalize the weights for the edges of the graph *G* using Eq. 7 for ensuring that the weights are scaled appropriately relative to each other. In this framework, *k*-mers, treated as nodes, and the interconnected edges between *k*-mers collectively form the weighted heterogeneous graph *G*.

$$\omega(v_i, u_j) = \frac{N(v_i, u_j) - N_{min}(v_i, u_j)}{N_{max}(v_i, u_j) - N_{min}(v_i, u_j)} + \frac{L(v_i)}{n(E)} + \frac{L(u_j)}{m(TSS)}, \tag{5}$$

$$\omega(v_p, v_q) = \frac{L(v_p \cap v_q)}{L(v_p)} + \frac{L(v_p \cap v_q)}{L(v_q)}, \tag{6}$$

$$\omega' = \frac{\omega - \omega_{min}}{\omega_{max} - \omega_{min}}, \tag{7}$$

where, $v_i$ is a *k*-mer belonging to enhancer sequences, $u_j$ is a *k*-mer belonging to promoter sequences, $\mathrm{N}(v_i, u_j)$ represents the number of sequence pairs in which $v_i$ and $u_j$ belong, $L(v_i)$ and $L(u_j)$ represent the number of neighborhoods for $v_i$ and $u_j$ separately, $n(E)$ and $m(TSS)$ represent the num of *k*-mers belonging to enhancers and promoters, respectively. $L(v_p \cap v_q)$ represents the num of union of $v_p$ and $v_q$, $L(v_p)$ and $L(v_q)$ represent the num of $v_p$ and $v_q$, separately. $\omega$ denotes the weights of edges in graph *G*, $\omega_{max}$ and $\omega_{min}$ represent the maximum and minimum weights of edges in graph *G*, respectively.

## 2.5 The acquisition of dense subgraphs

We apply a community discover detection algorithm to identify dense subgraphs. Firstly, we define the fitness function for evaluating the density of a subgraph. Let *S* be a connected subgraph of graph *G*, where *V* represents the vertex set of subgraphs *S*, and $E_S$ represents the edge set of S. Let $n_S = |V_S|$ and $m_S = |E_S|$. By adding $\frac{n_S(n_S-1)}{2} - m_S$ edges to *S*, we form a complete graph $S'$, with the newly added weights are set to the average weight of graph *G*. The density of subgraph *S* is assessed by considering the difference in weights between the existing edges in *S* and the newly added edges. Eq. 8 outlines the community evaluation function $f(S)$ for subgraph *S*:

$$f(S) = \sum_{v_i u_j \in E(S)} \omega(v_i u_j) - \frac{1}{2|E(G)|}\left(n_S(n_S-1) - 2m_S\right) \cdot \sum_{v_i u_j \in E(G)} \omega(v_i u_j). \tag{8}$$

Obviously, the larger $f(S)$, the denser the subgraph *S* in the given sense. For a node $v \notin V_S$, the fitness function $\delta_S(v) = f(S \cup \{v\}) - f(S)$ for $v$ in *S* is defined, and the node that the maximizes fitness function, i.e., $\delta_S(v) > 0$, is added to the existing subgraph *S*.

It is worth noting that when identifying dense subgraphs, we select the point with the highest number of neighborhoods in the core pairs, possessing the highest weight, as the initial point. An iterative process ensues, continuing until no node is found that satisfies the condition, resulting in the formation of the current dense subgraph. Subsequently, we select the two nodes from the pair with the highest weight among the remaining core pairs as the initial nodes for the iterative process. Nodes that do not belong to any dense subgraph are considered isolated points and are excluded from the analysis. For the resulting dense subgraph $C = \{C_1, C_2, \cdots, C_t\}$, where $t$ denotes the number of obtained dense subgraphs, we define the overlap degree of nodes of subgraph $C_i$ and subgraph $C_j\,(1 \le i, j \le t)$ as $|C_i \cap C_j| / \min\,(|C_i|, |C_j|)$. Simply put, it is the count of shared nodes between both $C_i$ and $C_j$ divided by the smaller of the two sets' node counts. If the overlap degree of nodes is greater than 0.5, we merge the two subgraphs $C_i$ and $C_j$. Additionally, during the process of obtaining a dense subgraph, we record the type to which each $k$-mer belongs in the subgraph, as well as the weight of a $k$-mer pair formed from two types of $k$-mer.

## 2.6 Merger and extension of **k**-mers

We extend the $k$-mers identified in the dense subgraphs obtained in the previous step. First, considering that we have recorded the type to which each $k$-mer belongs in the subgraph, we categorize all $k$-mers in each dense subgraph into two groups: enhancer $k$-mers and promoter $k$-mers. The two $k$-mers corresponding to the most weighted $k$-mer pair in each dense subgraph serve as the centers for the two types of $k$-mers. Next, we compare each $k$-mer in each type to the central $k$-mer, determining the position of each $k$-mer by assessing whether the relationship is a mismatch or an overlap. During the construction the position weight matrix (PWM), the frequency of each base corresponds to the frequency of its $k$-mer in the sequence. The two PWMs obtained from the dense subgraph constitute the initial DNA motif pairs. Subsequently, we use FIMO to scan the positions of the two PWMs in the real sequences. If the two PWMs appear in the sequence pair respectively, we consider them to be the final DNA motif pairs.

## 2.7 Evaluation methods for predicted DNA motif pairs

Three evaluation methods are introduced to assess the performance of the predicted DNA motif pairs. Two of these methods are utilized to evaluate the accuracy of the predicted DNA motif pairs, while the third method is employed to assess the enrichment of the predicted DNA motif pairs.

The first evaluation method is DNA motif pair distance ($MPD$), which is defined by MotifHyades and computed using Eq. 9 (Wong, 2017). The metric $MPD$ is employed to assess how well the predicted DNA motif pairs $M = \{(M_P^i, M_E^i) \mid i \in N, i \le K\}$ can be matched to the known DNA motif pairs $m = \{(m_P^i, m_E^i) \mid i \in N, i \le K\}$ inserted into simulated sequence pairs:

$$MPD = \frac{1}{K}\sum_{i=1}^{K} min_{\,x}\left(D\left(m_P^i, M_P^x\right) + D\left(m_E^i, M_E^x\right)\right), \tag{9}$$

where $D(H_1, H_2)$ denoted the standard DNA motif distance between DNA motif $H_1$ and $H_2$ (Wong et al., 2013).

The second evaluation metric is DNA motif pair found ratio ($MPFR$), which is computed by Eq. 10 and used to estimate how many statistically significant DNA motif pairs are found correctly. A DNA motif $H_1$ is deemed a statistically significant ($p < 0.005$) match to another DNA motif $H_2$ when the standard DNA motif distance $D(H_1, H_2)$ is less than 0.5 according to the empirical distribution of random DNA motif patterns (Wong et al., 2013).

$$\mathrm{MPFR} = \frac{1}{K}\sum_{i=1}^{K} I\left[D\left(m_P^i, M_P^{x'}\right) < 0.5 \wedge \left(m_E^i, M_E^{x'}\right) < 0.5\right], \tag{10}$$

where $x' = \arg min_{\,x}\left(D\left(m_P^i, M_P^x\right) + D\left(m_E^i, M_E^x\right)\right)$ and $\mathrm{I}[condition]$ is the Iverson bracket used in mathematical notation and represents logical true-or-false conditions.

The third evaluation metric involves assessing the statistical significance of the enrichment of the predicted TF pairs with known TF pairs through hypergeometric testing, as computed using Eqs 11, 12:

$$p_{value} = phyper\left(m, \frac{n(n-1)}{2}, M, \frac{N(N-1)}{2}\right), \tag{11}$$

where,

$$phyper\left(x_1, y_1, x_2, y_2\right) = \sum_{k=x_1}^{\min\left(y_1, x_2\right)} \frac{y_1!\left(y_2 - y_1\right)!x_2!\left(y_2 - x_2\right)!}{y_2!k!\left(y_1 - k\right)!\left(y_2 - x_2 - y_1 + k\right)!}, \tag{12}$$

$x_1$, $y_1$, $x_2$ and $y_2$ are any non-negative integers. $N$ corresponds to the number of TFs in the STRING database, while $M$ represents the number of TF pairs. Similarly, $n$ and $m$ denote the number of TFs in the predicted TF pairs and the number of predicted TF pairs, respectively.

# 3 Results

## 3.1 Benchmarking MPCHG on simulation datasets and real datasets

To assess the accuracy of predicted motif pairs, we generated a total of 9000 sets of simulated data for different parameters and computed both the DNA motif pair distance and DNA motif pair found ratio for these 9000 sets of simulated data. Additionally, to explore the biological significance of predicted motif pairs, we identified that they may contribute to chromatin interactions based on the transcription factors they bind. Finally, we compared the accuracy of predicted motifs with existing software and found MPCHG to exhibit higher accuracy.

### 3.1.1 The DNA motif pairs predicted by MPCHG obtained high quality DNA motif pair distance and DNA motif pair found ratio on different simulation data

We generate 9000 sets of simulation data to evaluate the performance of MPCHG. The simulated sequences follow a Gaussian distribution with a mean of 500 nucleotides and standard deviation of 20 nucleotides for basic benchmarking. The number of DNA sequence pairs T) is varied from 100 to 1000. Subsequently, we randomly select $H$ DNA motif profile matrices from the JASPAR database, and the number of DNA motif pairs is varied from 3 to 100 through random combinations. We select the base-generating string corresponding to the number with the highest probability based on the distribution of bases at each position within each profile matrix. Afterward, we randomly replace the selected strings in the sequence pairs with the generated string pairs. The complete performance spectrum is visualized in Figure 2A and Figure 2B. The DNA motif pairs identified by MPCHG consistently exhibited high-quality DNA

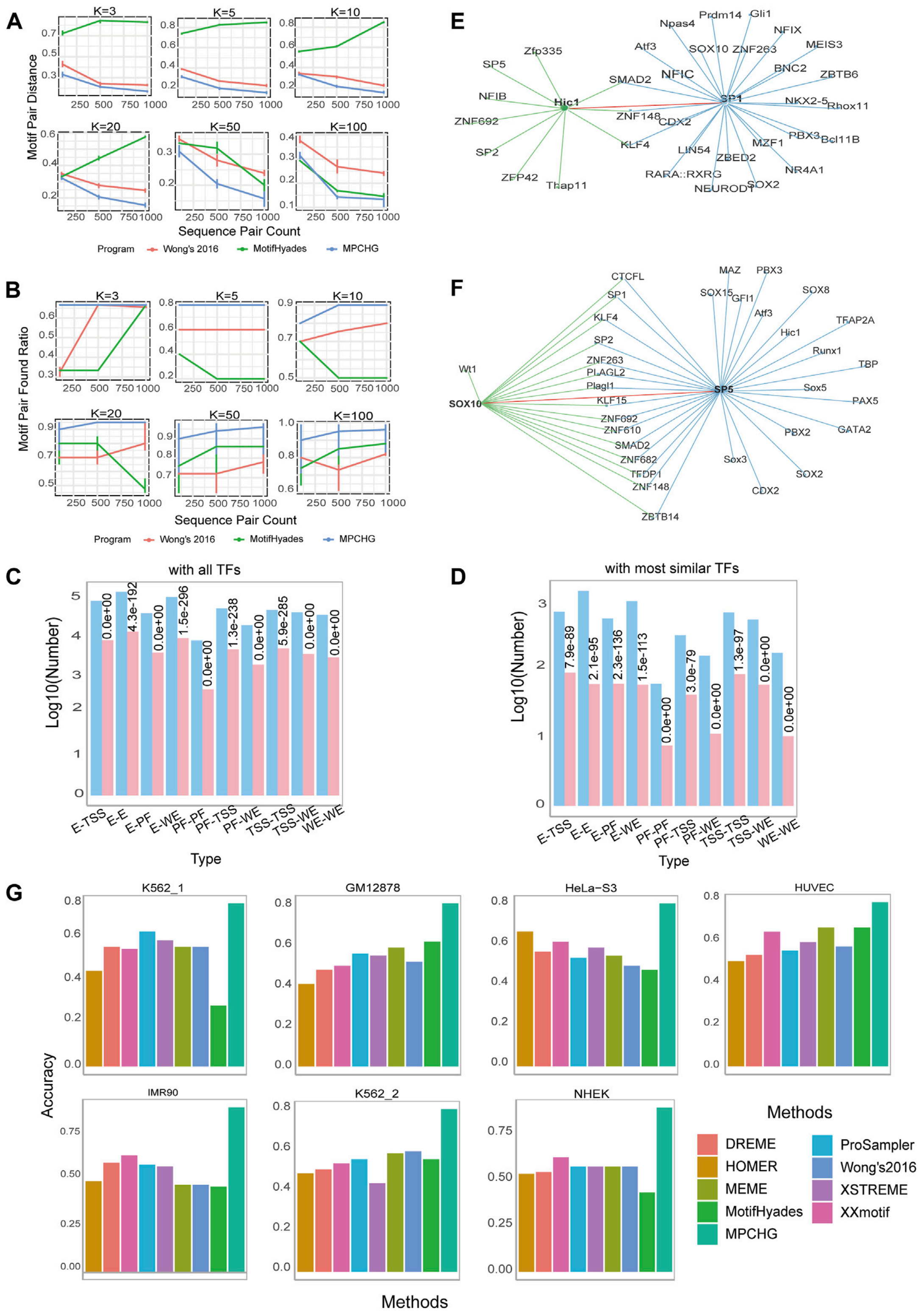


FIGURE 2
Performance of MPCHG on simulation datasets and real datasets. **(A)** Line chart for Motif Pair Distances (i.e., MPD). on known DNA motifs from JASPAR. **(B)** Line chart for Motif Pair Found Ratio (i.e., MPFR). on known DNA motifs from JASPAR. **(C,D)** Histogram on the predicted DNA motif pairs enriched with known interacting TF pairs. The red columns indicate the predicted log10 (TF pairs num) and the blue columns indicate the predicted log10 (TF pairs num supported by STRING database, which is experimentally proven). **(E)** The TF pair Hic1-SP1 in the network of the TFs corresponding to the predicted DNA motifs in K562_1 cell line. The green line represents the TFs interacting with TF Hic1 and the blue line represents the TFs interacting with TF SP1. The red line indicates the interaction between Hic1 and SP1, leaving out some of the lines between the interacting TFs. **(F)** The TF pair SOX10-

(*Continued*)

FIGURE 2 (Continued)
SP5 in the network of the TFs corresponding to the predicted DNA motifs in HUVEC cell line. The green line represents the TFs interacting with TF SOX10 and the blue line represents the TFs interacting with TF SP5. The red line indicates the interaction between SOX10 and SP5, leaving out some of the lines between the interacting TFs. **(G)** Histogram of the accuracy comparison between MPCHG and other six methods on seven datasets. The horizontal axis represents different methods, while the vertical axis indicates the accuracy of predicted DNA motifs.

motif pair distances and DNA motif pair found ratios across diverse simulation datasets. In addition, we can see from Figure 2A and Figure 2B that with more sequence pairs, MPD decreases while MPFR increases, suggesting MPCHG's better generalization on larger datasets and its potential for enhanced robustness, leading to more reliable and accurate predictions.

### 3.1.2 The interacting TF pairs obtained by predicted DNA motif pairs were significantly enriched with known interacting TF pairs and the TF pairs obtained by predicted DNA motif pairs may contribute to chromatin interactions

It is widely recognized that the interaction between DNA motifs is facilitated by transcription factors (TFs). Thus, we predict TF interactions based on the interactions between DNA motifs (Yu et al., 2006). Utilizing the JASPAR database, we can retrieve information about which TFs bind to each DNA motif. Subsequently, we compare the predicted DNA motifs with DNA motifs in the JASPAR (NON-REDUNDANT) DNA-JASPAR CORE (2022) vertebrates database to identify the TFs associated with the predicted DNA motifs. Based on the interactions between DNA motifs and the TFs bound by each DNA motif, we derive TF pairs. During the process of obtaining TF pairs from DNA motif pairs, it is noteworthy that a DNA motif may bind to multiple TFs. Therefore, we consider two approaches for the TFs associated with a predicted DNA motif: one involves including all TFs for the predicted DNA motif, while the other involves considering only 1 TF. For a given DNA motif, we first identify the most similar DNA motif in the database, i.e., the DNA motif corresponding to the lowest $p\_value$. Subsequently, we designate the TF of this most similar DNA motif as the TF of our predicted DNA motif. This yields two types of TF pairs corresponding to predicted DNA motif pairs.

To assess whether the predicted TF pairs are enriched with known TF pairs, we collect experimentally validated interacting TFs in STRING database (Szklarczyk et al., 2023). Then, we use hypergeometric testing to calculate the $p_{value}$, evaluating the statistical significance of the enrichment of the predicted TF pairs with known TF pairs. The findings for the K562_1 cell line are illustrated in Figure 2C and Figure 2D, while the results for the remaining 6 cell lines are presented in Supplementary Figures S2A, S2B. These figures unveil a notable and statistically significant enrichment of the predicted interacting transcription factor (TF) pairs with the established TF interactions in the STRING database.

The TF pairs we have predicted are likely to play a role in chromatin interactions. To illustrate, by comparing our predicted TF pairs with experimentally validated TF pairs in the STRING database, we identify a novel predicted TF pair, HIC1-SP1, as depicted in Figure 2E. HIC1 is a transcription factor (TF) classified as a member of the BTB/POZ (Broad complex, Tramtrack, Bric à brac or poxvirus and zinc finger) zinc finger family. These TFs are characterized by the presence of an N-terminal POZ domain involved in protein-protein interactions and a C-terminal zinc-finger binding domain for direct DNA interaction. A recent report reveals that HIC1 can act as both a transcriptional repressor and an activator during induction of human regulatory T cells (Ray and Chang, 2020). SP1, also known as specificity protein 1*, is a protein that in humans is encoded by the SP1 gene. The protein encoded by this gene is a zinc finger transcription factor that binds to GC-rich DNA motifs of many promoters (Al-Sarraj et al., 2005). Notably, Hypoxia repressed SIRT1 transcription through promoting the competition between Sp1 and HIC1 on the SIRT1 proximal promoter in a SUMOylation-dependent manner (Sun et al., 2013). Based on this, the competitive relationship between SP1 and HIC1 may regulate gene transcription by influencing chromatin structure and status. Furthermore, another novel DNA motif pair, SOX10-SP5, as illustrated in Figure 2F, is predicted in HUVEC cell line. Sox10 is present in all neural crest cells and plays a particularly vital role in determining the fate, viability, and maturation of Schwann cells originating from neural crest stem cells (Mao et al., 2014). SP5 binds to the GC box, a DNA motif present in the promoter of a very large number of genes (Harrison et al., 2000), and is an essential early regulator of neural crest specification in *xenopus* (Park et al., 2013). Furthermore, experimentally validated by Choi et al. demonstrated that knocking down Sp5 on the initial steps of neural crest development could result in complete loss or reduction of the expression of NC markers Sox10 (Park et al., 2013). Thus, it is likely that the interaction of SOX10-SP5 contributes to chromatin interactions, allowing their transcripts to co-localize in the neural crest region (Park et al., 2013).

### 3.1.3 MPCHG achieved a higher accuracy than existing methods in identifying DNA motifs

We finally assess the accuracy of the DNA motifs obtained in the intermediate process to understand the degree of overlap with existing DNA motifs. We conduct a comparative analysis of MPCHG against six state-of-the-art DNA motif-finding tools, namely, DREME (Bailey, 2011), HOMER (Heinz et al., 2010), MEME (Bailey et al., 2006), ProSampler (Li et al., 2019), XSTREME (Grant and Bailey, 2021), and XXmotif (Hartmann et al., 2013). All these tools utilize the JASPAR database as a reference and employed the TomTom software (Gupta et al., 2007) with default parameters for assessment. In particular, MEME requires user to specify the number of DNA motifs, and after systematic testing at output settings of 50, 100, 150, and 200 DNA motifs, the optimal parameter of 50 is determined (yielding the highest accuracy in comparison with the JASPAR database). The remaining parameters of the above-mentioned algorithm are set to their default values, the accuracy of each method can be observed in Figure 2G. The results show that the

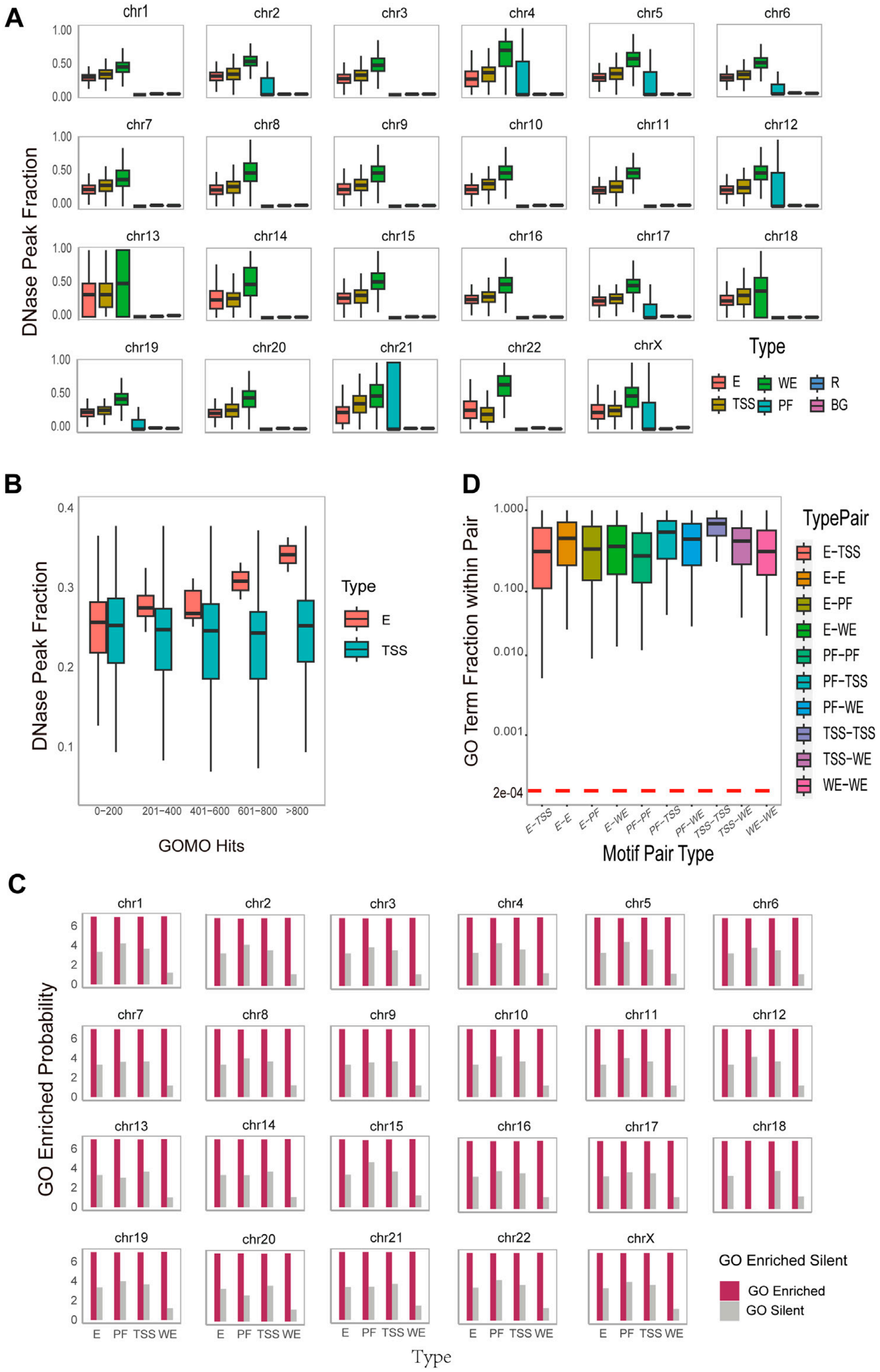


FIGURE 3
Functional and Spatial-level analysis of identified motif pairs. **(A)**. Box plots on the DNase hypersensitivity peak fraction of the DNA motifs found on different region types (i.e., WE (Weak Enhancer), E (Enhancer), TSS (Promoter), PF (Promoter-Flanking Region), R (Regulatory Region Background), BG (Background)) on different chromosomes. The horizontal axis represents different type of DNA motifs, while the vertical axis, DNase Peak Fraction, represents the ratio of the number of DNA motifs that overlap with DNase hypersensitive sites to the total DNA motifs. **(B)**. Box plots on the DNase hypersensitivity peak fraction of the DNA motifs found on different region (i.e., E (Enhancer) and TSS (Promoter)) with varying numbers of enriched Gene Ontology (GO) terms. **(C)**. Histogram of GOMO gene ontology enrichment results, with DNA motifs identified and sorted by type (horizontal axis), and the vertical axis is converted to 7+log(probability the proportion of DNA motifs with at least one GO term in each type). For each DNA motif, the term "GO Enriched" indicates that it has at least one statistically significant GO term identified by GOMO, while the term "Silent" indicates that there is no statistically significant GO term identified by GOMO. **(D)**. Boxplot on the overlap coefficients (Szymkiewicz-Simpson coefficients) between the enriched GO terms of (*Continued*)

FIGURE 3 (Continued)
the first DNA motif and those of the second DNA motif within each DNA motif pair. The arrangement is sorted by type on the horizontal axis. A horizontal red dashed line serves as a reference for the expected overlap coefficient under the null hypothesis. This assumption posits that the overlap is entirely random, featuring a uniform hit distribution for all identified GO terms in this study conducted by GOMO.

accuracy of motifs predicted by MPCHG on 7 sets of data ranges from 75.0% to 88.7%. In contrast, the accuracy of the other six methods range from 28.0% to 66.7%. Where, MotifHyades exhibits the lowest accuracy at 28.0% on K562_1 cell line. Overall, MPCHG demonstrates an improvement of around 60% in accuracy compared to the other six methods. Notably, the accuracy of MPCHG averaged around 80% across various cell lines, indicating its high robustness.

## 3.2 DNA motif spatial accessibility and functional correlations provide insights into predicted DNA motif pairs

Exploring the spatial accessibility of motif pairs can unveil their mechanisms of action in gene regulation. By assessing the spatial accessibility of these motif pairs, we can determine which gene regions are more prone to transcription factor binding, thus gaining deeper insights into key nodes within the gene regulatory network. Furthermore, investigating the functional correlations between motif pairs can reveal their synergistic roles and functional regulations in biological processes, thereby understanding their functions and regulatory mechanisms in specific biological processes.

### 3.2.1 DNA motifs predicted by MPCHG are spatial accessible and DNase peak fractions of different type of DNA motifs have different correlation to the number of enriched GO terms

Exploring the accessibility of DNA motifs is instrumental in identifying gene regions prone to transcription factor binding, offering insights into the underlying mechanisms of gene regulation. To investigate DNA motifs accessibility, we download the DNase Chip-seq peak-calling data (Supplementary Table S1) from the ENCODE consortium (Dunham et al., 2012) across 6 cell lines. We calculate how many DNA motifs overlap with DNase hypersensitive sites on the reference hg19 human genome. To measure the significance of DNase Peak Fraction, we adopt the approach used by Wong in 2016 (Wong et al., 2016). For each DNA motif instance, we randomly sample 100 sites of the same width from both the regulatory region and the entire region of the same chromosome. This process yields regulatory region background DNase peak fractions (denoted as R) and overall background DNase peak fractions (denoted as BG) for each chromosome, respectively. The result (Figure 3A) for K562_1 cell line and the result (Supplementary Figures S3–S8) for other 6 cell line illustrate the DNase Peak Fraction for different types of DNA motifs on each chromosome individually. As depicted in the Figure 3A, the DNase Peak Fraction consistently follows a pattern across various DNA motif types: WE motifs exhibit the highest DNase Peak Fraction, followed by TSS motifs, and the lowest PF, except for the 22nd chromosome. E motifs and TSS motifs have relatively almost the same size DNase Peak Fraction, but both are higher than R and BG motifs. This suggests that WE motifs are more inclined to be open, followed by TSS motifs, and this overlapping fraction is statistically significant. We conduct t-tests and Mann-Whitney tests to measure the statistical significance of the difference between the identified DNA motifs and those in the background region. The result indicates that all $p-values$ are less than 0.01, signifying a significant overlap between DNA motifs predicted by our method and DNase hypersensitive sites.

Furthermore, DNase peak fractions exhibit distinct correlations with the number of enriched GO terms for different type of DNA motifs. As illustrated in Figure 3B, for enhancer motifs, the DNase peak fraction displays a positively correlation with the number of enriched GO terms, while for TSS motifs, it remains almost unchanged. This observation may be attributed to the fact that enhancers, responsible for gene expression regulation, are typically located in open chromatin regions known as DNase hypersensitive sites. These sites, susceptible to nucleases like DNase I, represent chromatin regions that are not tightly bound in the nucleus, allowing easier access to DNA structures by regulatory elements such as transcription factors. Consequently, the increase in the number of GO terms associated with enhancer motif enrichment and their overlap ratio with DNase hypersensitive sites may be attributed to the likelihood of these enhancers being situated in open chromatin regions. This accessibility facilitates interactions with regulators, influencing the enrichment of GO terms. On the other hand, TSS motifs are commonly found in the promoter region of a gene, associated with the transcription start site. While these motifs play a crucial role in gene initiation, an increase in their number does not lead to a significant change in the overlap ratio with TSS. This is because the location of TSS motifs in the promoter region is relatively fixed, and there is no direct correlation with an increase in the number of GO terms. Despite the increase in enriched GO terms for enhancer motifs, these terms do not directly impact the distribution of TSS motifs. Therefore, the overlap ratio with TSS remains largely unchanged. This phenomenon underscores the importance of distinguishing between various regulatory elements and factors in the study of gene regulation. It emphasizes the necessity of considering their intricate interactions within the gene expression regulatory network.

### 3.2.2 DNA motifs predicted by MPCHG are enriched with GO terms and the two DNA motifs coupled within one DNA motif pair are functional consistency

Ontology enrichment analysis serves as a crucial bioinformatics tool, facilitating the identification of significant enrichment in a group of genes or gene-associated entities in biological functions and processes (Peng et al., 2019). This analysis provides comprehensive insights into the functional characteristics of the study subject, shedding light on its significant roles in biology. To conduct this analysis, we use GOMO software for Gene Ontology enrichment on

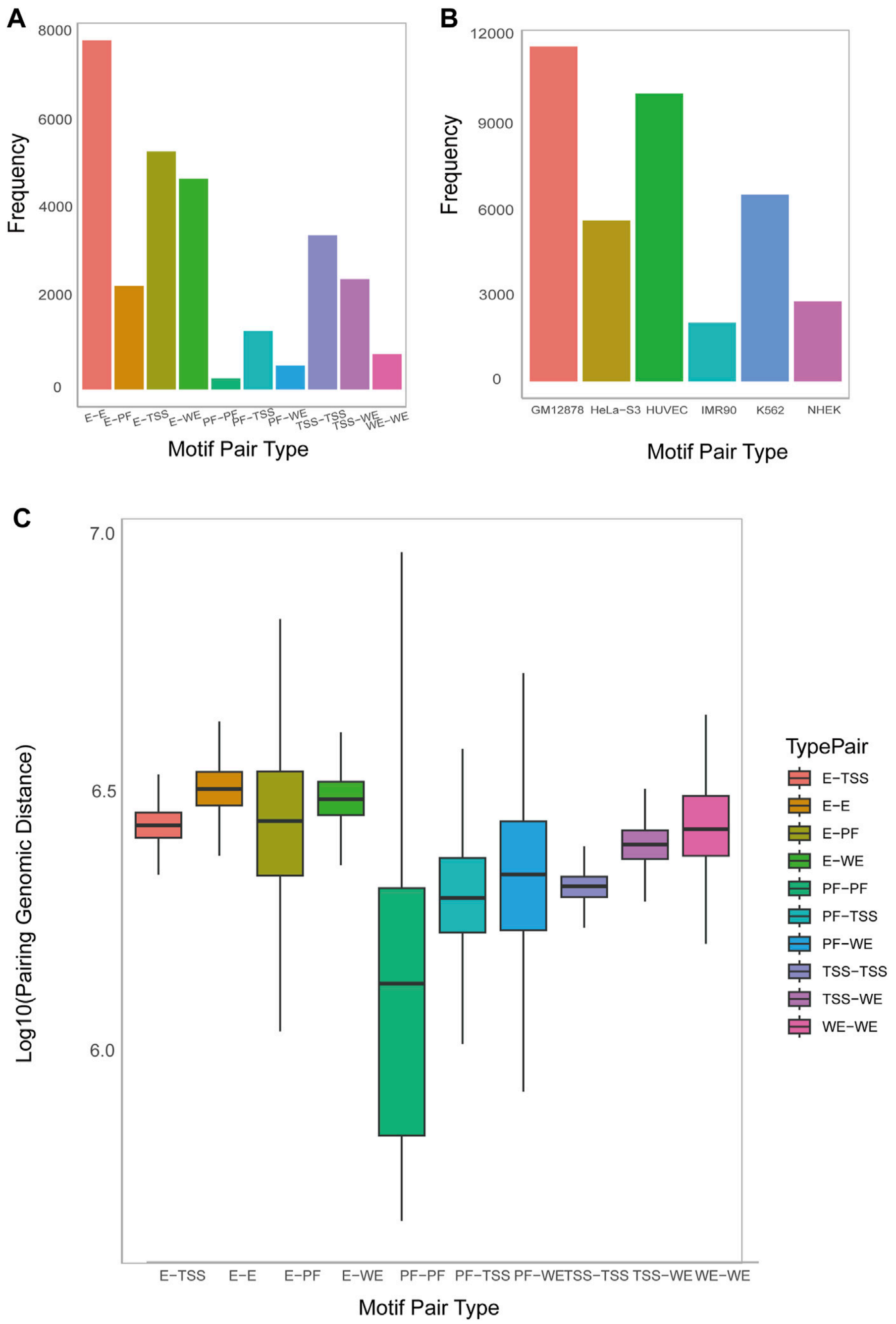


FIGURE 4
The num and spatial distribution relationship of identified DNA motif pairs. **(A)**. Histogram on the number of predicted motif pairs annotated to different types by ChromHMM and Segway (i.e., E (Enhancer), WE (Weak Enhancer), TSS (Promoter), and PF (Promoter-Flanking Region)). **(B)**. Histogram on the number of predicted motif pairs annotated by six cell lines. Their chromatin interaction type is exclusively Promoter-Enhancer. **(C)**. Boxplot on the average genomic distance between the motif instances of the first DNA motif and those of the second DNA motif within each DNA motif pair, sorted by type (horizontal axis).

each DNA motif obtained (Buske et al., 2010). In short, GOMO scans all promoters using the provided DNA motifs to determine if any DNA motif is significantly associated with genes linked to one or more Gene Ontology (GO) terms. This process is significant for understanding the biological roles of the DNA motifs. The results are depicted in Figure 3C and Supplementary Figures S9–S14. Notably, on average, more than 97% of DNA motifs exhibit enrichment for at least one GO term. This observation suggests that the predicted DNA motifs play a discernible role in gene regulation, cellular processes, or other biological functions, and their functions may be relatively extensive and universal. Among the top frequent terms, we observe the DNA motifs-related GO terms such as (GO:0048731 system development) (GO: 0048513 animal organ development), (GO:0030154 cell differentiation), and (GO:0003700 DNA-binding transcription factor activity).

Furthermore, our interest extends to the functional roles between the two DNA motifs within each DNA motif pair. To explore this, we calculate the overlap coefficient (Szymkiewicz-Simpson coefficient) between the enriched GO terms of the first DNA motif and those of the second DNA motif within each DNA motif pair. The results of the overlap coefficient are illustrated in Figure 3D and Supplementary Figure S15. The observed overlap coefficients are higher than expected, indicating a substantial overlap between the two DNA motif-related GO term set. This suggests a potential functional or biological correlation between the two motifs. Notably, the overlap coefficient for TSS-TSS interaction is the highest, implying that interactions between promoters may be functionally more closely related, involved in more common biological processes, and exhibit stronger functional correlations. These findings provide valuable insights for a deeper understanding of promoter interaction in gene regulatory network and biological processes. Additionally, they offer guidance for further functional annotation and research into regulatory mechanisms.

## 3.3 DNA motif pairs predicted by MPCHG unveiled genomic distance characteristics in human cell lines

To analyze genomic distance signatures within chromatin structures, we first counted the motif pairs predicted by MPCHG on seven cell lines. Through genomic distance analysis of the predicted motif pairs, MPCHG reveals the spatial relationships and interactions between the regulatory elements. Notably, these findings highlight the universality of long-distance regulatory mechanisms, and in particular enhancers play a key role in facilitating precise gene regulation.

### 3.3.1 DNA motif pairs were discovered by MPCHG on seven human cell lines

MPCHG has run on the seven cell lines (K562_1, GM12878, HeLa-S3, HUVEC, IMR90, K562_2, NHEK) to obtain ten thousand of DNA motif pairs. We counted the number of DNA motif pairs of 10 chromatin interaction types on the K562_1 cell line and the number of promoter-enhancer-pairs on the remaining six cell lines. The discovered DNA motif pairs are visualized Figure 4A and Figure 4B.

### 3.3.2 The genomic distance between DNA motifs pairs predicted by MPCHG revealed the interaction and relative position between the regulatory elements

Analyzing the distances between DNA motifs provides insights into the relative positioning and interactions of gene regulatory elements, indicating whether they are in close proximity or distantly located within the three-dimensional chromatin structure (Dekker and Misteli, 2015). This analysis enhances our understanding of the organization and spatial regulation of gene expression at the chromatin level. Therefore, Accordingly, we have computed the distance between DNA motifs of different interaction types. As depicted in Figure 4C, the interaction distance between E-E is the greatest, followed by E-WE, E-TSS, and E-PF and Supplementary Figure S16 also indicate that enhancers are far away from Promoters. This observation aligns with the widely accepted notion that enhancers are typically situated in regions far away from the genes they regulate, sometimes spanning millions of base pairs (bp). This long-distance regulatory action is facilitated through the establishment of chromatin loops, enabling effective and precise regulatory interactions.

# 4 Discussion

Identifying DNA motifs is of paramount importance in biology and computational biology. DNA motifs are short sequence patterns in protein or nucleic acid sequences that are functionally relevant. They are crucial for functional annotation, structure prediction, evolutionary relationships, and regulatory element recognition. Furthermore, the identification of DNA motif pairs in interacting sequences is also significant as it aids in predicting protein-protein interactions, drug design, and disease research. In conclusion, DNA motifs and their pairs play pivotal roles in biological research and medical applications.

Hence, we propose the MPCHG algorithm to identify tens of thousands of DNA motif pairs in the long-range chromatin interaction sequences. First, we use a 3-order Markov model to generate background sequences that matches the length and composition of the original sequence, ensuring statistical significance and rationality for *k*-mer seeds. In contrast to many algorithms that exhaustively determine DNA motif length within a specific range, our method extends the core DNA motif to both ends using a double-sample z-test. This approach aligns the predicted DNA motif more closely with real scenarios. At the same time, algorithms that set the DNA motif length in advance may miss some important DNA motifs or introduce high noise. Furthermore, we construct a composite heterogeneous graph for different types of *k*-mers (enhancer *k*-mers from enhancer sequences and promoter *k*-mers from promoter sequences). This graph connects *k*-mers of different types present in the same sequence pair. Simultaneously, it captures complex relationships among *k*-mers of the same type with mismatched or overlapping connections. To obtain a dense subgraph related to *k*-mers, we define the fitness function of the subgraph to assess its density. Nodes meeting specific conditions are extended to the current seed. Finally, we merge and extend the obtained subgraph to extract DNA motifs. Subsequently DNA motifs scanning enables the identification of DNA motif pairs.

Regarding the predicted DNA motif pairs, we conducted a thorough analysis covering various aspects. The accuracy rate, measured by comparing predicted DNA motifs with the JASPAR database using TOMTOM software, the accuracy of predicted DNA motifs ranged from 75.0% to 88.7%. Additionally, we employed GOMO software to explore the gene ontology enrichment of these predicted DNA motifs. The findings revealed that, on average, over 97% of DNA motifs are enriched for at least one GO term. This indicated that the predicted DNA motif play essential roles in gene regulation, cellular processes, or other biological functions, and their functions may be relatively extensive and universal. To further validate our predictions, we compare the predicted DNA motifs with DNase Chip-seq peak-calling data. The analysis demonstrated a significant overlap between DNA motifs predicted by our model and DNase hypersensitive sites. Notably, DNA motif pairs involving enhancer or weak enhancer regions exhibited greater distance, aligning with the common understanding that regulatory components in enhancer regions are typically located far from their interacting partners, often spanning a large genomic distance. We extended our analysis to predict TF interactions based on the predicted DNA motif pairs. The result indicated that the predicted interacting TF pairs are significantly enriched with the known interacting TF pairs in STRING, as determined by hypergeometric testing. Moreover, we unveiled new TF interaction information, such as the interaction between HIC1 and SP1, suggesting a potential role in facilitating chromatin interactions and promoting gene transcription. Finally, to evaluate the generalization performance of our model, we tested it on six additional E-TSS datasets representing different cell lines (GM12878, HeLa-S3, HUVEC, IMR90, K562, and NHEK). The results demonstrated consistently good performance across these diverse datasets.

The prediction of DNA motif pairs stands as a critical challenge in bioinformatics, offering valuable insights into various biological processes, including gene regulation, protein-protein interactions, and RNA structures. While significant strides have been made in this field, the future holds immense potential for further advancements. Firstly, the continuous evolution of deep learning and artificial intelligence techniques, including innovative algorithms and graph neural networks, is expected to elevate the accuracy and reliability of DNA motif pair predictions. Secondly, the exploration of cross-species DNA motif pair prediction presents an intriguing challenge, offering opportunities to uncover conserved sequence patterns and explore evolutionary variations. Thirdly, the integration of diverse data sources, such as epigenetic data and protein interaction information, will contribute to more comprehensive annotations for predicted results, enhancing our understanding of the intricacies of biological systems. Additionally, applying DNA motif pair predictions in disease research and precision medicine holds promise for identifying potential disease markers or therapeutic targets. Lastly, the combination of DNA motif pair predictions with network interactions and systems biology approaches will enable the construction of comprehensive biological regulatory network models. This integrative approach has the potential to deepen our understanding of the fundamental principles of biology. In conclusion, ongoing research in predicting DNA motif pairs has significant potential to drive breakthroughs in biotechnology and medical advancements, fostering progress in the fields of biology and medicine.

## Data availability statement

Original datasets are available in a publicly accessible repository: The original contributions presented in the study are publicly available. The input data of Hi-C data for six cell lines (K562, GM12878, HeLa-S3, HUVEC, IMR90, and NHEK) can be found here: https://doi.org/10.6084/m9.figshare.14192000. The protein-protein interaction data from the STRING database can be found here: https://cn.string-db.org/cgi/download?sessionId=blEDX7XXpqWC.

## Author contributions

QWu: Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft. YL: Conceptualization, Project administration, Supervision, Writing–review and editing. QWa: Investigation, Methodology, Visualization, Writing–review and editing. XZ: Investigation, Methodology, Visualization, Writing–review and editing. DS: Conceptualization, Project administration, Resources, Supervision, Writing–review and editing. BL: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing–review and editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. National Key R&D Program of China (2020YFA0712400); National Nature Science Foundation of China (62272270 and 11931008); Shandong University Multidisciplinary Research and Innovation Team of Young Scholars (2020QNQT017). National Natural Science Foundation of China [No. 62202269]; Science Foundation Program of the Shandong Province (2023HWYQ-012); the Program of Qilu Young Scholars of Shandong University.

## Acknowledgments

The authors would like to thank Yihua Wang for his assistance in downloading the data.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1424085/full#supplementary-material

# References

Al-Sarraj, A., Day, R. M., and Thiel, G. (2005). Specificity of transcriptional regulation by the zinc finger transcription factors Sp1, Sp3, and Egr-1. *J. Cell Biochem.* 94 (1), 153–167. doi:10.1002/jcb.20305

Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27 (12), 1653–1659. doi:10.1093/bioinformatics/btr261

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373. doi:10.1093/nar/gkl198

Bartlett, A., O'Malley, R. C., Huang, S. S. C., Galli, M., Nery, J. R., Gallavotti, A., et al. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.* 12 (8), 1659–1672. doi:10.1038/nprot.2017.055

Bhatia, S., and Kleinjan, D. A. (2014). Disruption of long-range gene regulation in human genetic disease: a kaleidoscope of general principles, diverse mechanisms and unique phenotypic consequences. *Hum. Genet.* 133 (7), 815–845. doi:10.1007/s00439-014-1424-6

Buske, F. A., Boden, M., Bauer, D. C., and Bailey, T. L. (2010). Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics* 26 (7), 860–866. doi:10.1093/bioinformatics/btq049

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 50 (D1), D165–D173. doi:10.1093/nar/gkab1113

Clauss, B., and Lu, M. (2023). A quantitative evaluation of topological motifs and their coupling in gene circuit state distributions. *iScience* 26 (2), 106029. doi:10.1016/j.isci.2023.106029

Daily, K., Patel, V. R., Rigor, P., Xie, X., and Baldi, P. (2011). MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *Bmc Bioinforma.* 12, 495. doi:10.1186/1471-2105-12-495

Dekker, J., and Misteli, T. (2015). Long-range chromatin interactions. *Csh Perspect. Biol.* 7 (10), a019356. doi:10.1101/cshperspect.a019356

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414), 57–74. doi:10.1038/nature11247

Eddy, S. R. (2004). What is a hidden Markov model? *Nat. Biotechnol.* 22 (10), 1315–1316. doi:10.1038/nbt1004-1315

Fulton, D. L., Sundararajan, S., Badis, G., Hughes, T. R., Wasserman, W. W., Roach, J. C., et al. (2009). TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.* 10 (3), R29. doi:10.1186/gb-2009-10-3-r29

Gold, L. (2015). SELEX: how it happened and where it will go. *J. Mol. Evol.* 81 (5-6), 140–143. doi:10.1007/s00239-015-9705-9

Grant, C. E., and Bailey, T. L. (2021). XSTREME: comprehensive motif analysis of biological sequence datasets. *BioRxiv* 2021-09. doi:10.1101/2021.09.02.458722

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biol.* 8 (2), R24. doi:10.1186/gb-2007-8-2-r24

Harrison, S. M., Houzelstein, D., Dunwoodie, S. L., and Beddington, R. S. (2000). Sp5, a new member of the Sp1 family, is dynamically expressed during development and genetically interacts with Brachyury. *Dev. Biol.* 227 (2), 358–372. doi:10.1006/dbio.2000.9878

Hartmann, H., Guthohrlein, E. W., Siebert, M., Luehr, S., and Soding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. *Genome Res.* 23 (1), 181–194. doi:10.1101/gr.139881.112

Hashim, F. A., Mabrouk, M. S., and Al-Atabany, W. (2019). Review of different sequence motif finding algorithms. *Avicenna J. Med. Biotechnol.* 11 (2), 130–148.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38 (4), 576–589. doi:10.1016/j.molcel.2010.05.004

Hellman, L. M., and Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat. Protoc.* 2 (8), 1849–1861. doi:10.1038/nprot.2007.249

Khodabandelou, G., Routhier, E., and Mozziconacci, J. (2020). Genome annotation across species using deep convolutional neural networks. *PeerJ Comput. Sci.* 6, e278. doi:10.7717/peerj-cs.278

Kim, S., and Wysocka, J. (2023). Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* 83 (3), 373–392. doi:10.1016/j.molcel.2022.12.032

Li, Y., Ni, P., Zhang, S., Li, G., and Su, Z. (2019). ProSampler: an ultrafast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. *Bioinformatics* 35 (22), 4632–4639. doi:10.1093/bioinformatics/btz290

Li, Y., Wang, Y., Wang, C., Ma, A., Ma, Q., and Liu, B. (2024). A weighted two-stage sequence alignment framework to identify motifs from ChIP-exo data. *Patterns* 5, 100927. doi:10.1016/j.patter.2024.100927

Liu, B., Yang, J., Li, Y., McDermaid, A., and Ma, Q. (2018). An algorithmic perspective of *de novo* cis-regulatory motif finding based on ChIP-seq data. *Brief. Bioinform* 19, 1069–1081. doi:10.1093/bib/bbx026

Lyu, H., Lium, E., and Wu, Z. (2020). Comparison of normalization methods for Hi-C data. *Biotechniques* 68 (2), 56–64. doi:10.2144/btn-2019-0105

Makolo, U. A., and Suberu, S. O. (2016). Gapped motif discovery with multi-objective genetic algorithm. *OALib* 03 (03), 1–6. doi:10.4236/oalib.1102293

Mao, Y., Reiprich, S., Wegner, M., and Fritzsch, B. (2014). Targeted deletion of Sox10 by Wnt1-cre defects neuronal migration and projection in the mouse inner ear. *Plos One* 9 (4), e94580. doi:10.1371/journal.pone.0094580

Mering, C. V., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31 (1), 258–261. doi:10.1093/nar/gkg034

Park, D. S., Seo, J. H., Hong, M., Bang, W., Han, J. K., and Choi, S. C. (2013). Role of Sp5 as an essential early regulator of neural crest specification in xenopus. *Dev. Dyn.* 242 (12), 1382–1394. doi:10.1002/dvdy.24034

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10 (10), 669–680. doi:10.1038/nrg2641

Peng, J., Lu, G., Xue, H., Wang, T., and Shang, X. (2019). TS-GOEA: a web tool for tissue-specific gene set enrichment analysis based on gene ontology. *Bmc Bioinforma.* 20, 572. doi:10.1186/s12859-019-3125-6

Pfreundt, U., James, D. P., Tweedie, S., Wilson, D., Teichmann, S. A., and Adryan, B. (2010). FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database. *Nucleic Acids Res.* 38, D443–D447. doi:10.1093/nar/gkp910

Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29 (2), 153–159. doi:10.1038/ng724

Ray, H., and Chang, C. (2020). The transcription factor Hypermethylated in Cancer 1 (Hic1) regulates neural crest migration via interaction with Wnt signaling. *Dev. Biol.* 463 (2), 169–181. doi:10.1016/j.ydbio.2020.05.012

Robasky, K., and Bulyk, M. L. (2011). UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 39, D124–D128. doi:10.1093/nar/gkq992

Spivak, A. T., and Stormo, G. D. (2012). ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic Acids Res.* 40 (D1), D162–D168. doi:10.1093/nar/gkr1180

Sun, L., Li, H., Chen, J., Dehennaut, V., Zhao, Y., Yang, Y., et al. (2013). A SUMOylation-dependent pathway regulates SIRT1 transcription and lung cancer metastasis. *J. Natl. Cancer Inst.* 105 (12), 887–898. doi:10.1093/jnci/djt118

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 51 (D1), D638–D646. doi:10.1093/nar/gkac1000

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., et al. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* (39), e1869. doi:10.3791/1869

Wang, S., Hu, H., and Li, X. (2022). A systematic study of motif pairs that may facilitate enhancer-promoter interactions. *J. Integr. Bioinform.* 19 (1), 20210038. doi:10.1515/jib-2021-0038

Wang, Y., Li, Y., Wang, C., Lio, C.-W. J., Ma, Q., and Liu, B. (2024). CEMIG: prediction of the cis-regulatory motif using the de Bruijn graph from ATAC-seq. *Brief. Bioinform* 25, bbad505. doi:10.1093/bib/bbad505

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158 (6), 1431–1443. doi:10.1016/j.cell.2014.08.009

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., et al. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* 28 (1), 316–319. doi:10.1093/nar/28.1.316

Wong, K. C. (2017). MotifHyades: expectation maximization for *de novo* DNA motif pair discovery on paired sequences. *Bioinformatics* 33 (19), 3028–3035. doi:10.1093/bioinformatics/btx381

Wong, K. C., Chan, T. M., Peng, C., Li, Y., and Zhang, Z. (2013). DNA motif elucidation using belief propagation. *Nucleic Acids Res.* 41 (16), e153. doi:10.1093/nar/gkt574

Wong, K. C., Li, Y., and Peng, C. (2016). Identification of coupling DNA motif pairs on long-range chromatin interactions in human K562 cells. *Bioinformatics* 32 (3), 321–324. doi:10.1093/bioinformatics/btv555

Yang, J., Ma, A., Hoppe, A. D., Wang, C., Liu, B., Ma, Q., et al. (2019). Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic Acids Res.* 47, 7809–7824. doi:10.1093/nar/gkz672

Yu, X., Lin, J., Zack, D. J., and Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.* 34 (17), 4925–4936. doi:10.1093/nar/gkl595

Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 July 2024
DOI 10.3389/fgene.2024.1405032

Check for updates

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Huiyan Sun,
Jilin University, China
Fa Zhang,
Beijing Institute of Technology, China
Shiwei Sun,
Institute of Computing Technology (CAS),
China
Chen Cao,
Nanjing Medical University, China

*CORRESPONDENCE
Zhenyu Cui,
✉ skd991578@sdust.edu.cn

RECEIVED 22 March 2024
ACCEPTED 20 June 2024
PUBLISHED 10 July 2024

CITATION
Su L, Yan Y, Ma B, Zhao S and Cui Z (2024), GIHP: Graph convolutional neural network based interpretable pan-specific HLA-peptide binding affinity prediction.
*Front. Genet.* 15:1405032.
doi: 10.3389/fgene.2024.1405032

COPYRIGHT
© 2024 Su, Yan, Ma, Zhao and Cui. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# GIHP: Graph convolutional neural network based interpretable pan-specific HLA-peptide binding affinity prediction

Lingtao Su[1], Yan Yan[2], Bo Ma[3], Shiwei Zhao[1] and Zhenyu Cui[1]*

[1]Shandong University of Science and Technology, Qingdao, China, [2]Shandong Guohe Industrial Technology Research Institute Co. Ltd., Jinan, China, [3]Qingdao UNIC Information Technology Co. Ltd., Qingdao, China

Accurately predicting the binding affinities between Human Leukocyte Antigen (HLA) molecules and peptides is a crucial step in understanding the adaptive immune response. This knowledge can have important implications for the development of effective vaccines and the design of targeted immunotherapies. Existing sequence-based methods are insufficient to capture the structure information. Besides, the current methods lack model interpretability, which hinder revealing the key binding amino acids between the two molecules. To address these limitations, we proposed an interpretable graph convolutional neural network (GCNN) based prediction method named GIHP. Considering the size differences between HLA and short peptides, GIHP represent HLA structure as amino acid-level graph while represent peptide SMILE string as atom-level graph. For interpretation, we design a novel visual explanation method, gradient weighted activation mapping (Grad-WAM), for identifying key binding residues. GIHP achieved better prediction accuracy than state-of-the-art methods across various datasets. According to current research findings, key HLA-peptide binding residues mutations directly impact immunotherapy efficacy. Therefore, we verified those highlighted key residues to see whether they can significantly distinguish immunotherapy patient groups. We have verified that the identified functional residues can successfully separate patient survival groups across breast, bladder, and pan-cancer datasets. Results demonstrate that GIHP improves the accuracy and interpretation capabilities of HLA-peptide prediction, and the findings of this study can be used to guide personalized cancer immunotherapy treatment. Codes and datasets are publicly accessible at: https://github.com/sdustSu/GIHP.

KEYWORDS

**HLA-peptide binding, model interpretation, GCNN, immunotherapy, affinity prediction**

## 1 Introduction

HLA also known as MHC (major histocompatibility complex) molecules, are responsible for presenting peptides derived from intracellular or extracellular proteins to T cells. It is a crucial step in understanding and predicting immune responses, such as antigen presentation and T-cell activation (Kallingal et al., 2023). HLA molecules are classified into two major classes: class I and class II. Each class has different subtypes, and their binding abilities vary depending on the specific HLA subtype. For HLA class I, the open binding groove close to both ends restrict the size of the bounded peptides between

8–12 residues, whereas HLA class II incorporates peptides of length 13–25 residues (Wang and Claesson, 2014). As a results, existing methods can be classified into allele-specific and pan-specific methods. Allele-specific methods focus on predicting the binding affinity between a specific HLA allele. Pan-specific methods aim to predict HLA-peptide binding in a more general way, without the need for allele-specific training data. (Gizinski et al., 2024).

Allele-specific methods train separate models for each MHC allele and make predictions for individual alleles. NetMHC (Lundegaard et al., 2008) is a widely used allele-specific method, which utilize machine learning algorithm to learn the relationship between peptide sequences and their binding affinities to specific MHC alleles. NetMHC 4.0 (Andreatta and Nielsen, 2016) is also a sequence-based allele-specific method, which uses both BLOSUM62 and sparse encoding schemes to encode the peptide sequences into nine amino acid-binding cores. In comparison with the HLA (around 360aa in length), peptides length are much shorter, and such methods must take insertion methods to reconcile or extend the original sequence. In addition, deep learning-based methods have also been developed for MHC-peptide binding prediction. DeepMHCII (You et al., 2022), which utilizes deep convolutional neural networks (CNNs) to capture complex sequence patterns and interactions between peptide and MHC class II molecules. It takes the peptide and MHC protein sequences as input and uses multiple layers of convolutional filters to extract features from the sequences. These filters scan the input sequences at different lengths, capturing both local and global patterns. The extracted features are then fed into fully connected layers to make predictions of the binding affinity. MHCAttnNet (Venkatesh et al., 2020) utilizes a combination of bidirectional long short-term memory (Bi-LSTM) and attention mechanisms to capture important features and dependencies in MHC- peptide interactions. The Bi-LSTM processes the sequences in both forward and backward directions, capturing the dependencies and context in the data. The attention mechanism allows the model weight different parts of the input sequences based on their relative importance. This enables the model to focus on the most relevant regions of the peptide and MHC sequences during the prediction process. SMM-align (Nielsen et al., 2007) utilizes structural and sequence-based features to predict binding affinities for MHC class I alleles. It employs a PSSM alignment algorithm to align target peptide sequences with known binders and derive binding predictions. MHC-NP (Giguere et al., 2013) also incorporate structure with sequence-based features and employs a random forest regression model to make predictions. Allele-specific methods are particularly useful when the focus is on specific alleles of interest, allowing for more accurate predictions tailored to those specific alleles. However, developing and maintaining separate models for each allele requires a significant amount of experimental binding data and computational resources.

On the other hand, pan-specific methods have the advantage of predicting binding affinities not only for alleles present in the training data but also for new, unseen alleles. NetMHCpan and NetMHCIIpan (Reynisson et al., 2020) are widely used pan-specific methods. They take sequence feature as input, utilizes artificial neural networks (ANNs) to learn the relationship between peptide sequences and their binding affinities to MHCs. They consider various sequence-based features, including amino acid composition, physicochemical properties, and binding motifs. In comparison with these two methods, another pan-specific method MHCflurry (O'Donnell et al., 2018; O'Donnell et al., 2020) integrates additional information, such as peptide processing predictions and binding affinity measurements from mass spectrometry-based experiments, to enhance its predictions. Some sequence-based methods, such as BERTMHC (Cheng et al., 2021), leverage the power of the BERT language model to improve their performance. The BERT language model is pre-trained on a vast corpus of text data, which enables it to capture intricate patterns and dependencies within input sequences effectively. One of the advantages of using BERT for encoding peptide sequences is its ability to capture long-range dependencies and contextual information. This is particularly important in MHC binding prediction, where specific amino acid positions within a peptide can significantly affect the binding affinity. Because structure determines the function of proteins, therefore, some methods also incorporate structure information into their predictions. MixMHCpred-2.0.1 (Gfeller et al., 2018) employs a deep learning architecture capable of learning complex patterns and relationships between peptide sequences and MHC binding affinities. The model is trained on a diverse set of MHC alleles and covers a wide range of peptide lengths. This allows it to make accurate predictions for a broad range of MHC-peptide combinations. NetMHCpan-4.0 (Jurtz et al., 2017) utilizes a combination of structural and sequence-based features. It incorporates information from MHC-peptide complex structures and uses a machine learning approach to make pan-specific predictions. RPEMHC (Wang et al., 2024) is a deep learning approach that aims to improve the prediction of MHC-peptide binding affinity by utilizing a residue-residue pair encoding scheme. In RPEMHC, the peptide sequence and MHC binding groove are encoded as one-hot vectors, representing each amino acid residue and its position. AutoDock is a widely used molecular docking software that can be employed for MHC-peptide binding prediction. It uses a Lamarckian genetic algorithm to explore the conformational space and predict the binding modes and affinities of peptides within the MHC binding groove. By modelling the docking between the HLA protein and peptide ligands these methods have achieved accurate binding prediction performance. However, docking methods rely on sampling different conformations of the peptide and MHC molecule to find the best binding pose. However, the conformational space of peptides and MHC molecules can be vast, and exhaustively sampling all possible conformations is computationally infeasible.

In fact, no matter allele-specific or pan-specific methods, they all can be broadly categorized into two main categories: sequence-based and structure-based methods. Sequence-based methods utilize machine learning techniques to capture the sequence motifs and physicochemical properties important for HLA-peptide binding. These methods employ various algorithms, such as support vector machines (SVMs), random forests, or ANNs, to learn the relationships between peptide sequences and binding affinities from large datasets. Sequence-based methods have the advantage of being computationally efficient and applicable to a wide range of HLA alleles and peptides. Structure-based methods leverage the three-dimensional structures of HLA molecules and peptides to predict binding affinities. Molecular docking algorithms, such as AutoDock, are commonly used to explore the conformational space and calculate binding energies. These methods require knowledge of the 3D structures of the HLA molecule and peptide, limiting their applicability to cases where experimental structures are unavailable. Recent advancements in deep learning, such as CNNs and recurrent neural networks (RNNs), have shown promise in HLA-peptide binding affinity prediction. Deep learning-based methods can effectively capture complex sequence patterns and structural features, leading to improved prediction accuracy (Wang et al., 2023). These models often incorporate encoding schemes to represent peptide sequences or structural features and are trained on large datasets to learn the relationships between sequences and binding affinities. Despite notable progress, HLA-peptide binding affinity prediction still faces challenges and have some limitations. First, deep learning models are often considered as black boxes, meaning they lack interpretability. It can be challenging to understand the specific features or patterns that contribute to the model's predictions. Interpretability is crucial in immunology research to gain insights into the molecular mechanisms underlying MHC-peptide interactions and to guide experimental studies; Second, existing methods often rely on sequence-based encoding schemes due to the limited availability of experimentally determined 3D structures for HLA-peptide complexes. While sequence information is informative, the exclusion of structural details may limit the accuracy and coverage of predictions, particularly for cases where structural features play a crucial role. Even some tools consider structure information, they seldom consider the structure features at the amino acids level. Besides, the length difference between the peptides that HLA can bind (typically around 8–15 amino acids) and the length of HLA molecules (which can be over 360 amino acids) poses a challenge in HLA-peptide binding affinity prediction. Furthermore, unlike HLAs, peptides are too short to form stable structures. All these drawbacks are not well solved by existing methods.

Considering all these limitations, we proposed GIHP, which is an interpretable GCNN-based algorithm for the prediction of peptides binding to pan HLA molecules. By representing peptide SMILE strings (Quiros et al., 2018; Meng et al., 2024) and HLA structures as attributed graphs, GCNNs can effectively model the pairwise interactions between amino acids and capture both local and global structural features. Furthermore, GIHP has a novel visual explanation method called Grad-WAM for HLA-peptide binding affinity prediction and interpretation. By analyzing the learned representations and interactions within the graph structure, the Grad-WAM technique can identify the key residues that contribute most significantly to the HLA-peptide binding process. Comprehensive comparative evaluation results demonstrate that the GIHP achieves good performance across diverse benchmark datasets. By applying the GIHP framework to several cancer immunotherapy datasets, we have identified numerous promising biomarkers that can effectively distinguish patients with and without treatment response. Moving forward, the insights gained from the GIHP analysis can be leveraged to guide the development of more personalized cancer immunotherapy strategies.

# 2 Materials and methods

## 2.1 Data collection and processing

We collected human HLA-peptide interaction datasets from published papers or publicly available databases. (Table 1).

**TABLE 1 Summary of the collected datasets after preprocessing.**

| Name | HLAs | Peptides | HLA-peptide interactions |
|---|---|---|---|
| Wang-2008 | 26 | 4,421 | 24,295 |
| Wang-2010 | 14 | 3,902 | 9,478 |
| Kim-2014 | 183 | 28,428 | 268,189 |
| Jurtz-2017 | 124 | 3,307,868 | 3,618,591 |
| Jensen-2018 | 72 | 15,965 | 131,008 |
| Zhao-2018 | 53 | 2,168 | 21,092 |
| Reynisson-2020 | 161 | 4,523,148 | 4,795,633 |

Wang-2008 Dataset (Wang et al., 2008): Experimentally measured peptide binding affinities for HLA class II molecules. The processed data set had 24,295 interaction entries in total with ligand length ranging from 16 to 37 and have 26 unique HLA molecules. HLA DP and DQ molecules are covered.

Wang-2010 Dataset (Wang et al., 2010): Experimentally measured peptide binding affinities for MHC class II molecules. After preprocessing, the dataset contains 9,478 measured affinities and covers 14 MHC class II alleles with peptides length ranging from 9 to 37.

Kim-2014 Dataset (Kim et al., 2014): this dataset was obtained from the Immune Epitope Database (IEDB) (Vita et al., 2019), including binding affinity data compiled in 2009 (BD 2009), 2013 (BD 2013) and also include a blind datasets. Blind datasets refer to data resulting after subtracting BD2009 from BD 2013. For all these three datasets, only human datasets were kept for training. After preprocessing the dataset contains 268,189 interactions in total, with peptides length ranging from 8 to 30.

Jurtz-2017 Dataset (Jurtz et al., 2017): this dataset is originally designed for training of NetMHCPan-4.0. The final processed dataset has 3,618,591 entries in total with ligand length ranging from 8 to 18.

Jensen-2018 Dataset (Jensen et al., 2018): this dataset is used for training of NetMHCIIpan-3.2 (Karosiene et al., 2013), which contains HLA class II binding affinities retrieved from the IEDB in 2016. The 2016 data set contains 131,008 data points, covering 36 HLA-DR, 27 HLA-DQ, 9 HLA-DP molecules and 15,965 unique peptides. The peptides length range from 9 to 33.

Zhao-2018 Dataset (Zhao and Sher, 2018): this dataset is compiled for training IEDB tools as well as the MHCflurry (O'Donnell et al., 2018). The dataset contains 21,092 binding relations, covering 18 HLA-DR, 19 HLA-DQ, 16 HLA-DP molecules and 2,168 unique peptides. The peptides length is 15.

Reynisson-2020 dataset (Reynisson et al., 2020): this dataset is originally collected for training NetMHCpan-4.1 and NetMHCIIpan-4.0 methods. The dataset covering 161 distinct HLA class I molecules, 4,523,148 distinct peptides, with peptides length ranging from 8 to 15.

For all the collected training datasets, only binding affinity values in IC50nM format are kept, which are log-transformed to fall in the range between 0 and 1 by applying 1−log (IC50 nM)/log (50k) as explained by Nielsen et al. (2003). When classifying the

**TABLE 2** Immunotherapy related dataset and three cancer datasets.

| Name | Type | Patients | SNP mutations |
|---|---|---|---|
| Samstein-2019 | Pan-cancer | 1,662 | 14,876 |
| Miao-2018 | Pan-cancer | 249 | 102,207 |
| Razavi-2018 | Breast cancer | 1,756 | 7,420 |
| Clinton-2022 | bladder Cancers | 1,245 | 24,277 |
| Aaltonen-2020 | Pan-cancer | 2,583 | 347,994 |

peptides into binders or non-binders a threshold of 500 nM is used. This means that peptides with log50k transformed binding affinity values greater than 0.426 are classified as binders. We consolidated all the collected datasets, removing any duplicate entries, to arrive at a final integrated dataset comprising 160,253 unique HLA-peptide interactions, covering 223 distinct HLA alleles and 35,481 peptide sequences. To further verify the generality of our method, we collected protein-peptide binding data from pepBDB (Wen et al., 2019) database, after deleting peptides short than 8aa, we got 12,655 interactions between 11,055 proteins and 7,811 peptides. Because our method takes HLA and protein structure as input, all the structure data are downloaded from the PDB (Berman et al., 2000) and AlphaFold database (Varadi et al., 2022) and some are predicted by alphafold2 (Jumper et al., 2021) and Rosettafold (Baek et al., 2021). Only high-resolution experimental structures (e.g., X-ray crystallography or cryo-EM data with resolution better than 3.0 Å) were included. All structural models, whether experimental or predicted, were subjected to validation using atomic contact evaluation, and overall model quality assessment. Only structures that passed these validation checks were retained for further analyses.

To evaluate whether the key binding residues identified by our method can effectively differentiate patients who benefit from immunotherapy, we collected relevant breast, bladder, and pan-cancer treatment datasets from the cBioPortal resource (Cerami et al., 2012), as shown in Table 2. Key binding residues mutation could lead to binding affinity change between HLA and peptides. Binding affinity change has been demonstrated as a biomarker of immunotherapy efficiency (Kim et al., 2020; Seidel et al., 2021; Murata et al., 2022). For each patient, only SNP mutations are kept, if the SNP locates on the key binding site of HLA or peptide, then we separate them in one group, otherwise in the other group. Then we conduct survival analysis for the two groups.

Samstein-2019 dataset (Samstein et al., 2019): The cohort consisted of 1,662 patients, received at least one dose of immune checkpoint inhibitor (ICI) therapy. The cohort encompassed a variety of cancer types with an adequate number of patients for analysis. In detail, 146 patients received anti-CTLA4, 1,447 received anti-PD1 or PD-L1, and 189 received both. This is a pan-cancer dataset, including 350 cases of non-small cell lung cancer (NSCLC), 321 cases of melanoma, 151 cases of renal cell carcinoma (RCC), 214 cases of bladder cancer, and 138 cases of head and neck squamous cell cancer.

Miao-2018 dataset (Miao et al., 2018): this dataset consists of 249 patient tumors from six different cancer types: melanoma (*N* = 151), non-small cell lung cancer (*N* = 57), bladder cancer (*N* = 27), head and neck squamous cell carcinoma (*N* = 12), anal cancer (*N* = 1), and sarcoma (*N* = 1). These patients were treated with anti-PD-1 therapy (*N* = 74), anti-PD-L1 therapy (*N* = 20), anti-CTLA-4 therapy (*N* = 145), or a combination of anti-CTLA-4 and anti-PD-1/L1 therapies (*N* = 10). A small proportion of patients (*N* = 7) received a combination of anti-PD-1, anti-PD-L1, or anti-CTLA-4 therapy with another immunotherapy, targeted therapy, or cytotoxic chemotherapy.

Razavi-2018 dataset (Razavi et al., 2018): This dataset is downloaded from cBioPortal: https://cbioportal-datahub.s3.amazonaws.com/breast_msk_2018.tar.gz.

Clinton-2022 dataset (Clinton et al., 2022): This dataset is downloaded from cBioPortal: https://cbioportal-datahub.s3.amazonaws.com/paired_bladder_2022.tar.gz.

Aaltonen-2020 dataset (Consortium et al., 2020): This dataset is downloaded from cBioPortal: https://cbioportal-datahub.s3.amazonaws.com/pancan_pcawg_2020.tar.gz.

## 2.2 Methods

The overall framework of GIHP is illustrated in Figure 1. GIHP takes HLA structure and peptide SMILE string as input. In the input representation module, HLA is represented as an attributed residue-level graph, while the peptide is represented as an attributed atom-level graph. Then a multi-layer GCNNs is used to learn the high-level features, and the learned features are contacted and fed into the MLP layer for final binding affinity prediction. To enhance the results interpretability, we introduced a novel visual interpretation method called Grad-WAM. Grad-WAM leverages gradient information from the last GCN layer to assess the significance of each neuron in determining affinity.

### 2.2.1 Input representation

Graph-based protein structure representation has inherent advantages over traditional sequence-based approaches in capturing true binding events. For each HLA molecular, we take both structure and sequence information into consideration. Given one of our key objectives is to identify the critical binding amino acid residues, we have represented the HLA proteins as residue-level relational graphs $G_H = (\nu, \varepsilon)$, where $\nu$ is the set of amino acids, $\varepsilon$ is the set of edges. As shown in Table 3, we describe the node attributes by integrating sequence and structural property, including amino acid type, chemical properties, charges, *etc.*, while the edge attributes encompass connection types, distances, and structural information. We consider four types of bond edges including Peptide Bonds, Hydrogen Bonds, Ionic Bonds and Disulfide Bridges.

Considering that the length of peptides binding to MHC class II is between 13–25 residues, and the length is around nine for peptides binding to MHC class I. Therefore, the peptide length is relatively short compared to HLAs (over 360aa). In this study, we represent peptides as SMILES-like sequences and then transform them into graphs using a molecular graph representation method inspired by RDKit (https://www.rdkit.org). The attributes of each node $\nu_i$ are shown in Table 4. $e_{ij} \in \varepsilon$ is covalent bonds between the *ith* and the *jth* atoms. The edge attributes depending on the electrons shared between atoms, resulting in single, double, or triple bonds, respectively.

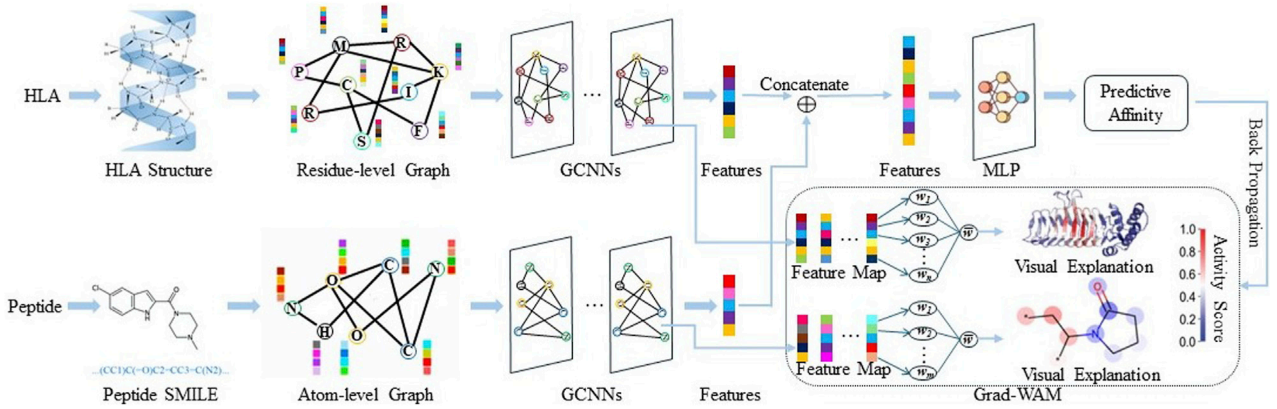


FIGURE 1
The overall framework of GIHP.

TABLE 3 The node features of HLA graph.

| Name | Description | Dim |
|---|---|---|
| Residue type | We utilize Blosum62, 20 types of amino acids plus 1 unknown | 21 |
| Structure mapping | Included α-helix (H), residue in isolated β-bridge (B), extended strand, participates in β ladder (E), hydrogen bonded turn (T),$3_{10}$ helix (G), π-helix (I), bend (S) and coil (C) | 8 |
| AA position | the position of α-carbon in each residue to record their 3D position | 3 |
| Hydrogen donor or acceptor | Donor: R, K, W. Acceptor: D, E. Donor and acceptor: N, Q, H, S, T, Y | 4 |
| Physicochemical properties | We utilize a set of 7 physicochemical properties for amino acid types (AAPHY7). These features include steric parameters, hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability | 7 |

TABLE 4 Node features of peptide graph.

| Name | Description | Dim |
|---|---|---|
| Atom type | [H, C, N, O, F, Cl, S, Br, I] (one-hot) | 9 |
| Atomic Num | The atomic number (integer) | 1 |
| Acceptor | Accepts electrons [0/1] (binary) | 1 |
| Donor | Donates electrons [0/1] (binary) | 1 |
| Aromatic | In an aromatic system [0/1] (binary) | 1 |
| Hybridization | [sp, sp2, sp3] (one hot) | 3 |
| Hydrogens | Number of connected hydrogens (integer) | 1 |
| Formal charge | Formal charge of the atom (integer) | 1 |
| Explicit valence | Explicit valence of the atom (integer) | 1 |
| Implicit valence | Implicit valence of the atom (integer) | 1 |
| Explicit Hs | Number of implicit Hs the atom is bound to (integer) | 1 |
| Radical electrons | Number of radical electrons for the atom (integer) | 1 |

### 2.2.2 Graph convolutional neural network module

Let $A$ be the adjacency matrix, and $X$ be the feature matrix of the given graph. Each GCN layer takes $A$ and node embeddings as input and outputs the final embeddings. As shown in Eqs 1, 2.

$$H^{(l+1)} = GCN\left(H^{(l)}, A\right) \tag{1}$$

$$H^{(l+1)} = ReLU\left(\hat{D}^{-0.5}\hat{A}\hat{D}^{-0.5}H^{(l)}W^{(l+1)}\right) \tag{2}$$

Where, $H$ is the embeddings, and $H^{(0)} = X$, $W^{(l+1)}$ are trainable weight matrix, $\hat{D}$ is the diagonal node degree matrix of $A$.

After obtaining the vector representations of HLA and peptide, they are concatenated and fed into a Multi-Layer Perceptron (MLP) to predict the binding affinity score. The MLP consists of three linear transformation layers, each followed by a Rectified Linear Unit (ReLU) activation function and a dropout layer with a dropout rate of 0.1, as in (Öztürk et al., 2019). The Mean Squared Error (MSE) is employed as the loss function to measure the discrepancy between predicted and actual affinity scores. MSE is defined in Eq. 3.

$$MSE = \frac{1}{n}\sum\nolimits_{i=1}^{n} (P_i - Y_i)^2 \tag{3}$$

Where, $n$ is the sample size, $P_i$ and $Y_i$ are the predictive and true values of the $ith$ interaction pair, respectively.

### 2.2.3 Gradient-weighted activation mapping

While Grad-CAM has been successfully applied to various computer vision tasks, it is not directly applicable to graph-structured data. Therefore, in this paper we proposed a novel results interpretation methods called Grad-WAM, which can be used for identifying key binding related residues. Grad-WAM measure the contribution of each residue for the decision of binding by taking use of the gradient information in the last GCN layer. Grad-WAM utilizes a weighted combination of the positive partial derivatives of the feature maps with respect to the interaction values to generate the corresponding visual explanations. Considering the contribution of each residue is not equal, different from the explanation method proposed in MGraphDTA (Yang et al., 2022), we introduce an additional weight $\omega$ (Eq. 4) gradient values.

$$\omega = \sum\nolimits_i [\alpha_i] \cdot ReLU\left(\frac{\partial P}{\partial T_i}\right), \forall \{i \mid i\epsilon T\} \tag{4}$$

Where, $ReLU$ is the activation function, $P$ is the predictive value as in Eq. 5. $T_i$ is the feature value of the $ith$ node on the feature map $T$ of the last GCN layer. $\alpha_i$ is the gradient value of the $ith$ node defined in Eq. 6. $\frac{\partial P}{\partial T_i}$ is the partial derivative as in Eq. 7.

$$P = \sum\nolimits_i \alpha_i \cdot ReLU\left(\frac{\partial P}{\partial T_i}\right) \cdot T_i \tag{5}$$

$$\alpha_i = \frac{\frac{\partial P}{\partial T_i}}{\frac{\partial P}{\partial T_i} + T_i \cdot \frac{\partial^2 P}{(\partial T_i)^2}} \tag{6}$$

$$\frac{\partial P}{\partial T_i} = \alpha_i \cdot \frac{\partial P}{\partial T_i} + T_i \cdot \alpha_i \cdot \frac{\partial^2 P}{(\partial T_i)^2} \tag{7}$$

In this way, the contribution of residues to the prediction of binding affinity is calculated. For visual explanation, residues are display utilizes colors, ranging from blue to red. A higher gradient value corresponds to a redder color, indicating the key role of that amino acid in the interaction.

# 3 Results

## 3.1 Performance comparisons with other methods

Four widely used performance metrics were employed to measure methods' performance. Including accuracy (*Acc*), Matthews Correlation Coefficient (*MCC*), sensitivity (*Sn*), and the specificity (*Sp*). The definitions of these four metrics are as follows: Eqs 8–11.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Sn = \frac{TP}{TP + FN} \tag{9}$$

$$Sp = \frac{TN}{TN + FP} \tag{10}$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

Where, *TP* is True Positives, *TN* is True Negatives, *FP* is False Positives, and *FN* is False Negatives. In addition, by comparing the predicted and true values, predictions were assessed to be true or false. The receiver operating characteristic curves (ROC) were generated for all the methods, and the performance of each algorithm to discriminate between binders and nonbinders was analyzed by calculating the area under the ROC curve (AUC) as an estimate of prediction performance.

We compare GIHP with state-of-the-art allele and pan-specific baselines including NetMHC-4.0 (Andreatta and Nielsen, 2016), NetMHCpan-4.0 (Jurtz et al., 2017), PickPocket-1.1 (Zhang et al., 2009), SMMPMBEC (Kim et al., 2009), MHCFlurry (O'Donnell et al., 2018), MixMHCpred-2.0 (Bassani-Sternberg et al., 2017) and NetMHCcons-1.1 (Karosiene et al., 2012). To eliminate the impact of data variations, all models were retrained and tested using our new collected and processed dataset. 10-fold cross-validation (CV) was applied. The data set is divided into 10 folds. During each iteration, one of the 10 partitions is designated as the validation dataset, while the remaining nine partitions are utilized to train the model. The final performance is determined by calculating the average performance across all 10 individual iterations. As shown in Figure 2, on average, GIHP outperform all the compared prediction methods. It is worth noting that not every method is suitable for every HLA and peptide length. To make the performance comparison fairer and more reasonable, we train allele-specific models with their required HLAs and peptide length, which included in our datasets.

To make comparisons more comparable and test methods performance on other protein-peptide binding datasets, a separate independent test is conducted using the data collected from pepBDB, which have no overlap with the above training data. This independent test data set serves as an unbiased validation source to assess the performance of different tools, which is relatively more objective, and can test models' generalization ability. 10-fold cross validation is applied, after each epoch average results are calculated. Results on the pepBDB independent test data is shown in Figure 3.

On average, GIHP achieved highest AUC value. In this independent test data, GIHP achieved the highest AUC of 0.88 and the highest Sp score of 0.98. In contrast, NetMHCPan-4.0 and Pickpocket-1.1 attained AUC values of 0.76 or lower, and Acc scores of 0.71 or lower when evaluated on this new dataset. Difference from the results on the above part, MHCflurry got AUC up to 0.8. Similar with our method, MHCflurry harness the power of deep learning and a comprehensive dataset to improve the prediction of HLA-peptide binding affinities. Our model outperforms both allele and pan-specific methods, demonstrate its ability to achieve higher prediction accuracy and robustness generality for all kinds of training data.

For evaluating the performance our method under different peptide length. We collected independent test set and external test set from TransPHLA, which can be downloaded from https://github.com/a96123155/TransPHLA-AOMP/tree/master/Dataset. In the collected datasets, 9-mer peptides comprising the largest proportion, while the number of 13-mer and 14-mer peptides is very small. Our model's performance on the independent test set and external test set for different peptide lengths are shown in Figures 4A, B respectively. As shown in Figure 4, our methods can achieve good performance on all kinds of peptide length.

## 3.2 Key binding residues on HLAs

The binding of peptides to HLA molecules occurs within specialized regions called binding pockets. HLA class I molecules have a peptide-binding groove formed by two alpha helices (α1 and

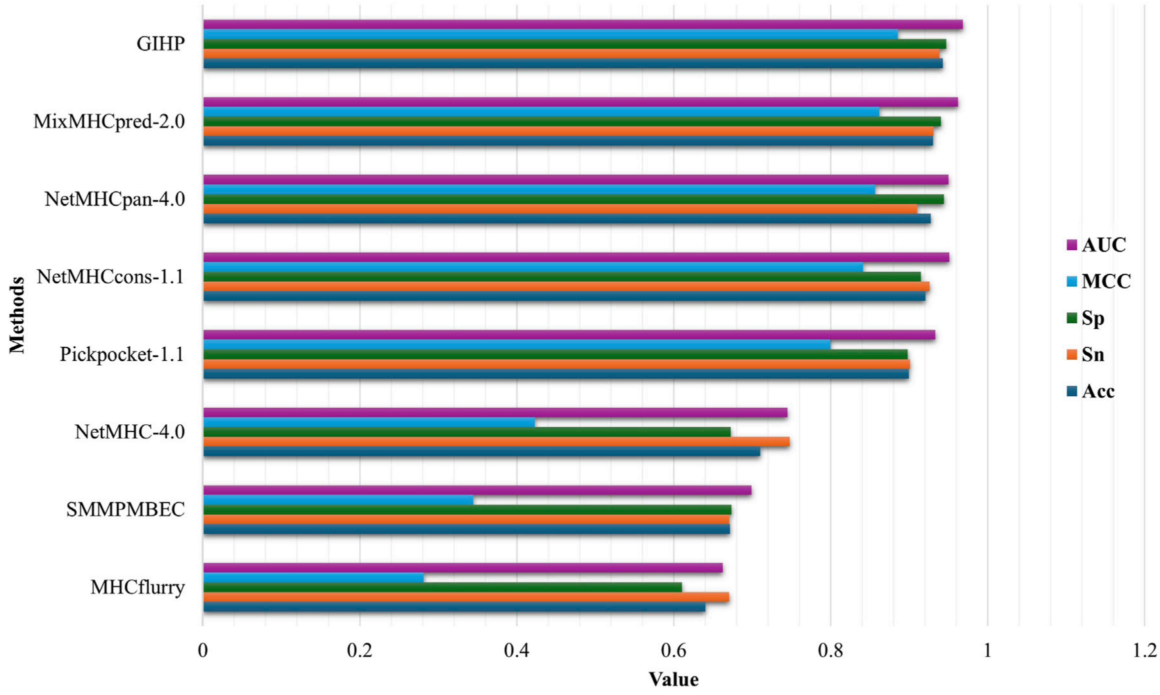


FIGURE 2
Performance comparison results.

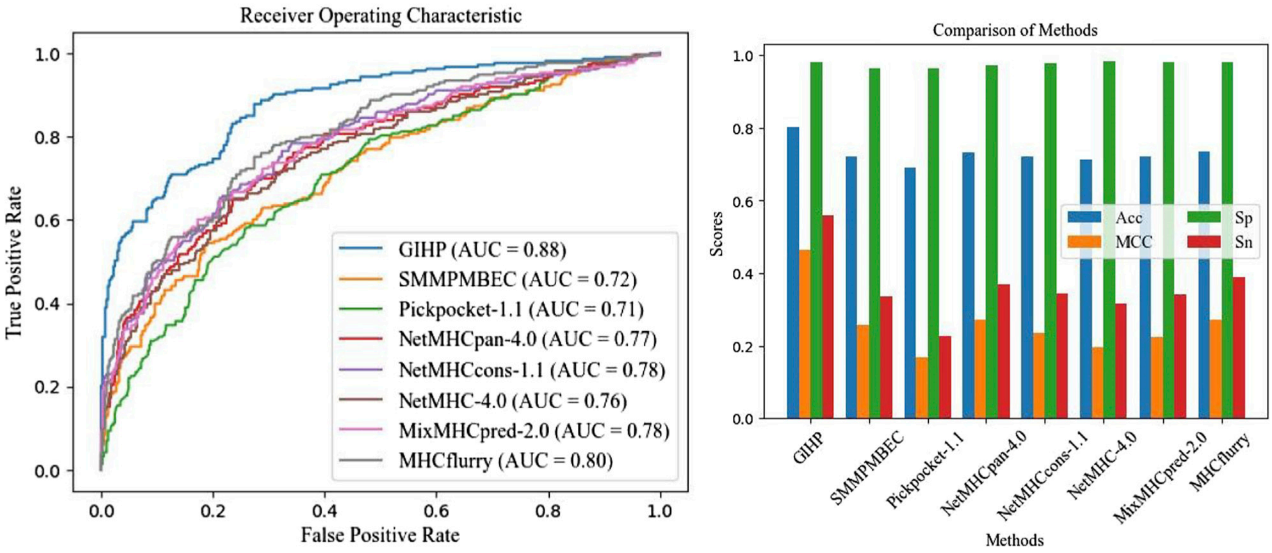


FIGURE 3
Independent test results on pepBDB datasets.

α2) and a beta sheet platform. Within this groove, there are seven pockets (numbered from A to F, shown in Figure 5A) that interact with specific amino acid residues of the bound peptide. HLA class II molecules are involved in presenting peptides derived from extracellular proteins to helper T cells. HLA class II binding pockets are formed by two chains: the alpha chain (α) and the beta chain (β). Each chain consists of two domains: the α1 and β1 domains form the peptide-binding groove, while the α2 and β2 domains provide structural support. The binding groove of HLA class II molecules is open at both ends, allowing longer peptides to bind compared to HLA class I molecules. The binding pockets in HLA class II molecules are referred to as P1, P4, P6, P7, P9 (Figure 5B). With our GIHP results interpret module, many key binding residues on both HLA class molecules and the corresponding peptides are identified. Although some residues with high activity scores locates outside of binding pockets, most

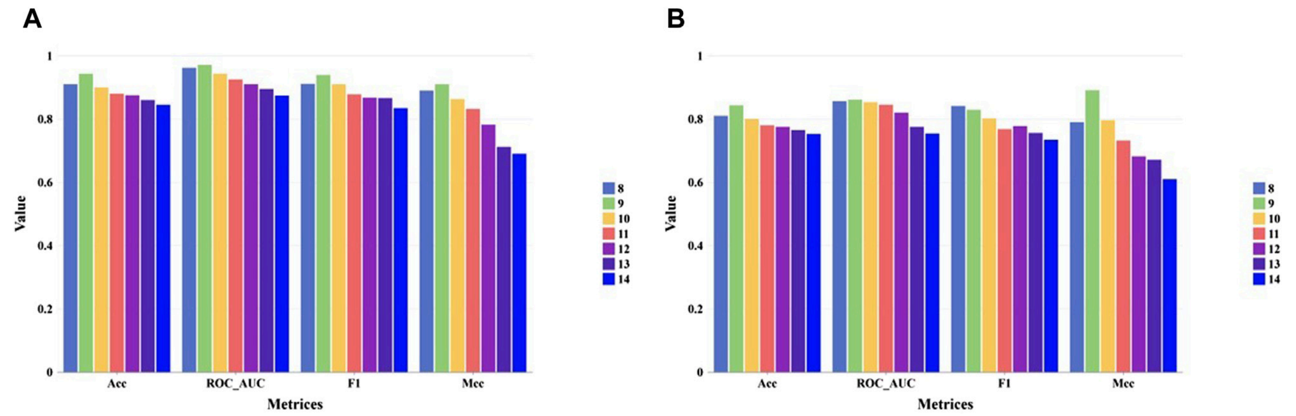


FIGURE 4
The performance of our model on the independent test set and external test set for the different peptide lengths. **(A)** Performance on the independent test set. **(B)** Performance on the external tet set.

of them locates on one of the binding pockets. As shown in Figures 5C,D, 45 residues with highest activity scores on HLAs are identified, among them 26 locates on HLA class I pockets, and 19 locates on HLA class II pockets.

Position 159 has the highest activity score on pockets A. Other positions including 59, 171, 167, seven and 66. According to current research, position seven is a pocket A's floor residue. This residue creates a hydrophobic environment within the pocket A and interact with the side chain of the anchor residue. Although residue on position 159 has no evidence of directly involved in peptide binding interactions, it has structural and functional implications for the overall stability and conformation of the pocket A region (Ma et al., 2020). It potentially contributes to the shape and electrostatic properties of the pocket, indirectly affecting the binding preferences and stability of the peptides presented by the HLA class I molecule. However, the specific role and impact of residue 159 on the pocket A's function vary among different HLA alleles and need further study for a comprehensive understanding. On pockets B, substitutions at positions 70 was found to yield a significantly distinct peptide-binding repertoire in HLA-A molecules when compared to HLA-B molecules. Positions 167 and position 67 on pocket B has been demonstrated as key peptide-binding residues. Besides, substitutions at positions 67 and nine exert a significant influence on the peptide-binding repertoire (van Deutekom and Keşmir, 2015). Position 97 has the highest activity score on pockets C. Position 97 is known to be a critical residue for peptide binding and presentation. This residue locates near the C-terminal anchor residue of the bound peptide and contributes to the formation of the peptide-binding groove. The amino acid at position 97 can significantly influence the peptide-binding specificity and affinity of the HLA molecule. Substitutions or variations at this position can alter the size, shape, or electrostatic properties of the pocket C, thereby affecting the recognition and binding of specific peptides. Several studies have investigated the impact of position 97 on peptide binding and immunological responses (Moutaftsi et al., 2006).

Considering the residues with high activity scores on HLA class II pockets, position nine is crucial for determining the peptide-binding specificity of the HLA class II molecule. The amino acid at position nine of the bound peptide interacts with residues in the P1 pocket, influencing the peptide-binding preferences. Position 86 plays a critical role in peptide binding and presentation (Brown et al., 1993). The amino acid at position 86 interacts with the peptide residue and contributes to the stability and specificity of the HLA-peptide class II complex (Stern et al., 1994). Among our identified important positions, positions 13 and 74 are critical for determining the peptide-binding specificity and stability of HLA class II molecules. The interactions between peptide residues and the residues in these pockets are essential for the recognition and presentation of antigenic peptides to CD4$^{+}$ T cells. Except these positions, we also prioritized many other residues, such as positions 63 and 57. These positions within the peptide-binding grooves of HLA class II molecules is crucial for understanding the molecular basis of antigen presentation and immune responses. Researchers can gain valuable information about the molecular interactions governing antigen presentation and T cell recognition. Furthermore, these results can help designing personalized immunotherapies (Boukouaci et al., 2024).

Figures 5E, F show the motif analysis results. In the two figures, the Y-axis describes the amount of information in bits. The X-axis shows the position in the alignment. At each position there is a stack of symbols representing the amino acid. Large symbols represent frequently observed amino acids, big stacks represent conserved positions and small stacks represents variable positions. Therefore, positions 2, 4 and nine have frequently observed amino acids in HLA class I and class II respectively.

## 3.3 Key binding residues on peptides and their corresponding genes

In this paper, we focus on finding immunotherapy efficiency related key residues and their corresponding genes. With the identified residue positions and the corresponding gene mutation, we try to verify whether they can be biomarkers to separate patients into different survival groups. We applied GIHP to immunotherapy related datasets (Samstein-2019 and Miao-2018 in Table 2). For each SNP mutation site, we extract the corresponding 9-mer peptide

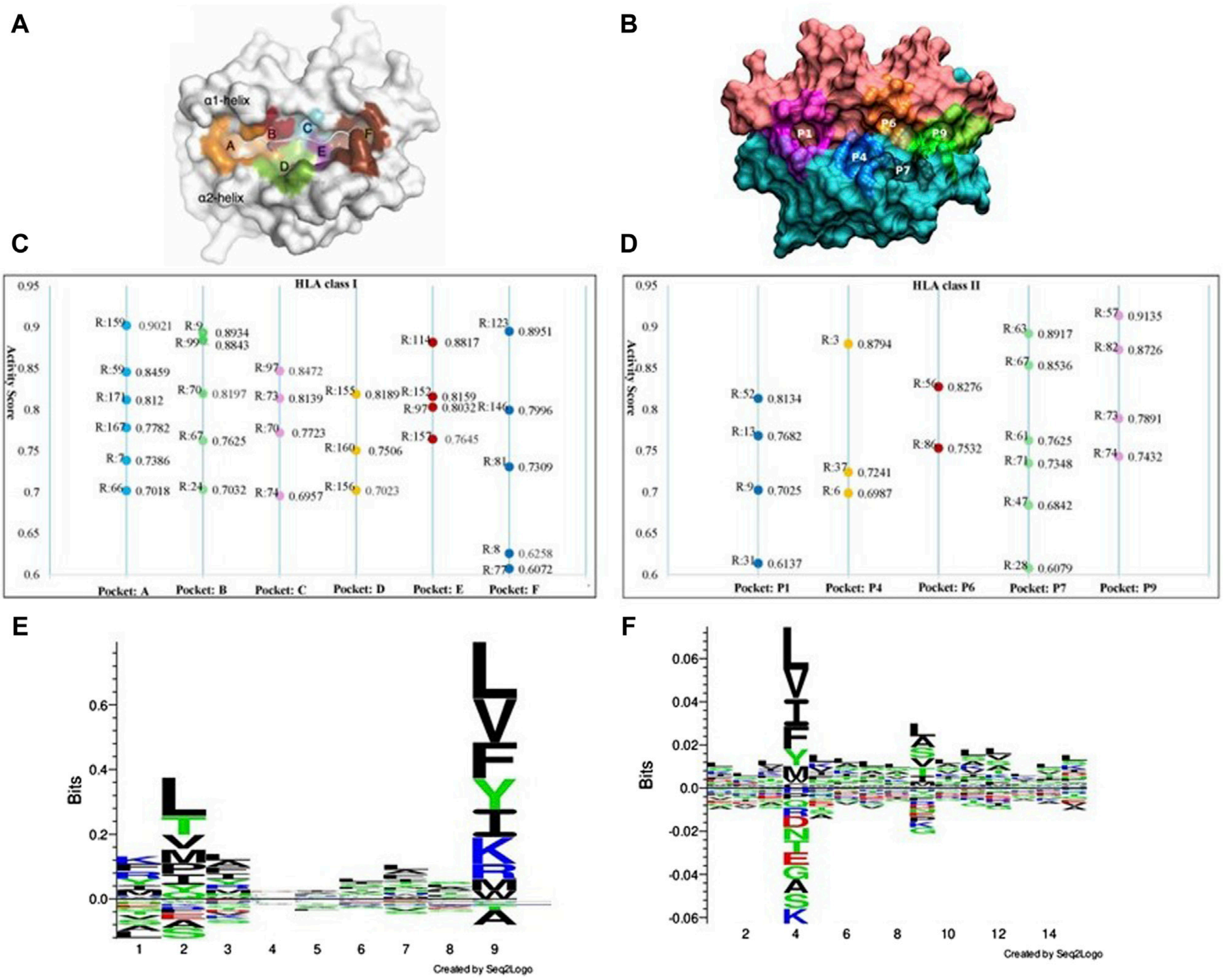


FIGURE 5

The key binding residues on HLA pockets and HLA binding peptides motif. **(A)** Binding pockets on HLA class I molecules. **(B)** Binding pockets on HLA class II molecules. **(C)** The identified key binding residue locations and activity scores on each pocket of HLA class I molecules, where R represent residue location analyzed HLA molecules. **(D)** The identified key binding residue locations and activity scores on each pocket of HLA class II molecules. **(E)** Distribution of preferred peptide residues of HLA class I molecules using Seq2logo2.0. **(F)** Distribution of preferred peptide residues of HLA class I molecules using Seq2logo2.0.

around it and predict the binding affinities with all the 223 HLAs. By paired *t*-test statistical comparing the binding affinity change before and after residue substitution, along with GIHP returned activity scores of each residue, significant key binding residues are identified. To get the functions of these mutation related genes, we conducted GO enrichment analysis by ShinyGO-0.80 (Ge et al., 2020). As shown in Figure 6, most of key residues locate on genes related to pathways in cancer and cancer related signaling pathways.

Since we interest in finding mutations related to immunotherapy response, therefore, we further analyzed key residues enriched in T cell receptor signaling pathway (Figure 6). The enriched genes include RHOA, HLA-B, HRAS, IL10, NRAS and KRAS. RHOA has been implicated in T cell activation and migration, which are critical for effective anti-tumor immune responses (Bros et al., 2019). Altered RHOA signaling could potentially impact T cell function and infiltration into the tumor microenvironment, influencing immunotherapy response. HLA-B plays a crucial role in immune recognition, as it presents peptide antigens derived from intracellular proteins to cytotoxic T cells. HRAS, NRAS, and KRAS are genes that belong to the RAS family of oncogenes. These genes encode proteins involved in intracellular signaling pathways regulating cell growth, survival, and proliferation. The presence of RAS mutations has been associated with poorer response rates to certain immunotherapies, including immune checkpoint inhibitors (East et al., 2022). IL10 can suppress the activity of cytotoxic T cells and natural killer (NK) cells, which are critical for tumor surveillance and elimination. High levels of IL10 in the tumor microenvironment have been associated with immunosuppression and reduced response to immunotherapy (Salkeni and Naing, 2023).

Next, we investigated the impact of biomarker gene mutations on patient survival outcomes using a cohort of individuals (Samstein-2019 dataset in Table 2) with immunotherapy treatment. The patients were categorized into two groups based

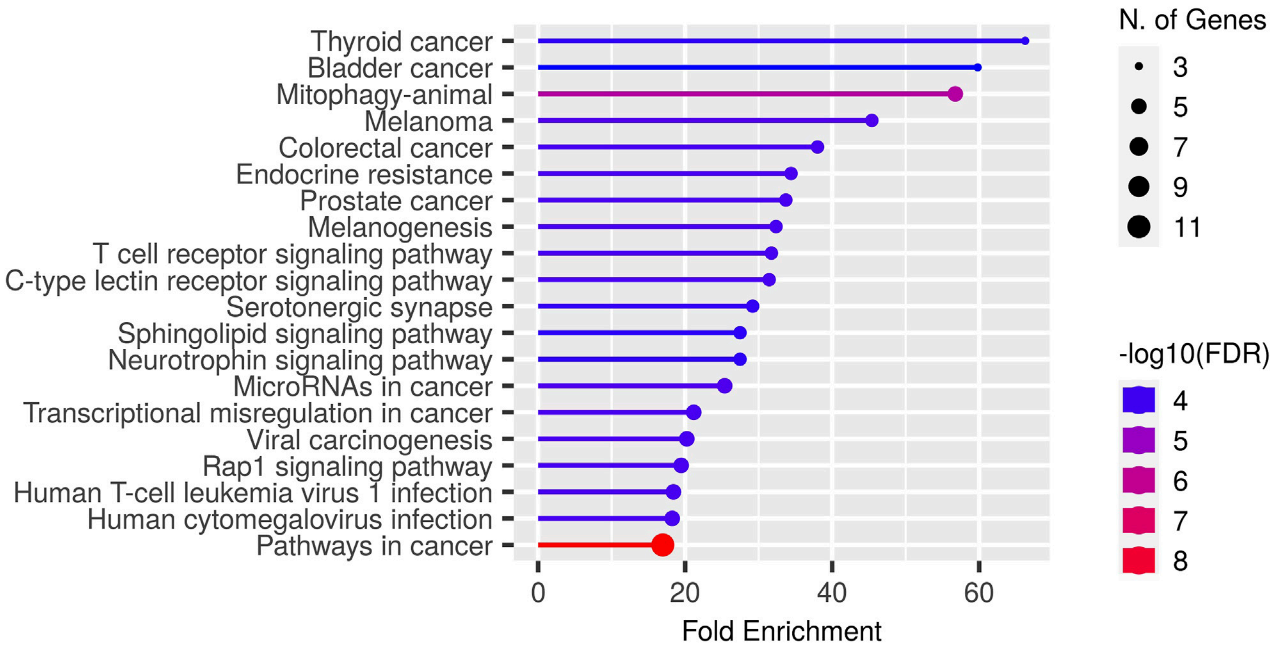


FIGURE 6
GO enrichment results of key residues related genes.

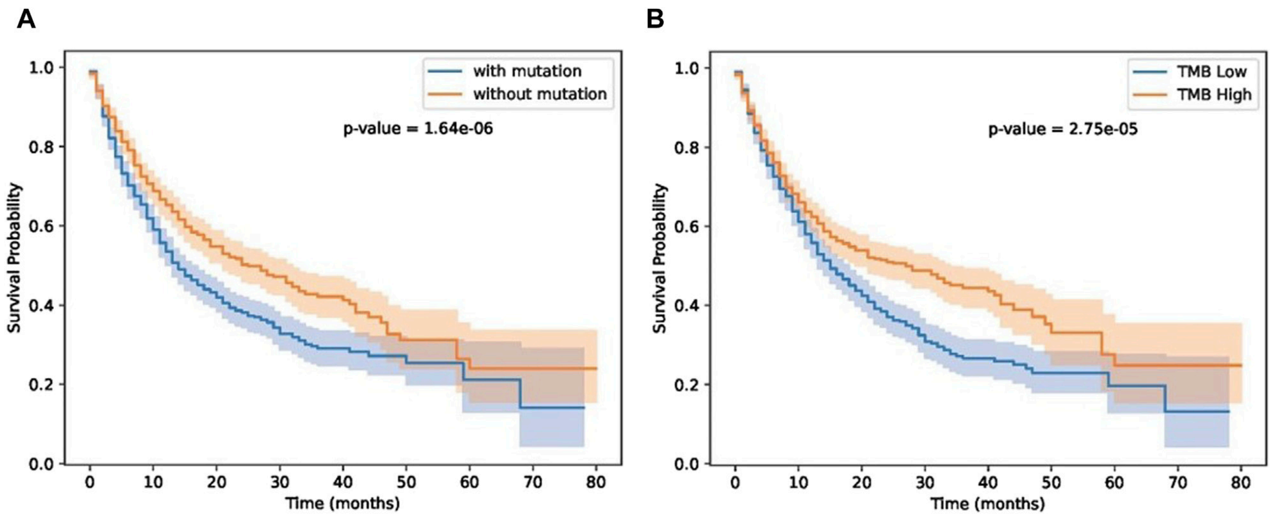


FIGURE 7
Results on immunotherapy data. **(A)** patient groups separated by GIHP identified biomarker mutations. **(B)** TMB separated patient groups.

on the presence or absence of the biomarker gene mutation. Kaplan-Meier survival curves were generated, and a log-rank test was performed to compare the survival between the two groups. The results revealed a significant difference in survival between the two groups, with patients harboring the biomarker gene mutation exhibiting a higher risk of adverse events compared to those without the mutation. These findings highlight the potential prognostic significance of the biomarker gene mutation and underscore its relevance in patient stratification and personalized treatment approaches. Furthermore, we compared our results with TMB score provided in (Samstein et al., 2019). As shown in Figure 7, patients with biomarker mutations tend to have poor survival status.

As shown in Figure 7, our methods can separate patients more significantly. Although TMB can separate patients, TMB is an overall measure, its hard to know which gene mutations play key roles in differentiating patients' response. Our methods not only can separate patients significantly, moreover, we also know which residue substitutions play key roles. To further test the

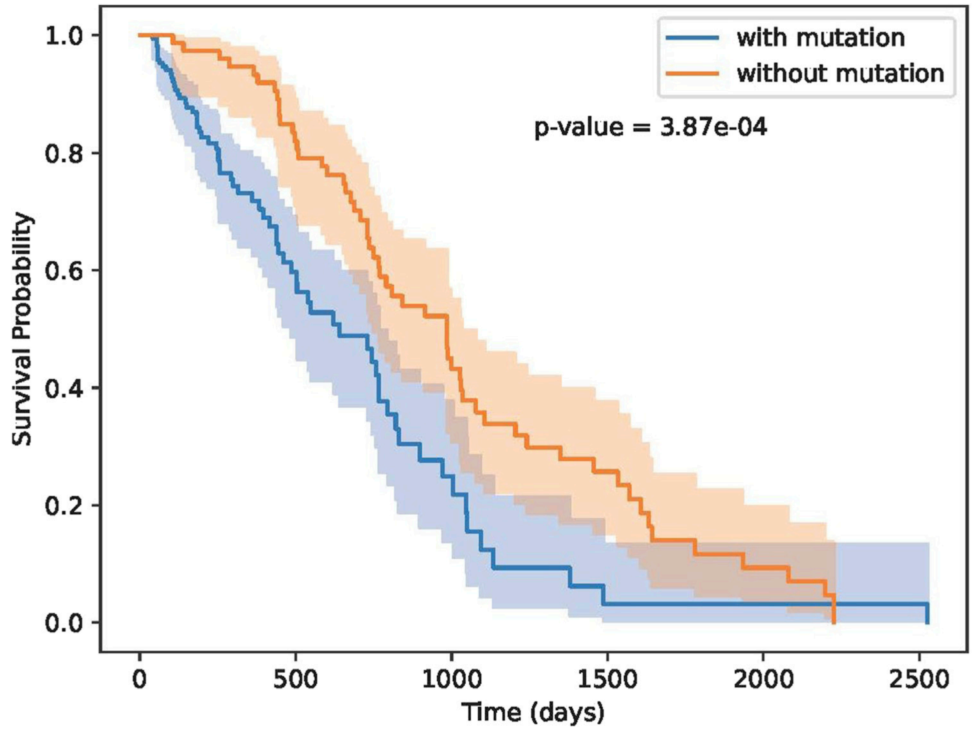


FIGURE 8
Results on Miao-2018 datasets.

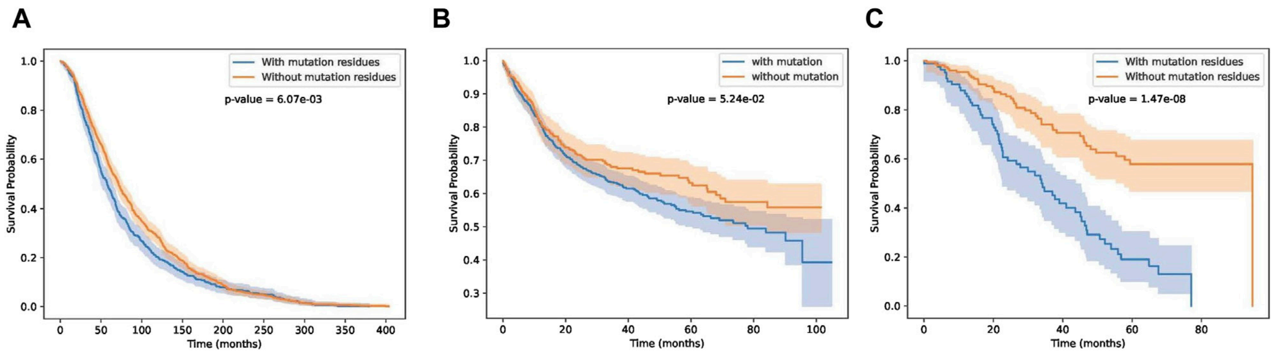


FIGURE 9
Survival curves on breast, bladder and pan cancer datasets.

performance of these biomarker genes, we analyzed Miao-2018 datasets (Table 2), results is show in Figure 8.

As illustrated in Figure 8, the identified biomarker mutations are also able to effectively separate patient groups with statistical significance. Our findings provide compelling evidence that the identified biomarker genes may possess valuable predictive power for immunotherapy response and patient survival outcomes. This highlights their potential as clinically relevant targets for the development of personalized treatment approaches. The results of this study advance the understanding of the underlying molecular mechanisms governing immunotherapy efficacy, and offer promising directions for future research and therapeutic interventions.

## 3.4 Performance on other cancer datasets

In this section, we test whether these key residue mutations and their corresponding genes can separate other cancer patients. Results are shown in Figures 9A–C. Detail information of these three cancer datasets are shown in Table 2. We can see that our biomarker genes can differentiate the three-cancer type significantly. Especially for the pan cancer datasets.

## 4 Conclusion

In summary, we proposed a new GCNN-based framework called GIHP for pan-specific HLA-peptide binding affinity prediction. GIHP harness both structure and sequence information and utilized Grad-WAM for visual interpretation. Extensive comparison with state-of-the-art methods verified the better performance of our methods. Collectively, the findings provide evidence that the GIHP framework has improved the generalization and interpretability capabilities of HLA-peptide binding prediction models. Furthermore, we have identified numerous key binding-related amino acid residues that can serve as potential biomarkers for differentiating patient groups based on immunotherapy response. When applying these identified biomarkers on datasets from other cancer types, they were also able to effectively differentiate patient groups with statistical significance. These findings highlight the potential prognostic significance of the biomarker gene mutation and underscore its relevance in patient stratification and personalized immunotherapy treatment approaches.

## Data availability statement

The data presented in the study are deposited in the Github, accession link: https://github.com/sdustSu/GIHP.

## Author contributions

LS: Funding acquisition, Methodology, Writing–original draft, Writing–review and editing. YY: Formal Analysis, Methodology, Validation, Visualization, Writing–review and editing. BM: Data curation, Formal Analysis, Investigation, Writing–review and editing. SZ: Formal Analysis, Methodology, Resources, Visualization, Writing–review and editing. ZC: Conceptualization, Project administration, Resources, Supervision, Writing–original draft, Writing–review and editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work is supported by Natural Science Foundation of Shandong Province (Youth Program, Grant No. ZR2022QF136), the Elite Program of Shandong University of Science and Technology and the National Science Foundation of China (Grant No. 62302277).

## Conflict of interest

Author YY was employed by Shandong Guohe Industrial Technology Research Institute Co. Ltd. BM was employed by Qingdao UNIC Information Technology Co. Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 32, 511–517. doi:10.1093/bioinformatics/btv639

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi:10.1126/science.abj8754

Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., et al. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput. Biol.* 13, e1005725. doi:10.1371/journal.pcbi.1005725

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Boukouaci, W., Rivera-Franco, M. M., Volt, F., Lajnef, M., Wu, C. L., Rafii, H., et al. (2024). HLA peptide-binding pocket diversity modulates immunological complications after cord blood transplant in acute leukaemia. *Br. J. Haematol.* 204, 1920–1934. doi:10.1111/bjh.19339

Bros, M., Haas, K., Moll, L., and Grabbe, S. (2019). RhoA as a key regulator of innate and adaptive immunity. *Cells* 8, 733. doi:10.3390/cells8070733

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L., et al. (1993). Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364, 33–39. doi:10.1038/364033a0

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi:10.1158/2159-8290.CD-12-0095

Cheng, J., Bendjama, K., Rittner, K., and Malone, B. (2021). BERTMHC: improved MHC-peptide class II interaction prediction with transformer and multiple instance learning. *Bioinformatics* 37, 4172–4179. doi:10.1093/bioinformatics/btab422

Clinton, T. N., Chen, Z., Wise, H., Lenis, A. T., Chavan, S., Donoghue, M. T. A., et al. (2022). Genomic heterogeneity as a barrier to precision oncology in urothelial cancer. *Cell Rep.* 41, 111859. doi:10.1016/j.celrep.2022.111859

Consortium, I. T. P., Abascal, F., Abeshouse, A., Aburatani, H., Adams, D. J., Agrawal, N., et al. (2020). Pan-cancer analysis of whole genomes. *Nature* 578, 82–93. doi:10.1038/s41586-020-1969-6

East, P., Kelly, G. P., Biswas, D., Marani, M., Hancock, D. C., Creasy, T., et al. (2022). RAS oncogenic activity predicts response to chemotherapy and outcome in lung adenocarcinoma. *Nat. Commun.* 13, 5632. doi:10.1038/s41467-022-33290-0

Ge, S. X., Jung, D., and Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* 36, 2628–2629. doi:10.1093/bioinformatics/btz931

Gfeller, D., Guillaume, P., Michaux, J., Pak, H. S., Daniel, R. T., Racle, J., et al. (2018). The length distribution and multiple specificity of naturally presented HLA-I ligands. *J. Immunol.* 201, 3705–3716. doi:10.4049/jimmunol.1800914

Giguere, S., Drouin, A., Lacoste, A., Marchand, M., Corbeil, J., and Laviolette, F. (2013). MHC-NP: predicting peptides naturally processed by the MHC. *J. Immunol. Methods* 400-401, 30–36. doi:10.1016/j.jim.2013.10.003

Gizinski, S., Preibisch, G., Kucharski, P., Tyrolski, M., Rembalski, M., Grzegorczyk, P., et al. (2024). Enhancing antigenic peptide discovery: improved MHC-I binding prediction and methodology. *Methods* 224, 1–9. doi:10.1016/j.ymeth.2024.01.016

Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., et al. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 154, 394–406. doi:10.1111/imm.12889

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* 199, 3360–3368. doi:10.4049/jimmunol.1700893

Kallingal, A., Olszewski, M., Maciejewska, N., Brankiewicz, W., and Baginski, M. (2023). Cancer immune escape: the role of antigen presentation machinery. *J. Cancer Res. Clin. Oncol.* 149, 8131–8141. doi:10.1007/s00432-023-04737-8

Karosiene, E., Lundegaard, C., Lund, O., and Nielsen, M. (2012). NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 64, 177–186. doi:10.1007/s00251-011-0579-8

Karosiene, E., Rasmussen, M., Blicher, T., Lund, O., Buus, S., and Nielsen, M. (2013). NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* 65, 711–724. doi:10.1007/s00251-013-0720-y

Kim, K., Kim, H. S., Kim, J. Y., Jung, H., Sun, J. M., Ahn, J. S., et al. (2020). Predicting clinical benefit of immunotherapy by antigenic or functional mutations affecting tumour immunogenicity. *Nat. Commun.* 11, 951. doi:10.1038/s41467-020-14562-z

Kim, Y., Sidney, J., Buus, S., Sette, A., Nielsen, M., and Peters, B. (2014). Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. *BMC Bioinforma.* 15, 241. doi:10.1186/1471-2105-15-241

Kim, Y., Sidney, J., Pinilla, C., Sette, A., and Peters, B. (2009). Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinforma.* 10, 394. doi:10.1186/1471-2105-10-394

Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O., and Nielsen, M. (2008). NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res.* 36, W509–W512. doi:10.1093/nar/gkn202

Ma, L., Zhang, N., Qu, Z., Liang, R., Zhang, L., Zhang, B., et al. (2020). A glimpse of the peptide profile presentation by *Xenopus laevis* MHC class I: crystal structure of pXela-UAA reveals a distinct peptide-binding groove. *J. Immunol.* 204, 147–158. doi:10.4049/jimmunol.1900865

Meng, Z., Chen, C., Zhang, X., Zhao, W., and Cui, X. (2024). Exploring fragment adding strategies to enhance molecule pretraining in AI-driven drug discovery. *Big Data Min. Anal.*, 1–12. doi:10.26599/bdma.2024.9020003

Miao, D., Margolis, C. A., Vokes, N. I., Liu, D., Taylor-Weiner, A., Wankowicz, S. M., et al. (2018). Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nat. Genet.* 50, 1271–1281. doi:10.1038/s41588-018-0200-2

Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D. C., Sidney, J., Bui, H. H., et al. (2006). A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. *Nat. Biotechnol.* 24, 817–819. doi:10.1038/nbt1215

Murata, K., Ly, D., Saijo, H., Matsunaga, Y., Sugata, K., Ihara, F., et al. (2022). Modification of the HLA-A*24:02 peptide binding pocket enhances cognate peptide-binding capacity and antigen-specific T cell activation. *J. Immunol.* 209, 1481–1491. doi:10.4049/jimmunol.2200305

Nielsen, M., Lundegaard, C., and Lund, O. (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinforma.* 8, 238. doi:10.1186/1471-2105-8-238

Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S. L., Lamberth, K., Buus, S., et al. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* 12, 1007–1017. doi:10.1110/ps.0239403

O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* 7, 129–132. doi:10.1016/j.cels.2018.05.014

O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020). MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst.* 11, 418–419. doi:10.1016/j.cels.2020.09.001

Öztürk, H., Ozkirimli, E., and Özgür, A. (2019). *WideDTA: prediction of drug-target binding affinity*. arXiv:1902.04166.

Quiros, M., Grazulis, S., Girdzijauskaite, S., Merkys, A., and Vaitkus, A. (2018). Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. *J. Cheminform* 10, 23. doi:10.1186/s13321-018-0279-6

Razavi, P., Chang, M. T., Xu, G., Bandlamudi, C., Ross, D. S., Vasan, N., et al. (2018). The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell* 34, 427–438. doi:10.1016/j.ccell.2018.08.008

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* 48, W449–W454. doi:10.1093/nar/gkaa379

Salkeni, M. A., and Naing, A. (2023). Interleukin-10 in cancer immunotherapy: from bench to bedside. *Trends Cancer* 9, 716–725. doi:10.1016/j.trecan.2023.05.003

Samstein, R. M., Lee, C. H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., et al. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* 51, 202–206. doi:10.1038/s41588-018-0312-8

Seidel, R. D., Merazga, Z., Thapa, D. R., Soriano, J., Spaulding, E., Vakkasoglu, A. S., et al. (2021). Peptide-HLA-based immunotherapeutics platforms for direct modulation of antigen-specific T cells. *Sci. Rep.* 11, 19220. doi:10.1038/s41598-021-98716-z

Stern, L. J., Brown, J. H., Jardetzky, T. S., Gorga, J. C., Urban, R. G., Strominger, J. L., et al. (1994). Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* 368, 215–221. doi:10.1038/368215a0

van Deutekom, H. W. M., and Keşmir, C. (2015). Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most? *Immunogenetics* 67, 425–436. doi:10.1007/s00251-015-0849-y

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50, D439–D444. doi:10.1093/nar/gkab1061

Venkatesh, G., Grover, A., Srinivasaraghavan, G., and Rao, S. (2020). MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. *Bioinformatics* 36, i399–i406. doi:10.1093/bioinformatics/btaa479

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2019). The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* 47, D339–D343. doi:10.1093/nar/gky1006

Wang, M., and Claesson, M. H. (2014). Classification of human leukocyte antigen (HLA) supertypes. *Methods Mol. Biol.* 1184, 309–317. doi:10.1007/978-1-4939-1115-8_17

Wang, P., Sidney, J., Dow, C., Mothe, B., Sette, A., and Peters, B. (2008). A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput. Biol.* 4, e1000048. doi:10.1371/journal.pcbi.1000048

Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M., et al. (2010). Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinforma.* 11, 568. doi:10.1186/1471-2105-11-568

Wang, X., Wu, T., Jiang, Y., Chen, T., Pan, D., Jin, Z., et al. (2024). RPEMHC: improved prediction of MHC-peptide binding affinity by a deep learning approach based on residue-residue pair encoding. *Bioinformatics* 40, btad785. doi:10.1093/bioinformatics/btad785

Wang, Y., Jiao, Q., Wang, J., Cai, X., Zhao, W., and Cui, X. (2023). Prediction of protein-ligand binding affinity with deep learning. *Comput. Struct. Biotechnol. J.* 21, 5796–5806. doi:10.1016/j.csbj.2023.11.009

Wen, Z., He, J., Tao, H., and Huang, S. Y. (2019). PepBDB: a comprehensive structural database of biological peptide-protein interactions. *Bioinformatics* 35, 175–177. doi:10.1093/bioinformatics/bty579

Yang, Z., Zhong, W., Zhao, L., and Yu-Chian Chen, C. (2022). MGraphDTA: deep multiscale graph neural network for explainable drug-target binding affinity prediction. *Chem. Sci.* 13, 816–833. doi:10.1039/d1sc05180f

You, R., Qu, W., Mamitsuka, H., and Zhu, S. (2022). DeepMHCII: a novel binding core-aware deep interaction model for accurate MHC-II peptide binding affinity prediction. *Bioinformatics* 38, i220–i228. doi:10.1093/bioinformatics/btac225

Zhang, H., Lund, O., and Nielsen, M. (2009). The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* 25, 1293–1299. doi:10.1093/bioinformatics/btp137

Zhao, W., and Sher, X. (2018). Systematically benchmarking peptide-MHC binding predictors: from synthetic to naturally processed epitopes. *PLoS Comput. Biol.* 14, e1006457. doi:10.1371/journal.pcbi.1006457

Frontiers in Genetics

TYPE Original Research
PUBLISHED 15 July 2024
DOI 10.3389/fgene.2024.1440448

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Xiaohua Wan,
Beijing Institute of Technology, China
Xinguo Lu,
Hunan University, China

*CORRESPONDENCE
Lingtao Su,
✉ sulingtao@163.com

RECEIVED 29 May 2024
ACCEPTED 24 June 2024
PUBLISHED 15 July 2024

CITATION
Zhao S, Cui Z, Zhang G, Gong Y and Su L (2024), MGPPI: multiscale graph neural networks for explainable protein–protein interaction prediction.
*Front. Genet.* 15:1440448.
doi: 10.3389/fgene.2024.1440448

COPYRIGHT
© 2024 Zhao, Cui, Zhang, Gong and Su. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# MGPPI: multiscale graph neural networks for explainable protein–protein interaction prediction

Shiwei Zhao, Zhenyu Cui, Gonglei Zhang, Yanlong Gong and Lingtao Su*

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China

Protein-Protein Interactions (PPIs) involves in various biological processes, which are of significant importance in cancer diagnosis and drug development. Computational based PPI prediction methods are more preferred due to their low cost and high accuracy. However, existing protein structure based methods are insufficient in the extraction of protein structural information. Furthermore, most methods are less interpretable, which hinder their practical application in the biomedical field. In this paper, we propose MGPPI, which is a Multiscale graph convolutional neural network model for PPI prediction. By incorporating multiscale module into the Graph Neural Network (GNN) and constructing multi convolutional layers, MGPPI can effectively capture both local and global protein structure information. For model interpretability, we introduce a novel visual explanation method named Gradient Weighted interaction Activation Mapping (Grad-WAM), which can highlight key binding residue sites. We evaluate the performance of MGPPI by comparing with state-of-the-arts methods on various datasets. Results shows that MGPPI outperforms other methods significantly and exhibits strong generalization capabilities on the multi-species dataset. As a practical case study, we predicted the binding affinity between the spike (S) protein of SARS-COV-2 and the human ACE2 receptor protein, and successfully identified key binding sites with known binding functions. Key binding sites mutation in PPIs can affect cancer patient survival statues. Therefore, we further verified Grad-WAM highlighted residue sites in separating patients survival groups in several different cancer type datasets. According to our results, some of the highlighted residues can be used as biomarkers in predicting patients survival probability. All these results together demonstrate the high accuracy and practical application value of MGPPI. Our method not only addresses the limitations of existing approaches but also can assists researchers in identifying crucial drug targets and help guide personalized cancer treatment.

KEYWORDS

**PPI prediction, multiscale GNN, model interpretability, key binding residues, Grad-WAM**

# 1 Introduction

PPIs are the basic components of protein complexes, which play crucial roles in cellular components and biological processes (Chen and Zacharias, 2023; Wu et al., 2023; Reed et al., 2024). Dysfunction interactions can lead to various chronic diseases and even cancer. Moreover, PPI contains a wealth of information, such as receptor binding and immune response, and can providing key insights of protein functionality and potential therapeutic targets for human cancer (Kjer-Hansen and Weatheritt, 2023; Rodina et al., 2023). As a results, numerous computational methods have been developed for PPI prediction, which are more efficient and low cost in comparison with traditional experimental based methods.

Early computational methods primarily use protein sequence as input (Ding and Kihara, 2018; Wang et al., 2021; Boldridge et al., 2023; Gainza et al., 2023; Shen et al., 2023). Such methods based on the hypothesis that proteins with similar sequences tend to have similar binding tendencies. With the increasing availability of protein structure data, especially with the emergence of Alphafold2 (Bryant et al., 2022), structure-based PPI prediction algorithms (Dong et al., 2019; Tang et al., 2023) have been growing year by year. In comparison with sequence-based methods, structure-based methods can capture more detailed binding structure information, therefore, the prediction accuracy are often more higher. PPI prediction methods can be categorized into two groups: machine learning and deep learning based methods. Among them, Support Vector Machines (SVM) (Guo et al., 2008; Wong et al., 2015; Bandyopadhyay and Mallick, 2016; Zhou et al., 2017) based methods aim to find an optimal hyperplane using protein sequence information (Chen and Jeong, 2009; Xia et al., 2010; Zahiri et al., 2013; You et al., 2015), 3D structure (Li et al., 2012), and domain information (Chen and Liu, 2005) to maximize the margin between different proteins for classification. Decision tree-based methods, on the other hand, utilize features such as protein 3D structure, primary sequence, and domain composition for PPI prediction. Compared to traditional machine learning-based methods, deep learning has the ability to automatically learn higher-level feature representations. Among them, DeepFE-PPI (Yao et al., 2019) proposes a novel residue representation method and deep learning network for protein-protein interaction prediction. Deep-Trio (Hu et al., 2022) introduces a sequence-based approach for PPI prediction, utilizing multiple parallel convolutional neural networks. GNN-PPI (Jha et al., 2022) leverages graph neural networks and language models (LM) to extract high-quality features from proteins for predicting protein interactions. HIGH-PPI (Gao et al., 2023) consists of bottom-level protein graph neural network (BGNN) representation learning and top-level PPI graph neural network (TGNN) representation learning. The vector representations obtained from both networks are concatenated to obtain the final prediction result.

While existing methods have achieved promising results on datasets from various species, most of them lack sufficient protein feature extraction and interpretability. Sequence-based models primarily focus on the one-dimensional sequence characteristics of proteins, while neglecting the higher-order structural properties. This can lead to incomplete accuracy in predicting PPIs since structural information plays a crucial role. Even when some models take into account the structural information of proteins, they often fail to adequately address how to extract both global and local structural information from proteins to contribute to PPI prediction. Additionally, some models incorporating interpretable modules solely rely on spatial biological arrangements of residues, introducing uncertainties and challenges to scientific validity and reliability. These limitations hinder their practical applications in PPI prediction. To address the issue of insufficient protein feature extraction, we choose to utilize graph convolutional layers to capture as much global structural information of protein graphs as possible. However, we need to find a suitable trade-off due to challenges such as over smoothing and gradient vanishing, which may arise when using multiple graph convolutional layers. Secondly, given the uniqueness of proteins, GNN should preserve the local structural information of proteins. Certain amino acid residues are crucial for protein interactions, and even the presence of specific residues determines protein functionality. Therefore, GNN should effectively distinguish between important residues and less relevant ones, enabling reasonable judgments in subsequent site prediction experiments. Secondly, the current interpretability of PPI models based on graph neural networks falls short in translating interactions into an understanding of function and mechanism. Moreover, explanations based solely on spatial biological arrangement information of residues are insufficient. Sequence information provides only static insights, disregarding the importance of protein structure and dynamic characteristics in determining the occurrence and stability of interactions.

To address the limitations in protein feature extraction and interpretability, we propose a novel framework called MGPPI, our main contributions are as follows.

- We represent both interacting proteins as amino acid level graphs, with amino acids as nodes and various relationships between them as edges, which allows MGPPI to capture the internal structure of proteins and their interactions more accurately.
- To address the issue of black-box features in existing deep learning models, we propose a novel interpretability module called Grad-WAM. Grad-WAM utilizes the gradient magnitudes generated by the final Graph Convolutional Network (GCN) layer of the model to calculate the contributions of each amino acid position in the PPI prediction. This information is then used to visualize the crucial amino acid residues that play a key role in the interaction between the two proteins.
- To address the issue of insufficient protein structure feature extraction in existing models, we propose a Multiscale Graph Convolutional Neural Networkk (MGCN) to learn both local and global protein structural representations. These representations are mapped into feature vectors for each protein, and the are combined for PPI prediction.

# 2 Methods

## 2.1 Input representation

The workflow of MGPPI is shown in Figure 1. The input to the model is paired protein structures, which are represented as amino

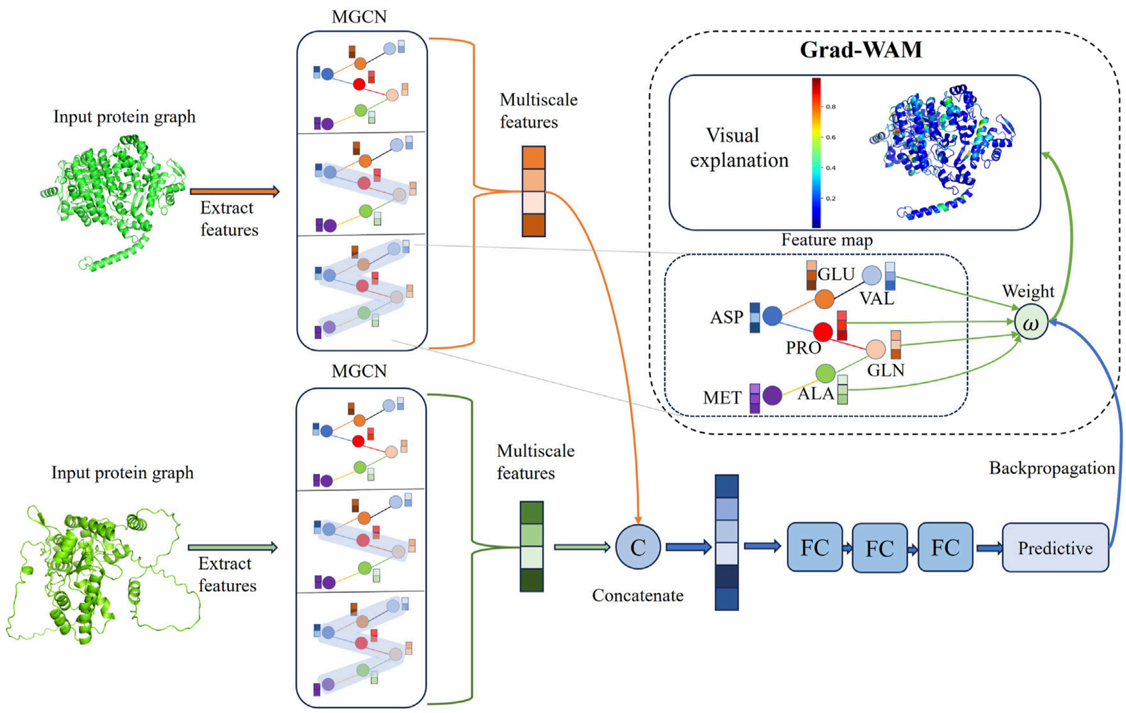


FIGURE 1
The model takes protein structures as input and employs graph representation learning to feed them into MGCN for extracting multi-scale features of proteins, resulting in the final protein representation vectors. The two protein representation vectors are then fused to obtain a combined representation of the protein pair. Subsequently, the combined representation is fed into fully connected layers to output predicted interaction scores. GradWAM utilizes the gradient information from the last graph convolutional layer of MGCN and the final predicted scores to analyze the importance of each amino acid for protein-protein interactions.

**TABLE 1** Protein graph representation at the amino acid level.

| Name | Size |
|---|---|
| Node features | |
| BLOSUM62 descriptors | 23 |
| AAPHY7 descriptors | 7 |
| One-hot encoded belonging to secondary structure | 6 |
| Solvent-accessible surface area | 1 |
| $\phi$ angle (divided by 180) | 1 |
| $\psi$ angle (divided by 180) | 1 |
| Hydrogen bond acceptor | 1 |
| Hydrogen bond donor | 1 |
| Edge features | |
| Covalent bond | 1 |
| Hydrophobic contact | 1 |
| Hydrogen bond | 1 |
| Aromatic bond | 1 |
| Ionic bond | 1 |
| Disulfide bond | 1 |

acid level graphs G = (N, E). Where nodes (N) are amino acid residues (i.e., amino acid-level graph representation) and the relationships between amino acids as edges (E). Node attributes including solvent-accessible surface area, $\phi$ angle,$\psi$ angle, secondary structure (alpha helix, isolated beta-bridge residue, strand, 3–10 helix, turn, bend), AAPHY7 and BLOSUM62 descriptors, hydrogen bond acceptor and donor information. Edge attributes including information about the existence of covalent bond, hydrophobic contact, ionic bond, disulfide bond, hydrogen bond, and aromatic bond relationships between amino acids as edge attributes. The details are shown in Table 1 below.

## 2.2 Graph neural network

We extract protein features using Graph Neural Networks and combine the feature vectors of two proteins to predict Protein-Protein Interactions. We map the protein graph representation to feature vectors through two stages: message passing and readout. In the message passing stage, as shown in Eq. 1, corresponding to Figure 2A, we update the feature vector of each node by incorporating the feature information from its neighboring nodes.

$$x_i^{(T)} = \sigma\left(\Phi_1 x_i^{(T-1)} + \Phi_2 \sum_{j\in\mathcal{N}(i)} x_j^{(T-1)}\right) \tag{1}$$

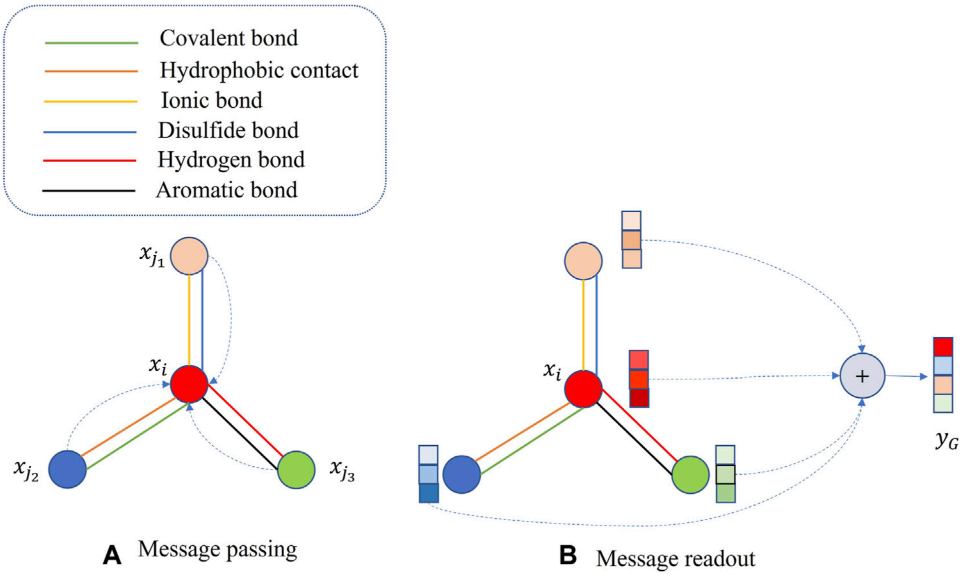


FIGURE 2
Protein graph representation learning. **(A)** message passing phase, **(B)** message readout phase.

$x_i^{(T)}$ represents the feature vector of the $i$th node at a time step T; $\Phi_1, \Phi_2$ symbolize learnable weight matrices shared by all nodes, initialized with small random values pre-training, and continuously optimized during training; $\sigma$ encompasses two operations: node-level batch normalization (Li et al., 2021) and the ReLU activation function; $N(i)$ refers to the set of neighboring nodes for the $i$th node. By utilizing Eq. 1, nodes can gradually capture more global information from the protein graph representation.

$$y_G = \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} x_i^{(T)} \tag{2}$$

$\mathcal{M}$ represents the set of all amino acids in the protein. During the readout stage, as shown in Eq. 2, corresponding to Figure 2B, we obtain the representation vector, denoted as $y_G$, for the entire protein graph $G$.

## 2.3 Multiscale graph convolutional neural network for protein encoding

Graph neural network extract features of target proteins through a layer-wise sampling approach. The layer-wise sampling approach allows the model to extract the node features after each convolutional layer, enabling the model to capture as much of the protein's global and local features as possible. A small receptive field allows nodes to observe only local protein structures, failing to capture global structural features. As a result, nodes fail to establish connections with the overall protein structure. On the contrary, when the receptive field is too large, nodes may absorb more irrelevant features that are unrelated to protein interactions. Additionally, it can lead to the homogenization of node features within a particular region, giving rise to the problem of oversmoothing. To effectively learn and integrate features from protein graph data at different receptive fields and granularities, we propose a Multiscale graph convolutional neural network (MGCN). MGCN consists of three multiscale blocks and three transition layers, the multiscale blocks as shown in Figure 3, the transition layer transfers the multiscale information of nodes to the next stage. At time step $n+1$, the transition layer as shown in Eq. 3.

$$\begin{aligned} x_i^{(n+1)} = \sigma\Big(&\Phi_1\big(x_i^{(0)}\|x_i^{(1)}\|\cdots\|x_i^{(n)}\big) \\ &+\Phi_2 \sum_{j\in\mathcal{N}(i)} \big(x_j^{(0)}\|x_j^{(1)}\|\cdots\|x_j^{(n)}\big)\Big) \end{aligned} \tag{3}$$

The purpose of the transition layers is to connect adjacent multiscale blocks, facilitating an increase in the depth of MGCN. The multiscale blocks allow gradients to propagate through skip connections, alleviating the issue of gradient vanishing. Additionally, MGCN enhances the representation capability of nodes by concatenating combinations of features from different receptive fields. After passing through the final transition layer, the feature vectors of amino acids are propagated to the readout stage. In the information readout stage, the feature vectors of all amino acids in a protein are integrated and transformed into a single feature vector representing that protein. This protein feature vector is then utilized for subsequent stages of protein-protein interaction prediction.

## 2.4 MGPPI network architecture

After obtaining the vector representations of proteins, we concatenate the vector representations of both proteins and feed them into fully connected layers to predict interaction scores. Since the predicted interaction scores fall within the range of 0–1, we need to set a threshold for classification. As we are going to predict key sites of protein interaction later, we need to minimize the false positive rate as much as possible to enhance the credibility of

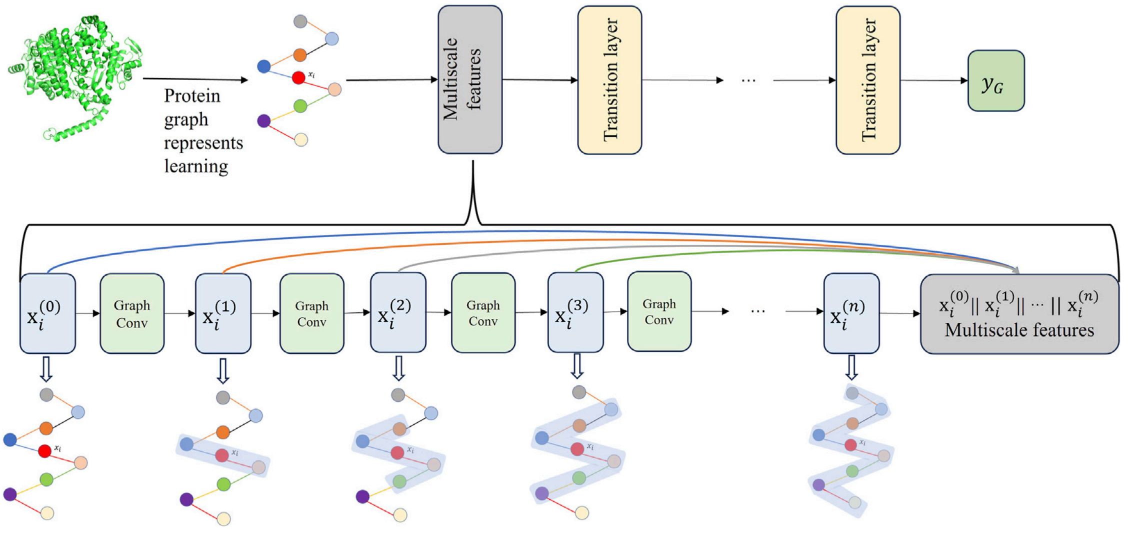


FIGURE 3
The upper part of the figure illustrates the workflow of MGCN, which consists of three multi-scale blocks and three transition layers. The lower part of the figure provides a detailed description of the multiscale blocks.

prediction results, and hence, we have set the classification threshold to 0.7. If the score is greater than this threshold, it is assigned a value of 1, indicating that the two proteins are likely to interact with each other. Conversely, if the score is below the threshold, it is assigned a value of 0, suggesting that the two proteins are less likely to interact or are unable to interact. Each fully connected layer is followed by a ReLU activation function and a dropout layer with a dropout rate of 0.1, consistent with previous studies. There are a total of three fully connected layers. Our loss function uses cross-entropy loss (Yang et al., 2021), defined as follows Eq. 4:

$$Cross\ Entropy\ Loss = -\sum_{i=1}^{n} Y_i \log P_i \tag{4}$$

$P_i$ represents the predicted interaction score for the $i$th protein pair, $Y_i$ represents the correct interaction score for the $i$th protein pair, and $n$ represents the total number of protein pairs.

## 2.5 Gradient weighted interaction activation mapping

To enhance the interpretability of the model, we propose an interpretable module called Gradient weighted interaction activation mapping (Grad-WAM) and visualize the results. Grad-WAM improves the identification of key interacting amino acid residues and explains the mechanism of protein-protein interactions, effectively addressing the issue of lack of interpretability in neural network predictions. Grad-WAM utilizes the gradient magnitudes generated by the last layer of graph convolutions and the final predicted scores to calculate the contributions of different amino acid positions in the protein structure to protein-protein interaction predictions.

Specifically, Grad-WAM uses a weighted combination of the positive partial derivatives of the feature maps with respect to the interaction values to generate the corresponding visual explanations. Since the contributions of each element are not equal, an additional weight is introduced to weight the gradient values. The calculation formula is as follows:

$$\omega = \sum_i \alpha_i \cdot ReLU\left(\frac{\partial P}{\partial T_i}\right), \quad \forall\{i \mid i\epsilon T\} \tag{5}$$

Where $\omega$ represents the weight, and positive gradient values indicate a positive influence on the predicted values, ensuring that $\omega$ is a weighted average rather than a global average. $\alpha_i$ corresponds to the gradient weight of the $i$th node. $ReLU()$ denotes the ReLU activation function. $T_i$ is the feature value of the $i$th node in the feature map $T$ of the last graph convolutional layer. $P$ represents the predicted protein-protein interaction value, and the calculation formula is as follows:

$$P = \sum_i \alpha_i \cdot ReLU\left(\frac{\partial P}{\partial T_i}\right) \cdot T_i \tag{6}$$

The derivative of Eq. 6 with respect to the variable yields the following Eq. 7:

$$\frac{\partial P}{\partial T_i} = \alpha_i \cdot \frac{\partial P}{\partial T_i} + T_i \cdot \alpha_i \cdot \frac{\partial^2 P}{(\partial T_i)^2} \tag{7}$$

Rearranging the terms in Eq. 7 yields the following Eq. 8:

$$\alpha_i = \frac{\frac{\partial P}{\partial T_i}}{\frac{\partial P}{\partial T_i} + T_i \cdot \frac{\partial^2 P}{(\partial T_i)^2}} \tag{8}$$

Substituting the weight $\alpha_i$ from Eq. 8 into Eq. 5, yields the final weight as shown in Eq. 9:

$$\omega = \sum_i \left[\frac{\frac{\partial P}{\partial T_i}}{\frac{\partial P}{\partial T_i} + T_i \cdot \frac{\partial^2 P}{(\partial T_i)^2}}\right] \cdot \quad ReLU\left(\frac{\partial P}{\partial T_i}\right), \ \forall\{i \mid i\epsilon T\} \tag{9}$$

TABLE 2 The quantity of positive and negative samples in each dataset.

| Dataset | Positive | Negative |
|---|---|---|
| HPRD | 35944 | 763115 |
| OPHID | 39412 | 63932 |
| BioGRID | 72367 | 82731 |
| STRING | 11810480 | 167224 |
| Multi-species | 25640 | 30332 |

The contribution of different amino acids at various positions in the protein structure to the prediction of protein-protein interactions can be calculated using Eq. 9. A key amino acid that plays a crucial role in protein-protein interactions is annotated and displayed in the protein structure. The color gradient from blue to green to red represents the contribution values of the amino acids, with higher contribution values indicated by a redder color, indicating their significant role in protein interactions. This method enhances the interpretability of the model. Grad-WAM calculates the contribution of different amino acid positions in protein-protein interaction prediction within protein structures by utilizing the gradient magnitude generated by the final graph convolutional layer and the predicted values propagated through backpropagation. Due to the local connectivity and weight sharing structure employed by the graph convolutional layer, it preserves spatial information lost in the fully connected layers. The last graph convolutional layer strikes a balance between high-order semantics and detailed spatial information (Selvaraju et al., 2017), considering both global and local features. Finally, the minimum-maximum normalization method is used to map the impact probabilities of each amino acid on protein interactions, ranging from 0 to 1.

## 2.6 Data

The datasets used in this study include the Human Protein Reference Database (HPRD) (Peri et al., 2003), the Online Predicted Human Interaction Database (OPHID) (Brown and Jurisica, 2005), the H. sapiens dataset from the Biological General Repository for Interaction Datasets (BioGRID) (Oughtred et al., 2019), and the STRING database (Szklarczyk et al., 2019). Additionally, the negative samples in the HPRD dataset are sourced from curated negative protein-protein interaction datasets. The negative protein-protein interaction datasets collected data on human protein pairs that did not exhibit interactions in large-scale yeast two-hybrid screening. The quantities of positive and negative samples after processing for each dataset are shown in Table 2.

For the aforementioned datasets, protein names were converted to UniProt (Bateman, 2019) ID, and the corresponding PDB (Berman et al., 2000) files were collected for training, testing, and validation purposes. We randomly sampled 25,000 positive examples and 25,000 negative examples from each human protein dataset, resulting in a final training set of 200,000 samples. The training set is independent of the subsequent test set.

## 2.7 Experimental environment configuration

The experimental environment consisted of Ubuntu 20.04.6LTS, an Intel (R) Core (TM) i5-10400 CPU, and an NVIDIA Corporation GP102G (Tesla P40) GPU. A batch size of 512 was set, and the Adam optimizer with a learning rate of 0.0005 was used to update the model parameters. The MGCN architecture comprised 15 graph convolutional layers, including three multiscale blocks. Each multiscale block consisted of N ($N = 4$) graph convolutional layers and three transition layers.

# 3 Result and discussion

## 3.1 Compare on human proteins datasets

We selected the HPRD as the benchmark dataset and compared MGPPI with several state-of-the-art PPI prediction methods for analysis. These methods include High-PPI, a hierarchical graph neural network-based PPI prediction method; GNN-PPI, a method that utilizes graph neural networks to learn PPI network topological structures; Deep-Trio, a deep learning framework based on a masked multiscale CNN architecture that learns multiscale contextual information from protein sequences; PIPR (Chen et al., 2019), an end-to-end framework based on recursive neural networks (RNNs) that incorporates pre-trained residue embeddings for protein representation; and DeepFE-PPI, a method that employs residue representation using the Res2vec [based on Word2vec (Mikolov et al., 2013)] approach. In Figure 4A, the precision-recall curves are provided, while Figure 4B presents the ROC curves. Across both evaluation metrics, MGPPI consistently achieved the best performance among all the compared methods, this also highlights the significance of protein structural information in PPI prediction. Furthermore, we conducted testing on the BioGRID dataset for MGPPI and five other methods, resulting in the confusion matrix shown in Figure 4C. The conclusions remain consistent with the previous findings.

To further validate the predictive capability of MGPPI for protein-protein interactions, we conducted performance comparisons between MGPPI and the five different PPI prediction methods on three other human protein datasets. The results as shown in Figure 5.

From Figure 4 and Figure 5, it can be observed that our proposed method, MGPPI, demonstrates favorable performance across various commonly used evaluation metrics. MGPPI aims to comprehensively consider the impact of oversmoothing and gradient vanishing while extracting as much global information as possible from protein graph structures. It simultaneously takes into account the preservation of local information within protein graph structures to enhance the prediction performance of PPI. On the other hand, High-PPI effectively utilizes a layered modeling approach. In this approach, the inner layer of the protein view consists of residues as nodes, with their physical adjacency forming the edges. The outer layer of the protein view considers proteins and their interactions as nodes and edges, respectively, in the PPI network structure. High-PPI, ranking second, highlights the significance of protein structural information in PPI prediction. In comparison to other sequence-based prediction methods such as

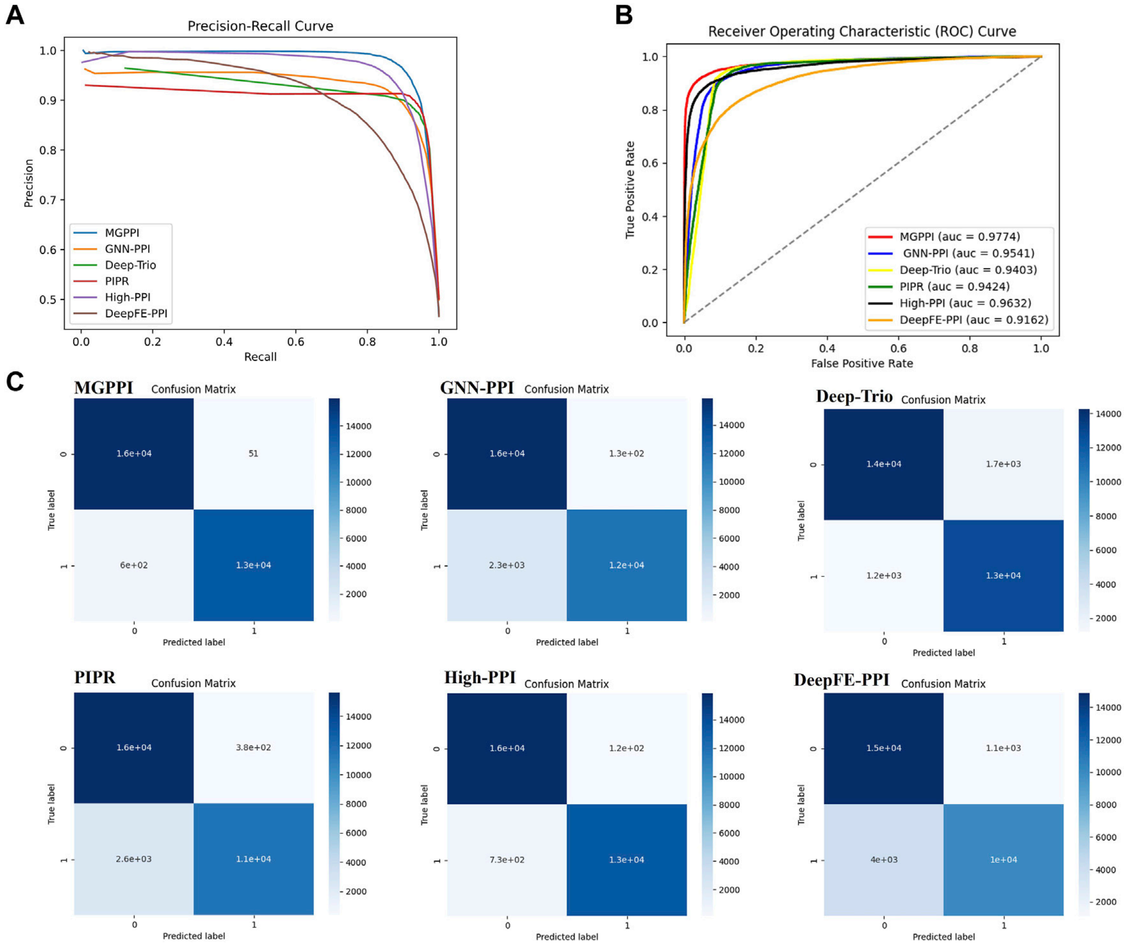


FIGURE 4
Here are the precision-recall curves **(A)**, ROC curves **(B)**, and AUC for the six methods on the HPRD dataset.**(C)** represents the 20% test set selected from the BioGRID dataset for evaluating MGPPI and five other methods. The confusion matrix was obtained by applying a threshold of 0.7 to the predicted values (as the predictions are continuous values between 0 and 1).

DeepFE-PPI, PIPR, and Deep-Trio, MGPPI exhibits significant advantages, further highlighting the importance of protein structure information in PPI prediction.

## 3.2 Compare on multi-species proteins datasets

The multi-species dataset consists of three species: *Caenorhabditis elegans* (Celegan), *Drosophila melanogaster* (*Drosophila*), and *Escherichia coli* (Ecoli). The utilization of multi-species datasets enables further exploration of the practical and generalization capabilities of MGPPI. Experimental evaluations conducted on these datasets assess the model's ability to generalize, as all previous models were trained and tested solely on human datasets. For MGPPI, protein network construction necessitates corresponding PDB files, obviating the necessity to establish distinct thresholds based on sequence similarity for data categorization. We standardized the dataset into a comprehensive multi-species dataset to assess the performance of each model effectively. The results as shown in Table 3.

Based on Table 3, it can be observed that MGPPI exhibits remarkable generalization ability. This can be attributed to its capability of performing graph representation learning on proteins from different species. MGPPI effectively extracts valuable structural information, enabling accurate prediction of protein-protein interactions. On one hand, sequence-based methods can accurately predict some PPI by identifying similarities between amino acid sequences of non-human species and those of human proteins, inferring similar functionalities and interaction tendencies. However, it should be noted that not all proteins from other species can be matched with similar sequences to human proteins. On the other hand, GNN-PPI might not have encountered multi-species data and thus struggles to accurately construct the PPI network structure, this leads to a performance deviation of the model from the anticipated expectations.

## 3.3 Ablation study

As the depth of GNN models increases, we have observed the issue of over-smoothing in certain cases. Over-smoothing refers to a

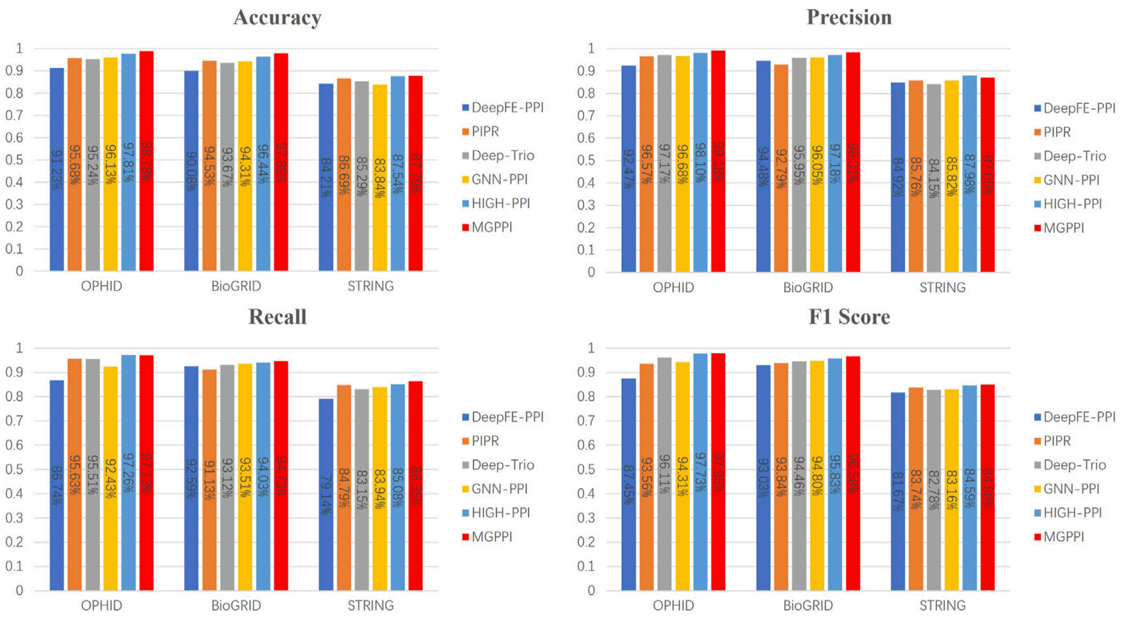


FIGURE 5
The overall performance of MGPPI compared to five other PPI prediction methods was evaluated on the OPHID dataset, the *H. sapiens* dataset from BioGRID, and STRING.

**TABLE 3** Comparison of MGPPI with other methods on a multi-species dataset (%).

| Method | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| DeepFE-PPI | 68.70 | 63.97 | 66.53 | 65.22 |
| PIPR | 73.92 | 65.98 | 73.15 | 68.69 |
| Deep-Trio | 71.29 | 69.57 | 70.67 | 68.54 |
| GNN-PPI | 67.15 | 61.53 | 64.82 | 62.10 |
| High-PPI | 75.18 | 73.26 | 75.21 | 73.52 |
| MGPPI | **82.53** | **81.77** | **83.68** | **82.17** |

situation where nodes incorporate an increasing amount of information from their neighboring nodes, causing the representation vectors of some nodes to converge towards the same value. When most or all nodes have representation vectors that converge to the same or a few values, it hinders the normal learning process of the model and renders the neural network's output insensitive to the input information. Therefore, in this paper, we propose MGCN to address this problem by integrating representation vectors from different temporal nodes, preserving a combination of receptive fields at different scales. This approach enhances the model's ability to represent nodes and alleviates the issue of over-smoothing.

In MGPPI, we alleviate the issues of over-smoothing and gradient vanishing by leveraging a multiscale module and batch normalization techniques to improve model performance. To demonstrate the individual contributions of the multiscale module and batch normalization, We conducted ablation study on the HPRD dataset. The study consisted of three experimental scenarios:

(1) The first scenario involved removing batch normalization while retaining the multi-scale module.
(2) The second scenario involved removing the multiscale module and utilizing the four graph convolution layers without it.
(3) The third scenario retained both the multiscale module and batch normalization for experimental analysis.

The results presented in Table 4 indicate that both the multi-scale module and batch normalization are essential components of MGCN. Furthermore, experiments conducted on the HPRD dataset aimed to investigate the impact of receptive field on the model's performance. Specifically, we progressively increased the number of graph convolutional layers (i.e., 2, 3, 4, 5, 6) within the multi-scale module to enlarge the network's receptive field.

From the results shown in Table 5, it can be observed that, overall, increasing the number of convolutional layers improves the overall performance of the model. However, when using six convolutional layers, some metrics are not as good as those achieved with four or five layers. This discrepancy may arise due to the inclusion of noise information from residues that do not participate in protein-protein interactions when increasing the number of convolutional layers. Considering that each additional convolutional layer introduces more computational operations, which could result in longer training and inference times, we aimed to strike a balance between the model's overall performance and its time complexity. Furthermore, we aimed to mitigate the impact of over-smoothing issues and gradient vanishing issues. Therefore, after comprehensive consideration, we opted for a compromise solution, retaining four convolutional layers for subsequent experiments.

**TABLE 4** Investigating the individual contributions of the multiscale module and batch normalization (%).

| Model | Loss | Accuracy | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|---|
| Without multiscale module | 1.50 | 98.16 | 99.31 | 95.86 | 97.53 | 95.07 |
| Without batch normalization | 1.34 | 98.28 | 99.29 | 95.61 | 97.45 | 95.15 |
| **MGPPI** | **1.16** | **98.74** | **99.52** | **96.11** | **97.78** | **95.41** |

**TABLE 5** Investigating the impact of different numbers of convolutional layers within the multiscale module on the model (%).

| Convolutional layer | Loss | Accuracy | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|---|
| 2 | 1.97 | 97.87 | 98.86 | 95.33 | 96.86 | 94.78 |
| 3 | 1.54 | 98.04 | 99.14 | 95.72 | 97.18 | 95.09 |
| 4 | 1.16 | **98.74** | 99.52 | **96.11** | **97.78** | 95.41 |
| 5 | **1.12** | 98.63 | **99.56** | 96.11 | 97.64 | 95.35 |
| 6 | 1.29 | 98.40 | 99.39 | 95.93 | 97.43 | **95.46** |

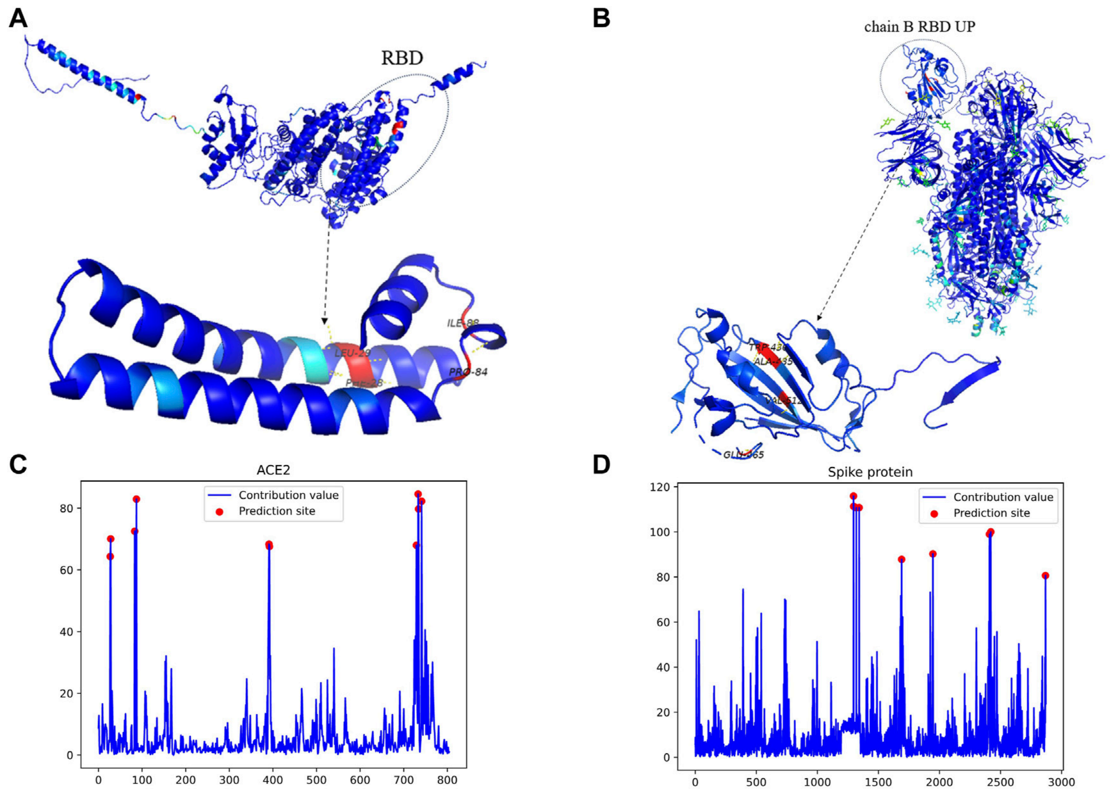


FIGURE 6
The key binding sites and amino acid contributions of the Spike protein and the human receptor ACE2. **(A)** represents the visualization results and predicted binding sites for the human receptor protein ACE2. **(B)** Represents the visualization results and predicted binding sites for the Spike protein. **(C)** Displays the contribution values of all amino acids in ACE2, represented by a single amino acid chain. **(D)** Displays the contribution values of all amino acids in the Spike protein, which consists of three amino acid chains: **(A, B, and C)**. Furthermore, the receptor binding domains on chains A and C are oriented downward, while the receptor binding domain on chain B is oriented upward. In the visualization, the region to the left of position 1,500 with higher scores corresponds precisely to the amino acids 435Ala, 436Trp, 512Val, and 465Glu in the B chain. Since the calculated contribution values are relatively small, we have proportionally magnified them for better color distinction during the visualization stage.

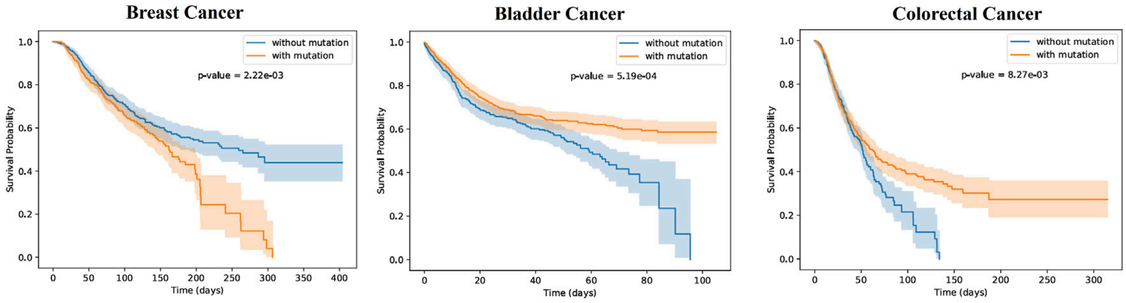


FIGURE 7
The patients with the three types of cancer were divided into two groups each, based on the occurrence of mutations in the key sites (mutated group) or the absence of mutations (non-mutated group). Survival probability differences over time were analyzed for the two patient groups of each cancer type, as depicted in the figures. The shaded regions around the curves represent the confidence intervals. The *p*-values were found to be 1.57e-03, 5.19e-04, and 8.27e-03, indicating statistically significant differences in survival probabilities between the two patient groups for all three types of cancer. This suggests that the presence of mutations has an impact on patient survival. In the case of breast cancer, the non-mutated group exhibited higher survival probabilities over time compared to the mutated group. Conversely, for the other two types of cancer, the mutated group showed higher survival probabilities as time progressed.

## 3.4 Predict and validate binding sites

To validate the interpretability and binding site prediction capability of MGPPI, we conducted an experiment using the interaction between the spike protein of SARS-COV-2 and the human ACE2 receptor protein as a case study. After feature extraction, graph convolution operations, and visualization, we discovered a high interaction score between the two proteins, indicating their ability to interact, which aligns with existing research findings. Additionally, using Grad-WAM visualization, we identified key amino acid residues that play a crucial role in their binding, elucidating the specific binding sites between Spike protein and Ace2. As show in Figure 6. We uniformly amplified the contribution values of all amino acids by a factor of 100 to facilitate color differentiation of the amino acids during the visualization process, ultimately resulting in the contribution value plots shown in Figures 6C,D.

As shown in Figure 6B, the residues 435Ala, 436Trp, 512Val, and 465Glu located in the receptor binding domain of the Spike protein's B chain exhibit the highest contribution values, indicating their significant impact in the interaction with the human receptor protein ACE2 (Chi et al., 2020). Have revealed that the receptor binding domains of the other two chains in the Spike protein are oriented downward, while the receptor binding domain on the B chain is oriented upward. This orientation suggests that the receptor binding domain on the B chain, which is typically the first to come into contact with the human receptor protein, further supports the reliability of the predicted binding sites. To further validate our findings, we mapped these predicted binding sites to functional domains and identified their corresponding functional annotations. The identified sites are located within the IPR018548 and IPR042578 functional domains, and the respective functional annotations of these domains are as follows:

The IPR018548 domain functions as the spike protein S1 subunit, receptor binding domain, and $\beta$-coronavirus. The IPR042578 domain functions as the spike protein S1, S2, and S2$'$, where S1 is responsible for binding to host cells and initiating infection, S2 is involved in cell membrane fusion, and S2' facilitates viral fusion.

The GO annotation (GO:0039654) indicates involvement in the fusion of viral membrane with the host endosomal membrane. The GO annotation (GO:0019064) suggests involvement in the fusion of viral membrane with the host ER membrane. The GO annotation (GO:0016020) indicates participation in the entry of the virus into host cells through endocytosis.

The functional domain and Gene Ontology annotations associated with these predicted binding sites provide additional evidence to substantiate the reliability and scientific validity of our predictions. These functional domains, such as IPR018548 and IPR042578, along with the corresponding GO annotations, further support the significance of the predicted binding sites in terms of their functional relevance and their involvement in critical viral-host interactions.

As shown in Figure 6A, The predicted binding sites, 28Phe, 29Leu, 84Pro, and 88Ile, are located on the edges of ACE2 and are prone to interact with other proteins. These sites can be mapped to specific functional domains, with corresponding Gene Ontology (GO) annotations. However, upon reviewing published literature, the actual binding sites are on the frontward-facing region of the protein, specifically the residues 28Phe, 29Leu, 84Pro, and 88Ile, with the following corresponding Gene Ontology (GO) annotations:

GO:0006508 (Protein catabolic process by peptide bond hydrolysis): This annotation suggests that these binding sites may be involved in the hydrolysis of peptide bonds, leading to the breakdown of larger polypeptides into smaller ones or amino acids. GO:0008237 (Metalloendopeptidase activity): This annotation indicates that the binding sites may possess the enzymatic activity of a metalloendopeptidase, which involves the cleavage of peptide bonds within a protein. GO:0008241 (Peptidyl-dipeptidase activity): This annotation suggests that these sites may catalyze the release of C-terminal dipeptides from peptide chains. GO:0016020 (Membrane and protein complex-associated within lipid bilayer): This annotation implies that the proteins containing these binding sites are embedded within the lipid bilayer and associated with protein complexes.

The functional domain allocation and GO annotation of predicted binding sites provide scientific evidence for the potential roles of these sites in the interaction between ACE2 and spike protein. This further strengthens the reliability of the predicted binding sites. Additionally, the use of Grad-WAM allows for intuitive visualization of the prediction results, enabling researchers to perform more targeted experimental validations based on the obtained amino acid contribution values. Furthermore, through MGPPI and Grad-WAM, we elucidated the roles of the two protein interactions in cellular processes and the functional significance of the predicted sites in the interaction process, thus enhancing the practicality and scientific rigor of MGPPI.

## 3.5 The impact of the predicted mutation site on cancer patients

To further validate the scientific significance of MGPPI, we collected three cancer patient datasets (breast cancer, bladder cancer, and colorectal cancer) and analyzed them individually. We selected protein-protein interaction samples related to cancer from the HPRD dataset based on proteins present in the patients' bodies. We successfully predicted the critical binding sites of these proteins using the MGPPI model. Subsequently, based on the incidence of mutations at these protein binding sites within the cancer patient dataset, we categorized each individual into one of two groups: patients with mutated binding sites and patients with non-mutated binding sites. We then analyzed the impact of these mutation sites on the survival time of patients in each cancer group. Finally, Kaplan-Meier curves were generated by analyzing the survival time and status of patients with each type of cancer, incorporating patient grouping information. As show in Figure 7:

From Figure 7, it is evident that there are significant differences in survival probabilities between the two patient groups for each type of disease, indicating the importance of these key sites for human survival. For example, in breast cancer patients, the occurrence of specific mutations at critical amino acid residues of certain proteins may suggest the presence of abnormalities or functional alterations in their bodies. As a result, their resistance against cancer could be weakened, rendering them more vulnerable to its effects and leading to a rapid decline in survival rates over time. Conversely, non-mutated patients at key sites demonstrate higher survival probabilities, indicating a potential survival advantage associated with those specific protein sites. However, in the case of bladder cancer and colorectal cancer patients, the Kaplan-Meier curve results are opposite to those of breast cancer patients, with the mutated groups showing higher survival probabilities over time. This indicates that mutations at certain amino acid residues may not necessarily be harmful to patients, and in some cases, they can even have a positive impact on the treatment of certain cancers, thereby increasing patients' survival probabilities.

In conclusion, MGPPI accurately predicts crucial amino acid sites in cancer patients that play a significant role in disease resistance, further validating the reliability and scientific soundness of the MGPPI model.

## 3.6 Limitations and future direction

Although MGPPI has demonstrated advantages in protein-protein interaction prediction and binding site prediction, there are still limitations in this study. Firstly, the output of the MGPPI model is a probability value that requires setting a threshold to convert probabilities into classifications. Choosing an inappropriate threshold may result in the model missing some true positive samples, leading to lower evaluation metrics than the actual values.

In future work, we will incorporate the 3D coordinate information of amino acids to predict binding sites during protein-protein interactions. In real-world scenarios, proteins exhibit diverse shapes, and certain amino acids may be located inside the protein due to protein folding or distortion. The likelihood of these amino acids interacting with other proteins is low. Therefore, when discussing PPI and predicting binding sites, it is essential to consider the actual coordinate information of amino acids.

# 4 Conclusion

This paper presents a novel PPI prediction framework called MGPPI based on chemical intuition. MGPPI utilizes MGCN, which consists of 15 graph convolutional layers, to capture the multiscale structure of proteins. It also employs Grad-WAM for visual interpretation. Extensive experiments validate the superiority of this method, demonstrating significant improvements over existing approaches on four human protein datasets and one multi-species dataset. The ability of MGPPI to represent proteins from various species as graph data greatly enhances the model's generalization capability. Furthermore, MGPPI successfully predicts the interaction between the spike protein of SARS-COV-2 and the human ACE2 receptor protein. By utilizing Grad-WAM, the importance of amino acids is visualized as labels, and the rationality of predicted binding sites is validated based on functional domain and Gene Ontology annotation. Finally, we screened for relevant proteins from samples of three cancer patients and used the MGPPI model to predict the binding sites of these proteins. Based on whether these sites undergo mutations, we divided each type of cancer patient into two groups and investigated the impact of these sites on the survival status of patients with the three types of diseases. The research results indicate that MGPPI enhances the overall generalization and interpretability of PPI prediction models, making it a highly practical tool.

# Data availability statement

The data presented in the study are deposited in the Github. accession link: https://github.com/Shiwei-Zhao/MGPPI.

## Author contributions

SZ: Software, Visualization, Writing–original draft, Writing–review and editing, Conceptualization, Methodology. ZC: Data curation, Methodology, Writing–review and editing, Software, Conceptualization. GZ: Validation, Writing–review and editing, Investigation, Data curation. YG: Resources, Writing–review and editing, Data curation, Methodology, Validation. LS: Methodology, Project administration, Supervision, Writing–review and editing, Funding acquisition, Conceptualization, Resources.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work is supported by Natural Science Foundation of Shandong Province (Youth Program, Grant No. ZR2022QF136), the National Natural Science Foundation of China (Grant No. 62302277).

## Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Bandyopadhyay, S., and Mallick, K. (2016). A new feature vector based on gene ontology terms for protein-protein interaction prediction. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 14, 762–770. doi:10.1109/TCBB.2016.2555304

Bateman, A. (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Boldridge, W. C., Ljubetič, A., Kim, H., Lubock, N., Szilágyi, D., Lee, J., et al. (2023). A multiplexed bacterial two-hybrid for rapid characterization of protein–protein interactions and iterative protein design. *Nat. Commun.* 14, 4636. doi:10.1038/s41467-023-38697-x

Brown, K. R., and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics* 21, 2076–2082. doi:10.1093/bioinformatics/bti273

Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved prediction of protein-protein interactions using alphafold2. *Nat. Commun.* 13, 1265. doi:10.1038/s41467-022-28865-w

Chen, M., Ju, C. J.-T., Zhou, G., Chen, X., Zhang, T., Chang, K.-W., et al. (2019). Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics* 35, i305–i314. doi:10.1093/bioinformatics/btz328

Chen, S.-Y., and Zacharias, M. (2023). What makes a good protein–protein interaction stabilizer: analysis and application of the dual-binding mechanism. *ACS Central Sci.* 9, 969–979. doi:10.1021/acscentsci.3c00003

Chen, X.-w., and Jeong, J. C. (2009). Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics* 25, 585–591. doi:10.1093/bioinformatics/btp039

Chen, X.-W., and Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics* 21, 4394–4400. doi:10.1093/bioinformatics/bti721

Chi, X., Yan, R., Zhang, J., Zhang, G., Zhang, Y., Hao, M., et al. (2020). A neutralizing human antibody binds to the n-terminal domain of the spike protein of sars-cov-2. *Science* 369, 650–655. doi:10.1126/science.abc6952

Ding, Z., and Kihara, D. (2018). Computational methods for predicting protein-protein interactions using various protein features. *Curr. Protoc. protein Sci.* 93, e62. doi:10.1002/cpps.62

Dong, S., Lau, V., Song, R., Ierullo, M., Esteban, E., Wu, Y., et al. (2019). Proteome-wide, structure-based prediction of protein-protein interactions/new molecular interactions viewer. *Plant physiol.* 179, 1893–1907. doi:10.1104/pp.18.01216

Gainza, P., Wehrle, S., Van Hall-Beauvais, A., Marchand, A., Scheck, A., Harteveld, Z., et al. (2023). *De novo* design of protein interactions with learned surface fingerprints. *Nature* 617, 176–184. doi:10.1038/s41586-023-05993-x

Gao, Z., Jiang, C., Zhang, J., Jiang, X., Li, L., Zhao, P., et al. (2023). Hierarchical graph learning for protein–protein interaction. *Nat. Commun.* 14, 1093. doi:10.1038/s41467-023-36736-1

Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids Res.* 36, 3025–3030. doi:10.1093/nar/gkn159

Hu, X., Feng, C., Zhou, Y., Harrison, A., and Chen, M. (2022). Deeptrio: a ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks. *Bioinformatics* 38, 694–702. doi:10.1093/bioinformatics/btab737

Jha, K., Saha, S., and Singh, H. (2022). Prediction of protein–protein interaction using graph neural networks. *Sci. Rep.* 12, 8360. doi:10.1038/s41598-022-12201-9

Kjer-Hansen, P., and Weatheritt, R. J. (2023). The function of alternative splicing in the proteome: rewiring protein interactomes to put old functions into new contexts. *Nat. Struct. Mol. Biol.* 30, 1844–1856. doi:10.1038/s41594-023-01155-9

Li, B.-Q., Feng, K.-Y., Chen, L., Huang, T., and Cai, Y.-D. (2012). *Prediction of protein-protein interaction sites by random forest algorithm with mrmr and ifs.*

Li, Y., Li, P., Yang, X., Hsieh, C.-Y., Zhang, S., Wang, X., et al. (2021). Introducing block design in graph neural networks for molecular properties prediction. *Chem. Eng. J.* 414, 128817. doi:10.1016/j.cej.2021.128817

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Adv. neural Inf. Process. Syst.* 26. doi:10.48550/arXiv.1310.4546

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., et al. (2019). The biogrid interaction database: 2019 update. *Nucleic acids Res.* 47, D529–D541. doi:10.1093/nar/gky1079

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., et al. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 13, 2363–2371. doi:10.1101/gr.1680803

Reed, T. J., Tyl, M. D., Tadych, A., Troyanskaya, O. G., and Cristea, I. M. (2024). Tapioca: a platform for predicting *de novo* protein–protein interactions in dynamic contexts. *Nat. Methods* 21, 488–500. doi:10.1038/s41592-024-02179-9

Rodina, A., Xu, C., Digwal, C. S., Joshi, S., Patel, Y., Santhaseela, A. R., et al. (2023). Systems-level analyses of protein-protein interaction network dysfunctions via epichaperomics identify cancer-specific mechanisms of stress adaptation. *Nat. Commun.* 14, 3742. doi:10.1038/s41467-023-39241-7

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shen, L., Feng, H., Qiu, Y., and Wei, G.-W. (2023). Svsbi: sequence-based virtual screening of biomolecular interactions. *Commun. Biol.* 6, 536. doi:10.1038/s42003-023-04866-3

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids Res.* 47, D607–D613. doi:10.1093/nar/gky1131

Tang, T., Zhang, X., Liu, Y., Peng, H., Zheng, B., Yin, Y., et al. (2023). Machine learning on protein–protein interaction prediction: models, challenges and trends. *Briefings Bioinforma.* 24, bbad076. doi:10.1093/bib/bbad076

Wang, Y., Li, Z., Zhang, Y., Ma, Y., Huang, Q., Chen, X., et al. (2021). Performance improvement for a 2d convolutional neural network by using ssc encoding on protein–protein interaction tasks. *BMC Bioinforma.* 22, 184–216. doi:10.1186/s12859-021-04111-w

Wong, L., You, Z.-H., Ming, Z., Li, J., Chen, X., and Huang, Y.-A. (2015). Detection of interactions between proteins through rotation forest and local phase quantization descriptors. *Int. J. Mol. Sci.* 17, 21. doi:10.3390/ijms17010021

Wu, D., Li, Y., Zheng, L., Xiao, H., Ouyang, L., Wang, G., et al. (2023). Small molecules targeting protein–protein interactions for cancer therapy. *Acta Pharm. Sin. B* 13, 4060–4088. doi:10.1016/j.apsb.2023.05.035

Xia, J.-F., Han, K., and Huang, D.-S. (2010). Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. *Protein Peptide Lett.* 17, 137–145. doi:10.2174/092986610789909403

Yang, Z., Zhao, L., Wu, S., and Chen, C. Y.-C. (2021). Lung lesion localization of covid-19 from chest ct image: a novel weakly supervised learning method. *IEEE J. Biomed. Health Inf.* 25, 1864–1872. doi:10.1109/JBHI.2021.3067465

Yao, Y., Du, X., Diao, Y., and Zhu, H. (2019). An integration of deep learning with feature embedding for protein–protein interaction prediction. *PeerJ* 7, e7126. doi:10.7717/peerj.7126

You, Z.-H., Chan, K. C., and Hu, P. (2015). Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PloS one* 10, e0125811. doi:10.1371/journal.pone.0125811

Zahiri, J., Yaghoubi, O., Mohammad-Noori, M., Ebrahimpour, R., and Masoudi-Nejad, A. (2013). Ppievo: protein–protein interaction prediction from pssm based evolutionary information. *Genomics* 102, 237–242. doi:10.1016/j.ygeno.2013.05.006

Zhou, C., Yu, H., Ding, Y., Guo, F., and Gong, X.-J. (2017). Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree. *PLoS One* 12, e0181426. doi:10.1371/journal.pone.0181426

# Frontiers in
# Genetics

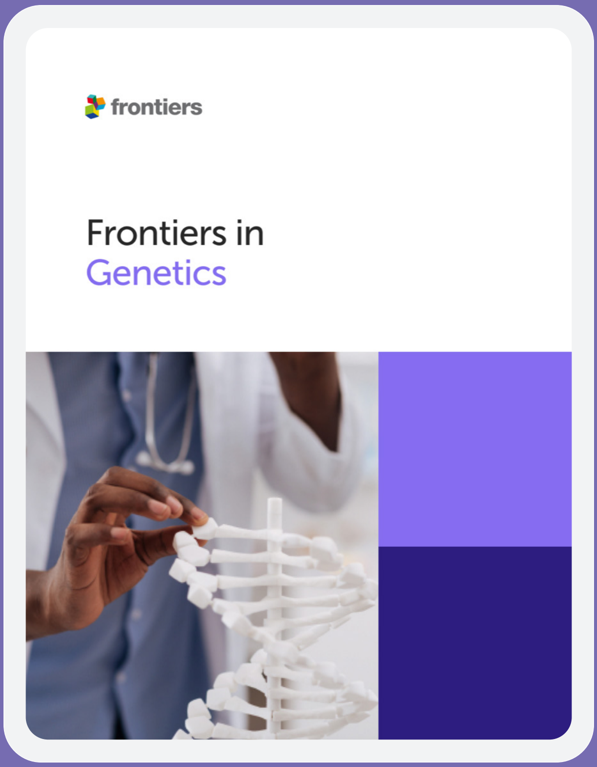
**Highlights genetic and genomic inquiry relating to all domains of life**

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

---

## Discover the latest Research Topics

See more →



frontiers

Frontiers in
Genetics

**Frontiers**

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

**Contact us**

+41 (0)21 510 17 00
frontiersin.org/about/contact

frontiers | Research Topics