Briefings in Bioinformatics, 2025, **26(1)**, bbae658
https://doi.org/10.1093/bib/bbae658
Problem Solving Protocol

MORE: a multi-omics data-driven hypergraph integration network for biomedical data classification and biomarker identification

Yuhan Wang[1,‡], Zhikang Wang[2,‡], Xuan Yu[3], Xiaoyu Wang[2], Jiangning Song[2,4,*], Dong-Jun Yu[1,*], Fang Ge[5,*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei, Nanjing 210094, China
[2]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Wellington Rd, Clayton, Melbourne, VIC 3800, Australia
[3]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong 999077, China
[4]Monash Data Futures Institute, Monash University, Wellington Rd, Clayton, Melbourne, VIC 3800, Australia
[5]State Key Laboratory of Organic Electronics and Information Displays & Institute of Advanced Materials (IAM), Nanjing University of Posts & Telecommunications, 9 Wenyuan, Nanjing 210023, China
*Corresponding authors. Fang Ge. E-mail: gfang0616@njupt.edu.cn; Dong-Jun Yu. E-mail: njyudj@njust.edu.cn; Jiangning Song. E-mail: jiangning.song@monash.edu
‡Yuhan Wang and Zhikang Wang contributed equally to this work and should be considered as co-first authors.

Abstract

High-throughput sequencing methods have brought about a huge change in omics-based biomedical study. Integrating various omics data is possibly useful for identifying some correlations across data modalities, thus improving our understanding of the underlying biological mechanisms and complexity. Nevertheless, most existing graph-based feature extraction methods overlook the complementary information and correlations across modalities. Moreover, these methods tend to treat the features of each omics modality equally, which contradicts current biological principles. To solve these challenges, we introduce a novel approach for integrating multi-omics data termed Multi-Omics hypeRgraph integration nEtwork (MORE). MORE initially constructs a comprehensive hyperedge group by extensively investigating the informative correlations within and across modalities. Subsequently, the multi-omics hypergraph encoding module is employed to learn the enriched omics-specific information. Afterward, the multi-omics self-attention mechanism is then utilized to adaptatively aggregate valuable correlations across modalities for representation learning and making the final prediction. We assess MORE's performance on datasets characterized by message RNA (mRNA) expression, Deoxyribonucleic Acid (DNA) methylation, and microRNA (miRNA) expression for Alzheimer's disease, invasive breast carcinoma, and glioblastoma. The results from three classification tasks highlight the competitive advantage of MORE in contrast with current state-of-the-art (SOTA) methods. Moreover, the results also show that MORE has the capability to identify a greater variety of disease-related biomarkers compared to existing methods, highlighting its advantages in biomedical data mining and interpretation. Overall, MORE can be investigated as a valuable tool for facilitating multi-omics analysis and novel biomarker discovery. Our code and data can be publicly accessed at https://github.com/Wangyuhanxx/MORE.

Keywords: comprehensive hyperedge group; multi-omics hypergraph encoding module; multi-omics self-attention mechanism; identify disease-related biomarkers

Introduction

Significant advancements in diverse high-throughput sequencing methods, e.g. DNA nanosphere sequencing, mRNA expression (mRNA), DNA methylation (meth), and miRNA expression (miRNA), have fundamentally revolutionized various biological analyses and facilitated biological discoveries [1, 2]. Current research has demonstrated diverse omics data contains both shared and unique knowledge of various biological processes [3, 4]. Combining multi-omics data enables deeper insights of analyses, leading to better clinical decisions and improved disease treatment [5, 6]. Several studies also highlight that multi-omics integration enhances the accuracy of disease prediction when compared to single-omics approaches [7–9]. Although a variety of approaches have been proposed, most of them only consider the commonalities, ignoring the complementarities across omics modalities. Consequently, there is a pressing need for an effective integration method to analyze and explore interactive and complementary information in multiple omics data.

Existing multi-omics integration approaches are categorized into unsupervised and supervised learning ones. Unsupervised learning approaches encode integrated multiple omics data into low-dimensional feature embeddings to perform classification and clustering tasks [10, 11]. However, due to the lack of labeled information for supervised training, the experiment results are usually unsatisfactory. In recent years, numerous studies have focused on employing supervised learning to explore various biological processes and mechanisms. For instance, Van De Wiel *et al.* [7] proposed an adaptive group-regularized ridge regression technique that incorporates methylation microarray data alongside curated annotations for the classification of cervical cancer. Moreover, Singh *et al.* [3] proposed Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO),

Received: August 19, 2024. **Revised:** November 18, 2024. **Accepted:** December 4, 2024

which extends sparse generalized canonical correlation analyses to supervised learning. DIABLO is capable of both discriminating different phenotypic groups and identifying the common information across different omics modalities. However, the simple linear correlation methods do not apply to complex disease studies.

Over the past few years, deep learning techniques have been widely utilized in integrating multiple omics data because of its powerful capability to learn nonlinear relationships across modalities [12, 13]. For example, Wang *et al.* [14] proposed Multi-Omics Graph cOnvolutional NETworks (MOGONET), which uses independent graph convolutional networks (GCNs) [15] to encode various omics datasets. Notably, it constructs a multiple omics intersection tensor to explore cross-omics correlations for effectively integrating multiple omics data. The limitation lies that an employed GCN can only model pairwise correlations between nodes, which is not optimal for biological analysis with complex correlations. Dong *et al.* [16] proposed Multi-Omics Graph Learning and Attention Mechanism (MOGLAM), which utilizes dynamic GCNs (FSDGCNs) to select the feature and omic-integrated representation learning (OIRL) to integrate multi-omics data. MOGLAM also builds an advanced sample similarity network, leading to the extraction of richer and more informative embeddings. Gong *et al.* [17] introduced Multi-Omics Attention Deep Learning Network (MOADLN), an approach that integrates multi-omics data using the self-attention mechanism [18] and a multi-omics association discovery network. However, MOGONET, MOGLAM, and MOADLN focus solely on the influence between the samples within each individual omics modality when constructing the sample similarity network. This strategy neglects the complementary information across different omics modalities, which can potentially result in inferior network performance.

Recently, hypergraphs have become an effective tool for modeling and exploring complicated correlations across different data modalities in numerous applications [19, 20]. Distinct from previous graph representations [15, 21], hypergraphs can utilize degree-free hyperedges to encode higher-order data correlations, which is a more efficient way to model complex correlations of nodes and enable more complex data analysis. To solve the aforementioned problems in multi-omics analysis, we introduce an innovative multi-omics integration approach, MORE. Specifically, MORE is a model comprising the Multi-Omics Hypergraph Encoding (MOHE) module to learn more representative omics-specific features and the Multi-Omics Self-Attention (MOSA) module to integrate valuable information across modalities to make the final prediction. Firstly, we constructed a hyperedge group for each omics modality by extensively mining the potential correlations within each omics modality. Subsequently, we fused the hyperedge groups from different modalities to construct a comprehensive hyperedge group. Next, the hypergraph, along with the features from each omics modality, was inputted into the MOHE module for omics-specific knowledge learning. Therefore, the MOHE module considers the correlations within and across different omics modalities, which could ensure a more discriminative omics-specific representation with richer information. Considering the varied contributions of different omics modalities to the final classification, the MOSA module was employed to adaptively integrate these features based on generated attention coefficients.

In the Introduction section, we proposed an innovative multi-omics integration method termed MORE. Then, in the Materials section, we introduced the datasets used and the data preprocessing method. In the Methods section, we described the construction of the different modules of the model in detail. In the Results and Discussions section, we demonstrated the effectiveness and potential applications of MORE based on extensive experiments. Benchmarking experiments demonstrated that MORE outperformed other multi-omics integration methods. Moreover, ablation experiments confirmed the essential contribution of the MOHE and MOSA modules to the performance of MORE. In the Identifying Biomarkers Using MORE section, we showed that MORE could identify important biomarkers relevant to biomedical problems, indicating its data mining and interpretation capabilities. Finally, we summarized the advantages and limitations of the model in the Conclusion section. Overall, MORE attains better performance than other existing advanced multiple omics integrated methods and can be potentially applied as a useful tool to facilitate community-wide efforts in multi-omics data analysis.

Table 1. Detailed information of the datasets in terms of disease categories and the number of samples.

Dataset	Categories	Number of features for training mRNA, meth, and miRNA
ROSMAP	NC: 169, AD: 182	200, 200, 200
BRCA	Normal-like: 105, Basal-like: 128, HER2-enriched: 44, Luminal A: 385, Luminal B: 146	1000, 1000, 503
GBM	Proneural: 66, Classical: 55, Mesenchymal: 66, Neural: 39	1000, 1000, 534

Materials

Datasets

Three datasets were used in this study: The Religious Orders Study and the Rush Memory and Aging Project (ROSMAP) for Alzheimer's disease (AD), normal control (NC) classification, invasive breast carcinoma (BRCA) for invasive breast carcinoma PAM50 subtype classification, and glioblastoma (GBM) for glioblastoma subtype classification. All datasets comprise omics data on mRNA, meth, and miRNA.

ROSMAP was retrieved from AMP-AD Knowledge Portal [14, 22], while BRCA was extracted from The Cancer Genome Atlas Program (TCGA) via Broad GDAC Firehose [14]. PAM50 identifies five molecular subtypes of breast cancer, including normal-like, basal-like, HER2-enriched, Luminal A, and Luminal B [23]. BRCA subtype data from PAM50 were accessed through TCGAbiolinks [24]. The GBM dataset was obtained from the benchmark cancer datasets [25]. The GBM dataset has four different subtypes including the proneural, classical, mesenchymal, and neural [26, 27]. A detailed description of the three datasets regarding disease categories and the sample amounts is provided in Table 1.

Preprocessing

Appropriate preprocessing of omics data is crucial to eliminate experimental errors and noise. First, we removed features with missing values (identified as NaN). Afterward, considering that probes for DNA methylation data might correspond to multiple genes, the probes corresponding to a single gene were reserved to ensure the data sensitivity. Subsequently, features with low variances or no signal were also filtered out. Particularly, different variance filtering thresholds were used for multiple omics data. For mRNA and meth data, the variance thresholds were set to 0.1 and 0.001, respectively. The miRNAs' amount in the expression data is much fewer than that in the other two modalities; thereby, only features with zero variance were filtered out.

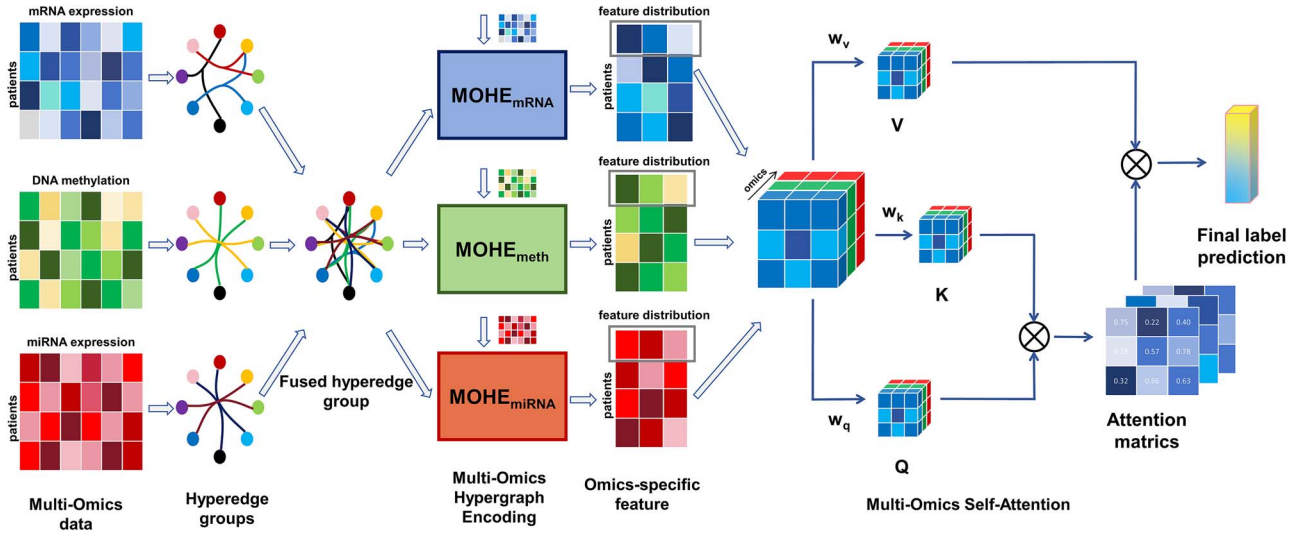

Figure 1. Overview of the proposed MORE for multi-omics analysis. The framework comprises the fused hyperedge group generation process, the multi-omics hypergraph encoding process, and the multi-omics self-attention process. The final prediction is based on the aggregated multi-omics features. For the binary classification task, the final prediction of the model represents AD and NC. For the multi-class classification tasks, the final prediction of the model represents different subtypes of the disease.

Although numerous preprocessing operations were employed, the high-dimensional omics data might still include needless information that could negatively impact the model performance. Consequently, the analysis of variance (ANOVA) was employed to further refine feature selection. ANOVA *F*-values were calculated for every omics data to determine the variance between categories. Subsequently, we selected the features that significantly varied between different categories. Eventually, we standardized all of omics data to [0,1].

Methods

This section introduces the proposed MORE method in detail. As illustrated in Fig. 1, MORE builds upon two main modules: MOHE and MOSA. The MOHE module is used to extract omics-specific knowledge and simultaneously reveal the correlations within and across omics modalities in the latent feature space. The MOSA module further efficiently integrates multi-omics features to generate the final prediction.

Multi-omics hypergraph construction and representation learning

The hypergraph neural network stands out in comparison with previous graph representations, which can effectively utilize degree-free hyperedges to encode complex correlations across data modalities [20, 28]. Generally, the hypergraph is represented as $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$, $\mathbf{V}$ and $\mathbf{E}$ denote the vertex and hyperedge set. The diagonal matrix of edge weights $\mathbf{W}$ assigns a weight to each hyperedge. $\mathbf{G}$ can be represented with a $|\mathbf{V}| \times |\mathbf{E}|$ incidence matrix $\mathbf{H}$, as follows:

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

where $v \in e$ denotes that node v is in the hyperedge e. For $v \in \mathbf{V}$, $d(v) = \sum_{e \in \mathbf{E}} w(e) h(v,e)$ represents its degree. For $e \in \mathbf{E}$, $d(e) = \sum_{v \in \mathbf{V}} h(v,e)$ represents its degree. Additionally, $\mathbf{D}_v$ and $\mathbf{D}_e$ represent the diagonal matrices of the vertex degrees and the edge degrees, respectively.

The following subsection introduces the generation process of $\boldsymbol{H}_{\text{mRNA}}$, $\boldsymbol{H}_{\text{meth}}$, and $\boldsymbol{H}_{\text{miRNA}}$ in detail. For each vertex, we computed its distance from all other vertices and constructed its hyperedge by connecting the k nearest ones. Here, k is an important hyper-parameter that represents the average amount of vertices connected by each hyperedge (also itself). If the two vertices are connected, the value of the corresponding position in the incidence matrix will be defined as 1, otherwise 0. We obtained $\boldsymbol{H}_{\text{mRNA}}$, $\boldsymbol{H}_{meth}$, and $\boldsymbol{H}_{\text{miRNA}}$ for the three omics modalities in this method. As shown in Fig. 2, a hypergraph can jointly employ multi-omics correlations for hyperedge group fusion by combining the incidence matrices [20]. To create the final multi-omics hypergraph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$, we concatenated the three incidence matrices from the three omics modalities. The global incidence matrix $\mathbf{H} \in R^{N \times 3N}$ can be represented as follows:

$$\mathbf{H} = \text{Concat}\left[\boldsymbol{H}_{\text{mRNA}}, \boldsymbol{H}_{\text{meth}}, \boldsymbol{H}_{\text{miRNA}}\right] \tag{2}$$

Furthermore, $\mathbf{G} \in R^{N \times N}$ can be defined as follows:

$$\mathbf{G} = \mathbf{D}_v^{-1/2} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^{\top} \mathbf{D}_v^{-1/2} \tag{3}$$

Eventually, we can construct a hypergraph convolutional layer $f(\mathbf{X}, \mathbf{W}, \Theta)$:

$$\mathbf{X}^{(l+1)} = \sigma\left(\mathbf{D}_v^{-1/2} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^{\top} \mathbf{D}_v^{-1/2} \mathbf{X}^{(l)} \Theta^{(l)}\right) \tag{4}$$

where $\mathbf{X}^{(l)} \in R^{N \times C}$ represents the input feature of a hypergraph at an l layer, $\mathbf{X}^{(0)} = \mathbf{X}$, Θ optimized throughout the training phase, σ is the nonlinear activation function.

Multi-omics integration learning

Existing graph-based multi-omics integration methods generally treat features from different modalities equally. Nevertheless, various omics modalities may offer unique contributions to different biological analysis processes [16]. Given the significant capability of the self-attention mechanism in representation learning

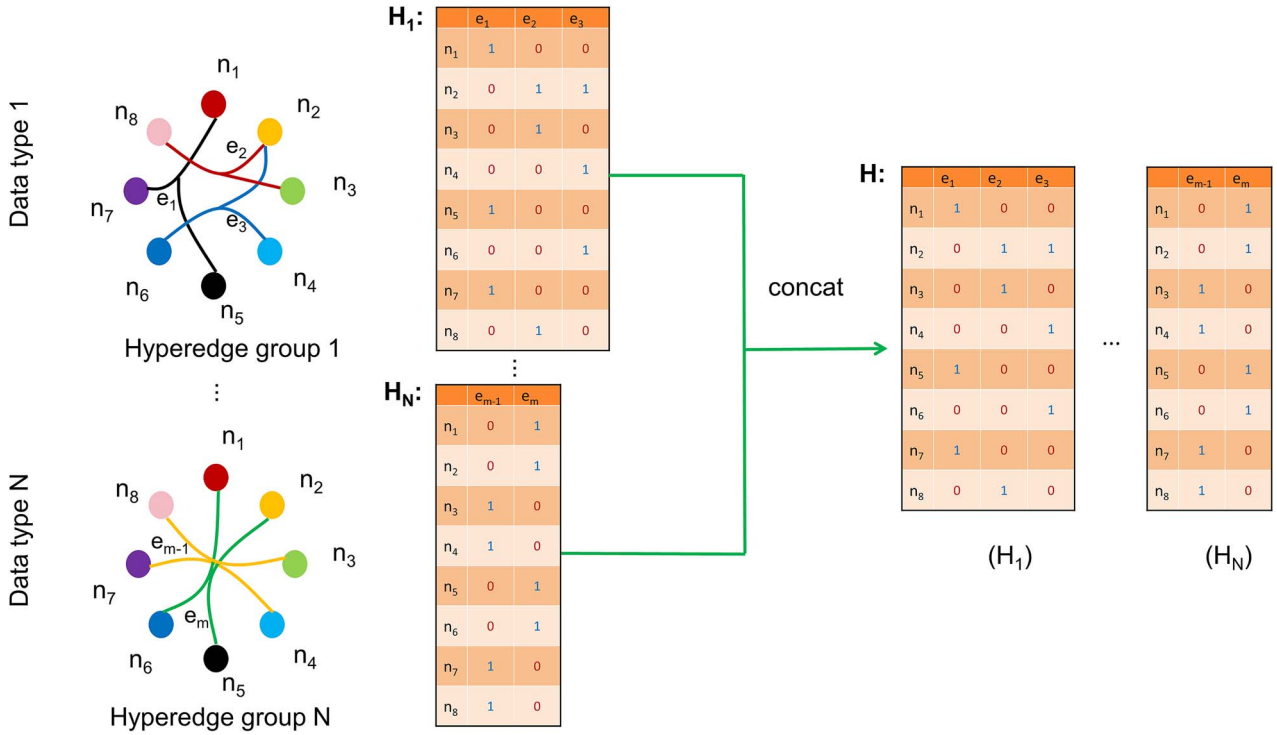

Figure 2. Fusion of hyperedge groups by combining the incidence matrices across modalities.

[29–32], we employed it to adaptatively establish the correlations across omics modalities.

For the patient j, its feature matrix is represented as $\hat{\mathbf{X}}_j = \left[\hat{\mathbf{x}}_j^{(1)}, \hat{\mathbf{x}}_j^{(2)}, ..., \hat{\mathbf{x}}_j^{(I)}\right]$, where $\hat{\mathbf{x}}_j^{(i)}$ denotes the feature vector of the i-th omics. Subsequently, the Query, Key, and Value can be generated by $\mathbf{Q}_j = w_q\hat{\mathbf{X}}_j + b_q = \left[q_j^{(1)}, q_j^{(2)}, ..., q_j^{(I)}\right]$, $\mathbf{K}_j = w_k\hat{\mathbf{X}}_j + b_k = \left[k_j^{(1)}, k_j^{(2)}, ..., k_j^{(I)}\right]$, and $\mathbf{V}_j = w_v\hat{\mathbf{X}}_j + b_v = \left[v_j^{(1)}, v_j^{(2)}, ..., v_j^{(I)}\right]$, respectively; w and b correspond to the optimizable parameters. The attention matrix among multi-omics can be defined as:

$$\mathbf{A}_j(m,n) = \frac{\exp\left[q_j^{(m)} \cdot \left(k_j^{(n)}\right)^\top / \sqrt{d_f}\right]}{\sum_{n=1}^{I} \exp\left[q_j^{(m)} \cdot \left(k_j^{(n)}\right)^\top / \sqrt{d_f}\right]} \tag{5}$$

where $\mathbf{A}_j(m,n)$ indirectly reflect the correlations between m-th and n-th omics in sample$j$. With the attention matrix $\mathbf{A}_j$, the final aggregation process can be formulated as follows:

$$\hat{\mathbf{V}}_j = \mathbf{A}_j \cdot \mathbf{V}_j^\top \tag{6}$$

The multihead operation was also utilized in this feature learning process, which aims to achieve an enriched feature collection from multiple perspectives. The final representation can be formulated as follows:

$$\hat{\mathbf{V}}_j^{cat} = Concat\left[\hat{\mathbf{V}}_j^{(1)}, \hat{\mathbf{V}}_j^{(2)}, ..., \hat{\mathbf{V}}_j^{(Z)}\right] \tag{7}$$

where Z refers to the number of heads. Eventually, a linear classification layer was employed to classify the patients using the generated representations.

Model optimization

To enhance omics-specific information and improve the overall network performance, we introduced a two-step learning strategy with the first step only optimizing the omics-specific MOHEs and the second step optimizing the whole network. The cross-entropy loss function was adopted for optimization. In the first training step, the loss function tailored to the i-th omics modality can be formulated as follows:

$$L_{\text{omics-specific}}^{(i)} = \sum_{j=1}^{N} L_{CE}\left(p_j^i, y_j\right) \tag{8}$$

where y_j represents true label, p_j^i is the prediction from the omics-specific layer, while N refers to the sample amounts, respectively. During the second phase, the total loss function is shown as follows:

$$L = \sum_{i=1}^{3} L_{\text{omics-specific}}^{(i)} + L_{\text{multi-omics}} \tag{9}$$

To demonstrate the superiority of the training strategy, we also performed experiments using an end-to-end training strategy. Experimental results and details are shown in the section Model Performance Based on Different Training Strategies.

Results and discussions

Evaluation metrics

In this study, each dataset was partitioned into nonoverlapping two subsets: 70% training and 30% testing. For the binary classification task, the performance was assessed using multiple performance metrics including accuracy (ACC), F1 score (F1), and the area under the receiver operating characteristic curve (AUC). For the multi-class classification tasks, we utilized ACC, the weighted F1 (F1_weighted), and the macro-averaged F1 (F1_macro). There is a thorough description of the performance assessment metrics in

Table 2. Classification performance of all methods on three datasets.

Dataset		Evaluation metrics		
ROSMAP	**Method**	**ACC**	**F1**	**AUC**
	KNN	0.661 ± 0.038	0.682 ± 0.029	0.701 ± 0.037
	SVM	0.787 ± 0.015	0.791 ± 0.017	0.790 ± 0.016
	RF	0.735 ± 0.029	0.741 ± 0.030	0.815 ± 0.029
	XGBoost	0.775 ± 0.039	0.784 ± 0.040	0.846 ± 0.036
	NN	0.771 ± 0.012	0.779 ± 0.014	0.842 ± 0.017
	MOGONET	0.814 ± 0.022	0.819 ± 0.021	0.882 ± 0.017
	MOGLAM	0.816 ± 0.014	0.822 ± 0.013	0.885 ± 0.017
	MOADLN	0.816 ± 0.014	0.823 ± 0.019	0.886 ± 0.018
	MORE	0.829 ± 0.018	0.836 ± 0.017	0.903 ± 0.010
BRCA	**Method**	**ACC**	**F1_weighted**	**F1_macro**
	KNN	0.741 ± 0.013	0.710 ± 0.018	0.662 ± 0.023
	SVM	0.737 ± 0.017	0.696 ± 0.021	0.637 ± 0.028
	RF	0.755 ± 0.010	0.741 ± 0.011	0.661 ± 0.013
	XGBoost	0.779 ± 0.007	0.771 ± 0.007	0.714 ± 0.012
	NN	0.770 ± 0.020	0.754 ± 0.031	0.690 ± 0.032
	MOGONET	0.815 ± 0.017	0.812 ± 0.016	0.742 ± 0.023
	MOGLAM	0.819 ± 0.012	0.813 ± 0.014	0.750 ± 0.014
	MOADLN	0.822 ± 0.018	0.816 ± 0.018	0.755 ± 0.019
	MORE	0.835 ± 0.020	0.820 ± 0.023	0.768 ± 0.021
GBM	**Method**	**ACC**	**F1_weighted**	**F1_macro**
	KNN	0.665 ± 0.033	0.618 ± 0.031	0.555 ± 0.037
	SVM	0.702 ± 0.019	0.637 ± 0.018	0.580 ± 0.017
	RF	0.670 ± 0.023	0.643 ± 0.026	0.590 ± 0.030
	XGBoost	0.709 ± 0.026	0.692 ± 0.027	0.634 ± 0.029
	NN	0.707 ± 0.028	0.694 ± 0.023	0.667 ± 0.024
	MOGONET	0.740 ± 0.025	0.727 ± 0.021	0.702 ± 0.021
	MOGLAM	0.741 ± 0.024	0.733 ± 0.023	0.725 ± 0.023
	MOADLN	0.743 ± 0.022	0.734 ± 0.025	0.727 ± 0.019
	MORE	0.762 ± 0.027	0.755 ± 0.025	0.736 ± 0.024

Text S1. The implementation details of MORE are given in Text S2. The experiment was conducted five times on each dataset, with the average of these five trials reported as the final performance.

Comparison with other multi-omics integration methods

We evaluated MORE with eight superior multi-omics integration approaches: (i) K-Nearest Neighbors (KNN) [33, 34], (ii) Support Vector Machine (SVM) [35, 36], (iii) Random Forest (RF) [37, 38], (iv) eXtreme Gradient Boosting (XGBoost) [39, 40], (v) fully connected Neural Network (NN) [41, 42], (vi) MOGONET [14], (vii) MOGLAM [16], and (viii) MOADLN [17]. The preprocessed multi-omic data were directly concatenated and utilized as input for KNN, SVM, RF, XGBboost, and NN.

From Table 2, we concluded that MORE attained the best performance on three classification tasks compared to other superior multiple omics integration approaches. For example, on ROSMAP, AUC of MORE was 1.7% higher than the second-best MOADLN. On BRCA, ACC of MORE was 1.3% higher than the second-best MOADLN. It is noteworthy that on the GBM dataset, even with small-size training samples, MORE still outperformed other machine learning and deep learning approaches, highlighting its generalization and robustness capability.

Ablation study

Herein, we evaluated and discussed the performance of proposed modules during the multi-omics analysis process. Extensive ablation experiments were performed to compare the performance of MORE with its three variants: (i) NN_NN: NNs for multiple omics feature learning and integration. (ii) NN_MOSA: NN for multiple omics feature learning and MOSA for integration. (iii) MOHE_NN: MOHE for multiple omics feature learning and NN for integration. In the ablation experiments, all NNs had the same number of layers as the replaced modules.

From Table 3 and Table S1, it was evident that MORE performed better than all three model variants in every task. For example, ACC of MORE was 0.9% higher than that of MOHE_NN, 3.5% higher than that of NN_MOSA, and 4.3% higher than that of NN_NN on the ROSMAP dataset. Although the ACC of MOHE_NN classification tasks was close to that of MORE on the ROSMAP dataset, MORE consistently yielded better metrics than MOHE_NN across all metrics for all tasks. Notably, MOHE_NN performed better than NN_MOSA and NN_NN in every task, highlighting the importance and effectiveness of MOHE for multi-omics feature learning. This indicates that MOHE can comprehensively utilize the complementary information across different omics modalities. This capability is underpinned by the fused hyperedge groups in the hypergraph learning process. Furthermore, after adding MOSA on MOHE_NN and NN_NN, the performance can be further improved on all three tasks, which indicates the effectiveness of MOSA. For example, on the BRCA dataset, the ACCs of MORE and NN_MOSA were 1.2% and 0.5% higher than those of MOHE_NN and NN_NN, respectively. It is noteworthy that the performance metrics of NN_MOSA and NN_NN were very close to each other on the GBM dataset. One possible reason is that the smaller number of disease categories and sample size may lead to the

Table 3. Ablation study of the key components on the ROSMAP and BRCA datasets.

Dataset		Evaluation metrics		
ROSMAP	**Method**	**ACC**	**F1**	**AUC**
	NN_NN	0.786 ± 0.022	0.790 ± 0.019	0.855 ± 0.019
	NN_MOSA	0.794 ± 0.020	0.799 ± 0.021	0.864 ± 0.022
	MOHE_NN	0.820 ± 0.020	0.825 ± 0.021	0.892 ± 0.021
	MORE	0.829 ± 0.018	0.836 ± 0.017	0.903 ± 0.010
BRCA	**Method**	**ACC**	**F1_weighted**	**F1_macro**
	NN_NN	0.797 ± 0.019	0.777 ± 0.021	0.726 ± 0.017
	NN_MOSA	0.802 ± 0.021	0.786 ± 0.024	0.731 ± 0.022
	MOHE_NN	0.823 ± 0.017	0.818 ± 0.020	0.757 ± 0.019
	MORE	0.835 ± 0.020	0.820 ± 0.023	0.768 ± 0.021

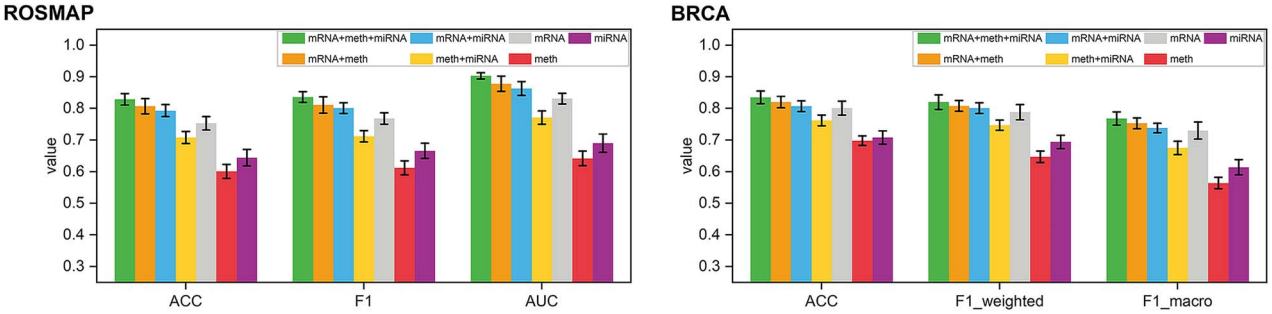

Figure 3. Classification performance with multi-omics data and single-omics data on the ROSMAP and BRCA datasets.

amplification of noise or errors in the absence of MOHE, thus potentially impacting the performance of MOSA. Nevertheless, the ablation study indicated that all the proposed modules in MORE are effective and can work synergistically.

Model performance under different omics settings

Within this subsection, we utilized three kinds of omics data: mRNA, meth, and miRNA. To verify the necessity of multiple omics integration for promoting model effectiveness, we particularly performed experiments under mRNA + meth + miRNA, mRNA + meth, mRNA + miRNA, meth + miRNA, mRNA, meth, and miRNA settings. As shown in Fig. 3 and Fig. S1, MORE integrating three kinds of omics data outperformed all other models in the aspect of ACC, F1, and AUC. Generally, MORE integrating two kinds of omics data outperformed that with only one kind of data. An exception was observed that the model integrating meth and miRNA performed worse than that using only mRNA, possibly due to the significant role of mRNA in the classification task. It is worth noting that the model integrating mRNA and meth on ROSMAP and BRCA dataset performed second best after integrating three kinds of omics data. However, on the GBM dataset, the model integrating mRNA and miRNA ranked the second best, suggesting that different classification tasks might require different combinations of omics data. Nevertheless, MORE that integrated all three kinds of omics data always attained the best performance in all tasks, further demonstrating the necessity and significance of multi-omics integration in various biological analyses.

Model performance with different hyper-parameter

A crucial hyper-parameter in MORE is k, denoting the average number of nodes connected by each hyperedge. When constructing a hypergraph, if only a small amount of vertices are connected by each hyperedge, the hyperedge groups may become too sparse, thereby missing important connections among samples. Conversely, if each hyperedge connects too many vertices, the hyperedge groups may become too dense, potentially introducing noise into the correlation analysis among samples. Therefore, the choice of the proper value of k is crucial for model performance. However, the optimal k value depends on the topology of the dataset and varies across different datasets. To assess the effect of k on MORE, we tested the proposed model on three datasets with different k values. As can be seen from Fig. 4 and Fig. S2, different values of k resulted in various performances. Particularly, on the ROSMAP dataset, MORE performed best when $k = 2$, while on the BRCA and the GBM datasets, it achieved the best performance based on $k = 3$. A possible explanation is that the AD classification task is a binary classification task with a relatively small amount of disease categories, thus requiring a smaller optimal k value than multi-class classification tasks. In addition, the small sample size may also have an influence on the optimal k value.

Model performance based on different training strategies

Within this subsection, we utilized a two-step training strategy. During the pretraining phase, we trained each individual omics-specific MOHE module separately to initialize the model parameters. Subsequently, during the formal training process, we trained both MOHE and MOSA modules for final classification using multi-omics data. To prove the superiority of our strategy, we compared the performance based on the end-to-end training strategy with that of the two-step training strategy. From Table 4 and Table S2, we discovered that our training strategy performed better than end-to-end training. For example, the ACC of MORE on the ROSMAP dataset was 1.4% higher than that of the

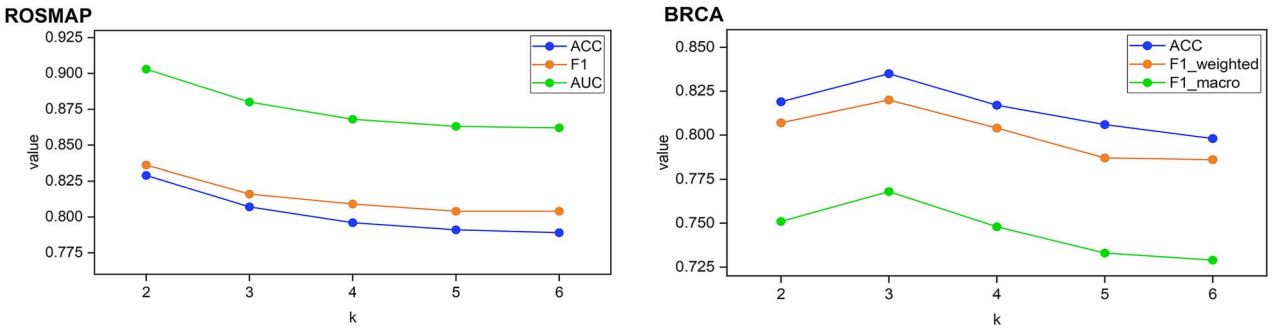

Figure 4. Classification performance of MORE with different hyper-parameters on the ROSMAP and BRCA datasets.

Table 4. Classification performance with different training strategies on the ROSMAP and BRCA datasets.

Dataset		Evaluation metrics		
ROSMAP	**Strategy**	**ACC**	**F1**	**AUC**
	End-to-end	0.815 ± 0.023	0.819 ± 0.023	0.885 ± 0.017
	Two-step (ours)	0.829 ± 0.018	0.836 ± 0.017	0.903 ± 0.010
BRCA	**Strategy**	**ACC**	**F1_weighted**	**F1_macro**
	End-to-end	0.819 ± 0.017	0.810 ± 0.020	0.747 ± 0.025
	Two-step (ours)	0.835 ± 0.020	0.820 ± 0.023	0.768 ± 0.021

end-to-end training model. Moreover, on the BRCA dataset, all the performance metrics of MORE were higher than the end-to-end training ones, especially on F1_macro with a 2.1% improvement. These results adequately demonstrated the feasibility and effectiveness of our training strategy.

Identifying biomarkers using MORE

Biomarkers are defined as biochemical indicators for physiological, pathological, or therapeutic processes that can be objectively measured and evaluated [43, 44]. Accordingly, identifying biomarkers is critical for interpreting the trained deep neural networks and understanding the underlying molecular mechanisms [45, 46]. Within this subsection, we first introduce the employed theory for biomarker identification following with the discussion of the results in detail.

Methods for identifying biomarkers

As the values of input features to MORE were scaled to $[0, 1]$ during data preprocessing, we could eliminate particular features by changing its value to 0. In this way, the effect of that particular feature on the predictive performance could be measured by calculating the accuracy fluctuation after its removal. This method has been widely used in neural networks for feature importance grading [47, 48]. Through this method, we explored the importance of features from different kinds of omics data. When the classification performance decreases significantly after eliminating a feature, it is considered an important biomarker in this task. We utilized F1 and F1_macro to assess the performance of MORE for binary and multi-class classification tasks, respectively. We ran the experiment on the dataset five times and summed the classification performance decrease across these repetitions for each feature as the final indicator to ensure our results were reliable. Tables 5 and 6 provide the list of the identified top 10 biomarkers for each omics modality from ROSMAP and BRCA. GBM (used as a proof-of-concept) is not further analyzed for detailed biomarker identification. In order to prove the reliability of MORE, the inner product regularization [16] was also applied to the feature indicator matrix for selecting critical biomarkers. The outcomes, presented in Tables S3 and S4, demonstrate a high degree of consistency between the biomarkers identified by the two methods, confirming the robustness of MORE. An additional comprehensive explanation is provided in Text S3.

For the most significant genes in mRNA and meth data, we conducted gene set functional enrichment analysis using the

Table 5. Important biomarkers identified by MORE on the ROSMAP dataset.

Omics type	Top 10 important biomarkers
mRNA expression	ARRDC2, CDK18, KIF5A, CXCR4, NPNT, LNCBRM, CNN3-DT, QDPR, TCEA3, APLN
DNA methylation	MBOAT7, DNAJC16, TMC4, PCDH12, CCL3, ABCB5, FGD4, RAB34, AGA, TM4SF18
miRNA expression	hsa-miR-132, hsa-miR-146b-5p, hsa-miR-33a, hsa-miR-129-5p, hsa-miR-UL70-3p, hsa-miR-143, hsa-miR-133a, hsa-miR-129-3p, hsa-miR-374a, hsa-miR-640

Table 6. Important biomarkers identified by MORE on the BRCA dataset.

Omics type	Top 10 important biomarkers
mRNA expression	SOX11, GART, NRTN, PGBD5, PI3, AKR1E2, SLC6A14, KRT6B, CPA4, MASTL
DNA methylation	ADAMTSL5, NFIL3, ATP10B, LIMK1, PABPC4L, TFF3, DLGAP5, PAPPA2, CAMK2N1, SNORD21
miRNA expression	hsa-mir-187, hsa-mir-205, hsa-mir-130b, hsa-mir-451, hsa-mir-215, hsa-mir-503, hsa-mir-1269, hsa-mir-577, hsa-mir-526b, hsa-mir-204

ToppGene Suite [49]. GO terms can be found through the ToppGene Suite. To make the results more convincing, we utilized the Benjamini–Hochberg procedure and reported the modified *P*-values. The biomarkers identified by MORE showed variation in their biological enrichment process and function across each dataset.

Identified biomarkers associated with Alzheimer's disease

For the biomarkers identified using mRNA expression data, significant enrichment was observed for GO terms relevant to CDK18 and APLN, such as cyclin-dependent protein kinase activity (GO: 0097472, $P = 1.012E-2$) and apelin receptor binding (GO: 0031704, $P = 2.062E-2$). According to earlier research, cyclin-dependent kinase 18 is bound by a cytosolic group of PLCβ, which facilitates tau phosphorylation and aggregation. Furthermore, by combining its catalytic activity and association with cyclin-dependent kinase 18, the PLCβ will lose and facilitate AD [50]. Additionally, Luo *et al.* [51] discovered that altering the amount of apelin can influence the course of neurodegenerative events like AD, indicating that apelin can become an ideal target to treat neurodegenerative diseases. Specifically, apelin's effects include the suppression of apoptosis, reduction of oxidative stress, inhibition of Ca2+ signaling, induction of autophagy, and suppression of inflammatory response. Furthermore, positive regulation of neuroinflammatory response (GO: 0150078, $P = 7.290E-2$) and positive regulation of microglial cell migration (GO: 1904141, $P = 1.861E-2$) were found to be significantly enriched for the biomarkers identified by DNA methylation data. Shao *et al.* [52] reported that due to an elevated concentration of neuroinflammatory cytokines in AD, neuroinflammatory plays a role in the pathophysiology of the disease. Additionally, research has demonstrated that the central nervous system's native innate immune cell population comprises microglia cells [53]. Microglial cell migration is thus closely linked to the progression of AD.

Moreover, in the AD patient classification task, some most significant biomarkers that were shown to be related to AD were also identified by MORE. Among the biomarkers identified by mRNA data, KIF5A has emerged as a possible candidate gene for regulating AD development. An essential component of the molecular machinery that facilitates anterograde axonal mitochondrial transport is kinesin-1, of which KIF5A is an isoform [54]. Studies have shown that brain mitochondrial defect is a significant feature of AD. Therefore, protecting the function of KIF5A is a potential treatment approach. Additionally, it is established that the chemokine CXC motif receptor4 (CXCR4) plays a role in the progression of AD. The etiology of AD involves complex factors, such as inflammation caused by microglia overactivation, and the expression of CXCR4 is elevated in astrocytes and microglia [55, 56]. For the biomarkers recognized by meth data, Hohman *et al.* [57] demonstrated an association between TMC4 and the development of AD. Yu *et al.* [58] reported a potential link between FGD4 and AD, verifying that FGD4 has a connection to actin cytoskeleton regulating mechanisms and may regulate synaptic loss in AD brain tissue. For the biomarkers identified by miRNA expression data, Nagaraj *et al.* [59] characterized different expressions of hsa-miR-33a in the plasma of AD and non-AD patients.

Identified biomarkers associated with breast cancer

For the biomarkers identified by mRNA data, we discovered some GO terms relevant to breast cancer were significantly enriched, including solute: sodium symporter activity (GO: 0015370, $P = 8.117E-3$) and positive regulation of osteoblast differentiation (GO: 0045669, $P = 2.267E-2$). It has been shown that sodium symporter protein is widely expressed in breast tumors, indicating the potential for breast cancer radiation treatment [60]. Additionally, Wu *et al.* [61] found that osteoblasts can deposit collagens to suppress Natural Killer (NK) cells through the inhibitory LAIR1 signaling and stimulate breast tumor colonization. Moreover, Adenosine Triphosate (ATP) dependent activity (GO: 0140657, $P = 5.500E-2$) was greatly enriched among the biomarkers identified by meth data. Studies have suggested that ATP-dependent activity is involved in breast cancer by showing the overexpression of ATPase phospholipid transporting 10B in breast cancer cells [62].

For invasive breast carcinoma subtype classification, there were also certain most significant biomarkers identified by MORE, which have proven to be associated with breast cancer. For the biomarkers identified by mRNA data, SOX11 has been shown to be associated with invasive cancer development. SOX11, usually inactive in mammary cells after birth, is expressed in estrogen receptor-negative Ductal Carcinoma In Situ (DCIS) lesions, particularly in basal-like clusters with increased aldehyde dehydrogenase activity and mammosphere formation capacity, and studies confirmed that SOX11 promotes the progression of DCIS to invasive cancer [63, 64]. In addition, Dunlap *et al.* [65] found that PI3 can promote the development of invasive breast carcinoma. For the biomarkers identified by meth data, TFF3 is strongly linked to invasive breast carcinoma. Studies showed that TFF3 is highly expressed in fibrocystic changes and papillomatous areas, with 89% expression in carcinomas *in situ* and 83% in invasive carcinomas [66, 67]. After analyzing tissue samples from individuals with invasive ductal carcinoma of the breast, Dietrich *et al.* [68] identified LIMK1 as a biomarker for invasive breast carcinoma. For the biomarkers identified by miRNA data, Gupta *et al.* [69] discovered that hsa-mir-503 is expressed in many types of tumors, like breast cancer and hepatocellular carcinoma. Their research demonstrated that hsa-mir-503 exerts its tumor-suppressing effect through its action on target genes. Xiao *et al.* [70] investigated the relationship between hsa-mir-205 and breast cancer and found that the gene has a tumor suppressor effect.

Conclusion

Recent advances in multi-omics sequencing techniques have enabled multiview characterization of various complex biological processes and diseases. Herein, we have proposed an innovative multi-omics integration method, termed MORE, for a more efficient and accurate multi-omics analysis. Specifically, MORE is developed based on two major modules MOHE and MOSA, which are designed to extract the omics-specific features and integrate the multiple omics information, respectively. The effectiveness of MORE was benchmarked on three classification tasks. The results showed that it significantly outperformed several existing integrated multi-omics approaches. The ablation study further demonstrated the significance and contribution of each proposed module during the analysis process. Furthermore, we found that the model combining three kinds of omics data had the best classification performance, proving the value of integrating diverse omics data. Additionally, MORE was effective in identifying potential biomarkers associated with various diseases, including AD, invasive breast carcinoma, and glioblastoma, and also had excellent capability in model interpretation. Despite the advantages, the proposed MORE method has certain limitations in terms of its capability of modeling other data modalities, e.g.

medical image data and clinical reports. In further research, we aim to design a more robust and comprehensive framework to facilitate disease-oriented multi-omics data processing.

Key Points

- We propose an innovative multi-omics integration method termed "MORE" to enable the integration of different omics data modalities for efficient analysis.
- MORE utilizes the MOHE module to effectively learn omics-specific features and the MOSA module to integrate valuable information across different data modalities, respectively.
- The core component of the MOHE module is a hypergraph, which can encode higher-order data correlations with its degree-free hyperedges. The MOHE module can simultaneously reveal correlations within and across omics modalities, and the MOSA module considers the varied contributions of diverse omics modalities to the last forecast.
- We verify the excellent performance of MORE in contrast with several current superior methods through extensive benchmarking experiments.
- We show the predictive power of MORE in effectively identifying disease-related biomarkers, highlighting its excellent biomedical data mining and interpretation capabilities.

Supplementary data

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

Funding

This work was supported by the National Natural Science Foundation of China (62372234, 62072243), the Natural Science Foundation of Jiangsu (BK20201304), Major and Seed Inter-Disciplinary Research project awarded by Monash University, and the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY223062).

References

1. He X, Liu X, Zuo F. *et al.* Artificial intelligence-based multi-omics analysis fuels cancer precision medicine. *Semin Cancer Biol* 2023;**88**:187–200. https://doi.org/10.1016/j.semcancer.2022.12.009.
2. Solari FA, Krahn D, Swieringa F. *et al.* Multi-omics approaches to study platelet mechanisms. *Curr Opin Chem Biol* 2023;**73**:102253. https://doi.org/10.1016/j.cbpa.2022.102253.
3. Singh A, Shannon CP, Gautier B. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019;**35**:3055–62. https://doi.org/10.1093/bioinformatics/bty1054.
4. Sathyanarayanan A, Mueller TT, Ali Moni M. *et al.* Multi-omics data integration methods and their applications in psychiatric disorders. *Eur Neuropsychopharmacol* 2023;**69**:26–46. https://doi.org/10.1016/j.euroneuro.2023.01.001.
5. Ning L, Zhou YL, Sun H. *et al.* Microbiome and metabolome features in inflammatory bowel disease via multi-omics integration analyses across cohorts. *Nat Commun* 2023;**14**:7135. https://doi.org/10.1038/s41467-023-42788-0.
6. Chong D, Jones NC, Schittenhelm RB. *et al.* Multi-omics integration and epilepsy: towards a better understanding of biological mechanisms. *Prog Neurobiol* 2023;**227**:102480. https://doi.org/10.1016/j.pneurobio.2023.102480.
7. van de Wiel MA, Lien TG, Verlaat W. *et al.* Better prediction by use of co-data: adaptive group-regularized ridge regression. *Stat Med* 2016;**35**:368–81. https://doi.org/10.1002/sim.6732.
8. Vandereyken K, Sifrim A, Thienpont B. *et al.* Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet* 2023;**24**:494–515. https://doi.org/10.1038/s41576-023-00580-2.
9. Baysoy A, Bai Z, Satija R. *et al.* The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023;**24**:695–713. https://doi.org/10.1038/s41580-023-00615-w.
10. Zhang Y, Kiryu H. MODEC: an unsupervised clustering method integrating omics data for identifying cancer subtypes. *Brief Bioinform* 2022;**23**:1–11. https://doi.org/10.1093/bib/bbac372.
11. Tini G, Marchetti L, Priami C. *et al.* Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief Bioinform* 2019;**20**:1269–79. https://doi.org/10.1093/bib/bbx167.
12. Han K, Wang J, Chu Y. *et al.* Deep learning based method for predicting DNA N6-methyladenosine sites. *Methods* 2024;**230**:91–8. https://doi.org/10.1016/j.ymeth.2024.07.012.
13. Li Z, Jiang M, Wang S. *et al.* Deep learning methods for molecular representation and property prediction. *Drug Discov Today* 2022;**27**:103373. https://doi.org/10.1016/j.drudis.2022.103373.
14. Wang T, Shao W, Huang Z. *et al.* MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun* 2021;**12**:3445. https://doi.org/10.1038/s41467-021-23774-w.
15. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations, ICLR 2017*, Toulon, France. 2016.
16. Ouyang D, Liang Y, Li L. *et al.* Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. *Comput Biol Med* 2023;**164**:107303. https://doi.org/10.1016/j.compbiomed.2023.107303.
17. Gong P, Cheng L, Zhang Z. *et al.* Multi-omics integration method based on attention deep learning network for biomedical data classification. *Comput Methods Prog Biomed* 2023;**231**:107377. https://doi.org/10.1016/j.cmpb.2023.107377.
18. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA. 2017.
19. Nguyen DA, Nguyen CH, Mamitsuka H. Central-smoothing hypergraph neural networks for predicting drug-drug interactions. *IEEE Trans Neural Netw Learn Syst* 2023;**35**:11620–5. https://doi.org/10.1109/TNNLS.2023.3261860.
20. Gao Y, Feng Y, Ji S. *et al.* HGNN(+): general hypergraph neural networks. *IEEE Trans Pattern Anal Mach Intell* 2023;**45**:3181–99. https://doi.org/10.1109/TPAMI.2022.3182052.
21. Li XS, Liu X, Lu L. *et al.* Multiphysical graph neural network (MP-GNN) for COVID-19 drug design. *Brief Bioinform* 2022;**23**:1–10. https://doi.org/10.1093/bib/bbac231.
22. Hodes RJ, Buckholtz N. Accelerating medicines partnership: Alzheimer's disease (AMP-AD) knowledge portal aids Alzheimer's drug discovery through open data sharing. *Expert Opin Ther Targets* 2016;**20**:389–91. https://doi.org/10.1517/14728222.2016.1135132.

23. Parker JS, Mullins M, Cheang MC. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 2009;**27**:1160–7. https://doi.org/10.1200/JCO.2008.18.1370.
24. Colaprico A, Silva TC, Olsen C. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016;**44**:e71. https://doi.org/10.1093/nar/gkv1507.
25. Leng D, Zheng L, Wen Y. *et al.* A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol* 2022;**23**:171. https://doi.org/10.1186/s13059-022-02739-2.
26. Behnan J, Finocchiaro G, Hanna G. The landscape of the mesenchymal signature in brain tumours. *Brain* 2019;**142**:847–66. https://doi.org/10.1093/brain/awz044.
27. Jovčevska I. Sequencing the next generation of glioblastomas. *Crit Rev Clin Lab Sci* 2018;**55**:264–82. https://doi.org/10.1080/10408363.2018.1462759.
28. Wang F, Pena-Pena K, Qian W. *et al.* T-HyperGNNs: hypergraph neural networks via tensor representations. *IEEE Trans Neural Netw Learn Syst* 2024;**14**:1–13.
29. Zhu J, Tan Y, Lin R. *et al.* Efficient self-attention mechanism and structural distilling model for Alzheimer's disease diagnosis. *Comput Biol Med* 2022;**147**:105737. https://doi.org/10.1016/j.compbiomed.2022.105737.
30. Pan T, Li C, Bi Y. *et al.* PFresGO: an attention mechanism-based deep-learning approach for protein annotation by integrating gene ontology inter-relationships. *Bioinformatics* 2023;**39**:btad094. https://doi.org/10.1093/bioinformatics/btad094.
31. Wang Z, Gao Q, Yi X. *et al.* Surformer: an interpretable pattern-perceptive survival transformer for cancer survival prediction from histopathology whole slide images. *Comput Methods Prog Biomed* 2023;**241**:107733. https://doi.org/10.1016/j.cmpb.2023.107733.
32. Wang Z, Bi Y, Pan T. *et al.* Targeting tumor heterogeneity: multiplex-detection-based multiple instance learning for whole slide image classification. *Bioinformatics* 2023;**39**:btad114. https://doi.org/10.1093/bioinformatics/btad114.
33. Zhang S, Li X, Zong M. *et al.* Efficient kNN classification with different numbers of nearest Neighbors. *IEEE Trans Neural Netw Learn Syst* 2018;**29**:1774–85. https://doi.org/10.1109/TNNLS.2017.2673241.
34. Ehsani R, Drabløs F. Robust distance measures for kNN classification of cancer data. *Cancer Informat* 2020; **19**:1176935120965542. https://doi.org/10.1177/1176935120965542.
35. Kaur A, Verma K, Bhondekar AP. *et al.* Implementation of bagged SVM ensemble model for classification of epileptic states using EEG. *Curr Pharm Biotechnol* 2019;**20**:755–65. https://doi.org/10.2174/1389201020666190618112715.
36. Mustaqeem A, Anwar SM, Majid M. Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants. *Comput Math Methods Med* 2018;**2018**:1–10. https://doi.org/10.1155/2018/7310496.
37. Paul A, Mukherjee DP, Das P. *et al.* Improved random Forest for classification. *IEEE Trans Image Process* 2018;**27**:4012–24. https://doi.org/10.1109/TIP.2018.2834830.
38. Chowdhury AR, Chatterjee T, Banerjee S. A random Forest classifier-based approach in the detection of abnormalities in the retina. *Med Biol Eng Comput* 2019;**57**:193–203. https://doi.org/10.1007/s11517-018-1878-0.
39. Lu C, Xie M. LDAEXC: LncRNA-disease associations prediction with deep autoencoder and XGBoost classifier. *Interdiscip Sci* 2023;**15**:439–51. https://doi.org/10.1007/s12539-023-00573-z.
40. Li J, Shi Z, Liu F. *et al.* XGBoost classifier based on computed tomography Radiomics for prediction of tumor-infiltrating CD8(+) T-cells in patients with pancreatic ductal adenocarcinoma. *Front Oncol* 2021;**11**:671333. https://doi.org/10.3389/fonc.2021.671333.
41. Gálvez JA, Jalali A, Ahumada L. *et al.* Neural network classifier for automatic detection of invasive versus noninvasive airway management technique based on respiratory monitoring parameters in a Pediatric Anesthesia. *J Med Syst* 2017;**41**:153. https://doi.org/10.1007/s10916-017-0787-3.
42. Zhang Y, Lin H, Yang Z. *et al.* Neural network-based approaches for biomedical relation classification: a review. *J Biomed Inform* 2019;**99**:103294. https://doi.org/10.1016/j.jbi.2019.103294.
43. Stevenson-Hoare J, Heslegrave A, Leonenko G. *et al.* Plasma biomarkers and genetics in the diagnosis and prediction of Alzheimer's disease. *Brain* 2023;**146**:690–9. https://doi.org/10.1093/brain/awac128.
44. Amin M, Tang S, Shalamanova L. *et al.* Polyamine biomarkers as indicators of human disease. *Biomarkers* 2021;**26**:77–94. https://doi.org/10.1080/1354750X.2021.1875506.
45. Huang JW, Chen YH, Phoa FKH. *et al.* An efficient approach for identifying important biomarkers for biomedical diagnosis. *Biosystems* 2024;**237**:105163. https://doi.org/10.1016/j.biosystems.2024.105163.
46. Wang Y, Liu ZP. Identifying biomarkers for breast cancer by gene regulatory network rewiring. *BMC Bioinformatics* 2022;**22**:308. https://doi.org/10.1186/s12859-021-04225-1.
47. Ashtawy HM, Mahapatra NR. Boosted neural networks scoring functions for accurate ligand docking and ranking. *J Bioinforma Comput Biol* 2018;**16**:1850004. https://doi.org/10.1142/S021972001850004X.
48. Setiono R, Liu H. Neural-network feature selector. *IEEE Trans Neural Netw* 1997;**8**:654–62. https://doi.org/10.1109/72.572104.
49. Chen J, Bardes EE, Aronow BJ. *et al.* ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009;**37**:W305–11. https://doi.org/10.1093/nar/gkp427.
50. Garwain O, Yerramilli VS, Romero K. *et al.* The Gαq/ phospholipase Cβ signaling system represses tau aggregation. *Cell Signal* 2020;**71**:109620. https://doi.org/10.1016/j.cellsig.2020.109620.
51. Luo H, Han L, Xu J. Apelin/APJ system: a novel promising target for neurodegenerative diseases. *J Cell Physiol* 2020;**235**:638–57. https://doi.org/10.1002/jcp.29001.
52. Shao P. MiR-216a-5p ameliorates learning-memory deficits and neuroinflammatory response of Alzheimer's disease mice via regulation of HMGB1/NF-κB signaling. *Brain Res* 2021;**1766**:147511. https://doi.org/10.1016/j.brainres.2021.147511.
53. Khairallah MI, Kassem LA, Yassin NA. *et al.* Activation of migration of endogenous stem cells by erythropoietin as potential rescue for neurodegenerative diseases. *Brain Res Bull* 2016;**121**:148–57. https://doi.org/10.1016/j.brainresbull.2016.01.007.
54. Wang Q, Tian J, Chen H. *et al.* Amyloid beta-mediated KIF5A deficiency disrupts anterograde axonal mitochondrial movement. *Neurobiol Dis* 2019;**127**:410–8. https://doi.org/10.1016/j.nbd.2019.03.021.
55. Li H, Wang R. A focus on CXCR4 in Alzheimer's disease. *Brain Circ* 2017;**3**:199–203. https://doi.org/10.4103/bc.bc_13_17.
56. Wang QL, Fang CL, Huang XY. *et al.* Research progress of the CXCR4 mechanism in Alzheimer's disease. *Ibrain* 2022;**8**:3–14. https://doi.org/10.1002/ibra.12026.
57. Hohman TJ, Dumitrescu L, Cox NJ. *et al.* Genetic resilience to amyloid related cognitive decline. *Brain Imaging Behav* 2017;**11**:401–9. https://doi.org/10.1007/s11682-016-9615-5.
58. Yu QS, Feng WQ, Shi LL. *et al.* Integrated analysis of cortex single-cell transcriptome and serum proteome reveals the novel biomarkers in Alzheimer's disease. *Brain Sci* 2022;**12**:1022. https://doi.org/10.3390/brainsci12081022.

59. Nagaraj S, Laskowska-Kaszub K, Dębski KJ. *et al.* Profile of 6 microRNA in blood plasma distinguish early stage Alzheimer's disease patients from non-demented subjects. *Oncotarget* 2017;**8**:16122–43. https://doi.org/10.18632/oncotarget.15109.
60. Beyer S, Lakshmanan A, Liu YY. *et al.* KT5823 differentially modulates sodium iodide symporter expression, activity, and glycosylation between thyroid and breast cancer cells. *Endocrinology* 2011;**152**:782–92. https://doi.org/10.1210/en.2010-0782.
61. Wu Q, Tian P, He D. *et al.* SCUBE2 mediates bone metastasis of luminal breast cancer by modulating immune-suppressive osteoblastic niches. *Cell Res* 2023;**33**:464–78. https://doi.org/10.1038/s41422-023-00810-6.
62. Dartier J, Lemaitre E, Chourpa I. *et al.* ATP-dependent activity and mitochondrial localization of drug efflux pumps in doxorubicin-resistant breast cancer cells. *Biochim Biophys Acta Gen Subj* 2017;**1861**:1075–84. https://doi.org/10.1016/j.bbagen.2017.02.019.
63. Oliemuller E, Newman R, Tsang SM. *et al.* SOX11 promotes epithelial/mesenchymal hybrid state and alters tropism of invasive breast cancer cells. *elife* 2020;**9**:e58374. https://doi.org/10.7554/eLife.58374.
64. Oliemuller E, Kogata N, Bland P. *et al.* SOX11 promotes invasive growth and ductal carcinoma in situ progression. *J Pathol* 2017;**243**:193–207. https://doi.org/10.1002/path.4939.
65. Dunlap J, Le C, Shukla A. *et al.* Phosphatidylinositol-3-kinase and AKT1 mutations occur early in breast carcinoma. *Breast Cancer Res Treat* 2010;**120**:409–18. https://doi.org/10.1007/s10549-009-0406-1.
66. Ahmed ARH, Griffiths AB, Tilby MT. *et al.* TFF3 is a normal breast epithelial protein and is associated with differentiated phenotype in early breast cancer but predisposes to invasion and metastasis in advanced disease. *Am J Pathol* 2012;**180**:904–16. https://doi.org/10.1016/j.ajpath.2011.11.022.
67. Al-Salam S, Sudhadevi M, Awwad A. *et al.* Trefoil factors peptide-3 is associated with residual invasive breast carcinoma following neoadjuvant chemotherapy. *BMC Cancer* 2019;**19**:135. https://doi.org/10.1186/s12885-019-5316-y.
68. Dietrich D, Lesche R, Tetzner R. *et al.* Analysis of DNA methylation of multiple genes in microdissected cells from formalin-fixed and paraffin-embedded tissues. *J Histochem Cytochem* 2009;**57**:477–89. https://doi.org/10.1369/jhc.2009.953026.
69. Gupta G, Chellappan DK, de Jesus Andreoli Pinto T. *et al.* Tumor suppressor role of miR-503. *Panminerva Med* 2018;**60**:17–24. https://doi.org/10.23736/S0031-0808.17.03386-9.
70. Xiao Y, Humphries B, Yang C. *et al.* MiR-205 dysregulations in breast cancer: the complexity and opportunities. *Noncoding RNA* 2019;**5**:53. https://doi.org/10.3390/ncrna5040053.