Fundamental Research 4 (2024) 738–751

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

Integrating multi-omics data of childhood asthma using a deep association model

Kai Wei [a,b], Fang Qian [a], Yixue Li [a,c,d,e,*], Tao Zeng [c,d,*], Tao Huang [a,*]

[a] Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[b] Guoke Ningbo Life Science and Health Industry Research Institute, Ningbo 315000, China
[c] Guangzhou National Laboratory, Guangzhou 510000, China
[d] GMU-GIBH Joint School of Life Sciences, The Guangdong-Hong Kong-Macau Joint Laboratory for Cell Fate Regulation and Diseases, Guangzhou Laboratory, Guangzhou Medical University, Guangzhou 510000, China
[e] Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China

ARTICLE INFO

Article history:
Received 23 June 2023
Received in revised form 6 March 2024
Accepted 17 March 2024
Available online 2 April 2024

Keywords:
Deep subspace reconstruction
Deep non-negative matrix factorization
Deep canonical correlation analysis
Multi-omics
Interpretable machine learning
Childhood asthma

ABSTRACT

Childhood asthma is one of the most common respiratory diseases with rising mortality and morbidity. The multi-omics data is providing a new chance to explore collaborative biomarkers and corresponding diagnostic models of childhood asthma. To capture the nonlinear association of multi-omics data and improve interpretability of diagnostic model, we proposed a novel deep association model (DAM) and corresponding efficient analysis framework. First, the Deep Subspace Reconstruction was used to fuse the omics data and diagnostic information, thereby correcting the distribution of the original omics data and reducing the influence of unnecessary data noises. Second, the Joint Deep Semi-Negative Matrix Factorization was applied to identify different latent sample patterns and extract biomarkers from different omics data levels. Third, our newly proposed Deep Orthogonal Canonical Correlation Analysis can rank features in the collaborative module, which are able to construct the diagnostic model considering nonlinear correlation between different omics data levels. Using DAM, we deeply analyzed the transcriptome and methylation data of childhood asthma. The effectiveness of DAM is verified from the perspectives of algorithm performance and biological significance on the independent test dataset, by ablation experiment and comparison with many baseline methods from clinical and biological studies. The DAM-induced diagnostic model can achieve a prediction AUC of 0.912, which is higher than that of many other alternative methods. Meanwhile, relevant pathways and biomarkers of childhood asthma are also recognized to be collectively altered on the gene expression and methylation levels. As an interpretable machine learning approach, DAM simultaneously considers the non-linear associations among samples and those among biological features, which should help explore interpretative biomarker candidates and efficient diagnostic models from multi-omics data analysis for human complex diseases.

1. Introduction

Childhood asthma is a severe and heterogeneous inflammatory disease whose pathogenesis remains unclear, although some researches have shown that complex inflammatory pathways are associated with childhood asthma [1]. Recently, the increasing multi-omics data provide a new chance to investigate the pathogenesis of childhood asthma from a collaborative perspective, which should help explore interpretative biomarker candidates and efficient diagnostic models.

Previously, Soliai et al. assessed transcriptional and epigenetic responses to rhinovirus (an asthma-promoting pathogen) based on an upper airway epithelial cell culture model, and provided specific transcriptional and epigenetic response mechanisms for explaining variants found in asthmatic GWASs annotation [2]. Forno et al. identified several top-ranked IL5RA SNPs associated with transcriptional factors of asthma by logistic regression models on childhood asthma omics data [3]. However, these analyses assumed a linear relationship between the data samples and could not completely capture the rich prior information. Indeed, suitable prior information can improve analysis model's performance to a certain extent and guide biologically meaningful discovery. For example, Zhang et al. fused the relationship matrix of data samples into the association analysis by network regularization and verified such

* Corresponding authors.
E-mail addresses: yxli@sibs.ac.cn (Y. Li), zeng_tao@gzlab.ac.cn (T. Zeng), huangtao@sibs.ac.cn (T. Huang).

https://doi.org/10.1016/j.fmre.2024.03.022

strategy's effectiveness [4]. Besides, the weighted gene co-expression network analysis (WGCNA) aggregated gene and metabolite signatures into multiple co-expression modules and found asthma-associated genes in different gene modules [5], but this analysis obviously cannot detect nonlinear associations between different omics data (e.g. genes and metabolites).

For efficiently integrating disease-related multi-omics and clinical data by considering diverse correlation information among samples and that among molecules/features, various computational schemes have been developed [6–8]. Especially, the integration method based on the Non-negative Matrix Factorization (NMF) technique has attracted wide attention due to its low time complexity and strong interpretability advantages. Zhang et al. proposed the Joint Non-negative Matrix Factorization (JNMF) algorithm to integrate cancer-related multi-omics data [9]; however, the JNMF algorithm adopts a strategy of random initialization of parameters, resulting in non-uniqueness in solving the objective function and making it sensitive to data noise, which affects the stability of the results. Deng et al. further proposed a Joint Sparse Network-Regularization Multiple NMF (JSNMNMF) algorithm to construct a ceRNA network and other variant models [10,11], which incorporates various prior information [12]. Wang et al. further incorporated more network regularization constraints into the algorithm and applied it to ceRNA network construction, revealing the interaction of three cancer-related RNAs [13]. Wei et al. proposed a Joint Connectivity-Negative Matrix Factorization (JCB-SNMF) algorithm to search for Alzheimer's disease-related biomarkers, by introducing connectivity information from imaging and genetic data into JNMF model [14]. However, these models they proposed assume a linear relationship between the omics data and would underestimate the nonlinear relationship. By contrast, Diego Salazar et al. proposed a Kernel Joint Non-negative Matrix Factorization, which integrated the factorization of the original matrices into a high-dimensional space and obtained better clustering and interpretation results than JNMF [15]. Sehwan Moon et al. proposed a Joint Deep Semi-Nonnegative Matrix Factorization (JDSNMF) algorithm and applied it to various multi-omics integration tasks, achieving improved algorithmic performance and biological results [16]. However, these methods do not make full use of available diagnostic information.

In addition, the above work generally conducted an overall analysis and could not efficiently evaluate the element/feature importance so as to reduce the model interpretability. Canonical correlation analysis (CCA) would be an effective method for conducting multivariate correlation analysis on a group of features. Most CCA-based methods are unsupervised and require a separate downstream analysis of diagnostic groups [17,18]. CCA based on assumptions within linear association may not be able to explain the implicit nonlinear associations. Although a few CCA variants based on kernel methods, e.g. KCCA [19] and grad KCCA [20] have been proposed to use kernel function for non-linearly data transformation, the limited kernel choices and parameter optimization are still not satisfactory enough.

Inspired by these computational questions and analysis issues, we aim to simultaneously identify nonlinear associations between multi-omics data and combine diagnostic information within such integrative analysis, which can help reduce the influence of unnecessary data signals/noises and provide an interpretable diagnostic model. In this work, we designed and implemented an interpretable machine learning approach, i.e. deep association model (DAM), and applied it to integrate transcriptome and DNA methylation data of childhood asthma. DAM includes three main steps: (i) the deep subspace reconstruction (DSR), which considers the clinical diagnosis information of patients as a priori information to correct the data distribution, enhancing clinical signal and reducing noisy signal; (ii) the deep joint NMF inference (JDSNMF), which performs deep correlation analysis on different omics datasets, identifying their potential nonlinear relationship within collaborative modules and discriminating candidate biomarkers; (iii) a new deep orthogonal canonical correlation analysis (DOCCA), which performs multivariate correlation analysis between reconstructed elements (e.g. gene and methylation loci) in the collaborative module, providing the top-ranked features to construct a robust diagnostic model based on multi-omics signatures.

Using DAM, we deeply analyzed the transcriptome and methylation data of childhood asthma as a detailed case study. First, the effectiveness of DAM is verified from the perspectives of algorithm performance and biological significance on both the train and test datasets, by wide ablation experiments and comparison with many baseline methods. Then, after efficient learning of DER and JDSNMF, the prediction AUC of DAM-induced diagnostic model with logistic regression can achieve 0.912, which is larger than those of many other alternative methods. Next, relevant pathways and biomarkers of childhood asthma are recognized by DOCCA to be collectively altered on the gene expression and methylation levels, which are also assessed on test data and independent data. Collectively, DAM should be effective for multi-omics data analysis, simultaneously providing discriminative molecular features and nonlinear biological correlations/explanations for deeply understanding human complex diseases.

2. Methods

2.1. Deep association model

2.1.1. Deep subspace reconstruction with prior information integration for cleaned data

Previous studies have confirmed that prior information can improve the association analysis performance of multi-omics integration models and algorithms. However, most multi-omics integration schemes used various penalty terms directly on original data with technical and biological noise, resulting in the estimation bias during integration [21]. Therefore, considering the multi-subspace structure at the bottom of the data, DAM is first to integrate the subjects/samples' diagnostic information into the original data by deep subspace reconstruction (DSR). Specifically, DSR applies the self-expressive nature of the data; and it guarantees that the samples with the same diagnosis labels will gather in the same subspace and the samples with different diagnosis labels should distribute in different subspaces.

Obviously, applying self-expressivity directly to original data would ignore nonlinear relationships among data points/samples. By contrast, DSR first makes the original data go through a multi-layer feed-forward neural network to perform a nonlinear transformation on the original data and then reconstructs the embedding data in the subspace at the output layer of the network. On the training/discovery data, DSR combines the transcriptome matrix $X_1 \in R^{N \times p}$, and DNA methylation matrix $X_2 \in R^{N \times q}$, and the diagnostic labels of samples as input of multi-layer feed-forward neural network, where N, p, and q are the number of samples, transcriptome genes, and DNA methylation sites/loci, respectively. The clean data after non-linearly transformation is output at the output layer, along which the subspace is iteratively calculated.

Briefly, the subspace reconstruction of $X_1 = [x_1, x_2, \dots x_i, \dots, x_N]$ is used as an example to illustrate the detailed calculation. First group X_1 by label, i.e. $X_1 = [x_1^{(1)}, x_2^{(1)}, \cdots, x_{N_1}^{(1)}, \dots, x_1^{(2)}, x_2^{(2)}, \cdots, x_{N_2}^{(2)}]$, where $x_i^{(1)} (i = 1, 2, .., N)$ represents the first label of the first class $i-th$ samples. Let $x_i^{(1)} = h_i^{(0)(1)} \in R^p$, set $\theta = \{W^{(m)}, b^{(m)}, C, m = 1 : M, i = 1 : N\}$ is the hyperparameter in the feedforward neural network, where m represents the current number of layers in the network, and the output of the i_{th} sample of the first-class label of the m_{th} layer is defined as follows:

$$h_i^{(m)(1)} = G\left(W^{(m)} h_i^{(m-1)(1)} + b^{(m)}\right) \quad (1)$$

Given d_m represents the dimension of the m_{th} layer of the neural network, $W^{(m)} \in R^{d_m \times d_{m-1}}$ and $b^{(m)} \in R^{d_m}$ represent the weight matrix and bias matrix of this layer, respectively. The output of the last layer is $H^{(M)(1)} = [h_1^{(M)(1)}, h_2^{(M)(1)}, \cdots, h_{N_1}^{(M)(1)}, \dots, h_1^{(M)(2)}, h_2^{(M)(2)}, \cdots, h_{N_2}^{(M)(2)}]$,

and then apply the nonlinearly transformed output layer data to perform the subspace reconstruction. Among them, $h_i^{(M)(1)}$ represents the i_{th} sample in the first class after the nonlinear transformation of the multi-layer neural network. The following is the objective function of reconstructing the i_{th}sample of the d_{th} class label with the samples of the same class:

$$\min_{\{W(m),b(m)\}_{m=1}^{M},C_i^{(d)}} \frac{1}{2}\sum_{i=1}^{N_d} \| h_i^{(M)(d)} - C_i^{(d)} H^{(M)(d)} \|_F^2 + \lambda \| C_i^{(d)} \|_1 \quad (2)$$

Among them, $C_i^{(d)}$ is the vector composed of the self-expression coefficients of the i_{th} nonlinearly transformed sample of the d_{th} class label. Then, the process of updating $W^{(m)}$, $b^{(m)}$ and $C_i^{(d)}$ is carried out as section 1.1 in the supplementary material. Finally, the self-expression coefficient matrix C of all samples can be expressed as

$$C = \left[C_1^{(1)},\ldots,C_{N_2}^{(2)}\right] + \left[C_1^{(1)},\ldots,C_{N_2}^{(2)}\right]^T \quad (3)$$

C has a block structure, which can reflect the similarity structure of the cleaned data by reconstructing the original data with schematic diagram shown in Fig. S1 in supplementary material.

Finally, the cleaned multi-omics data from DSR is subsequently a new data matrix as follows:

$$f(X_i) = C_i X_i \ (i = 1, 2) \quad (4)$$

C_1 and C_2 represent the self-expression coefficient matrix of two training datasets X_1 and X_2, respectively. Thus, $f(X_1)$ and $f(X_2)$ are the reconstruction matrices from DSR. In the execution process of the DSR algorithm, the integration of diagnostic label information is mainly manifested in the following aspects.

When learning the self-expression matrix, the learning process of the self-expression matrix takes into account label information. Specifically, through Eq. 2, for each sample, the model uses data and label information from other samples to learn the self-expression matrix. Thus, the self-expression matrix is not only used for data reconstruction but also takes into account the label relationships between samples. In the process of data reconstruction, the learned self-expression matrix is used to reconstruct the original data. In addition to label information, our model also takes into account the similarity and correlation between samples. The weight matrix W in the self-expression matrix is used to capture relationships between samples, including label information. The DSR is capable of reducing noise in raw data, improving the performance of subsequent algorithms while discovering more robust diagnostic biomarkers. A detailed explanation of this aspect can be further found in Section 1.2 of the supplementary material.

The difference between DSR and traditional classification or regression models is that DSR can learn hidden representation of samples before predicting labels. In fact, a representation is learned that preserves the features of the original data as much as possible, taking into account the label information. This learned representation can be used for subsequent tasks such as classification or regression. Unlike traditional models that are trained directly based on original data and labels, DSR emphasizes the incorporating self-expression relationship of samples during the representation learning process. This makes DSR more flexible and can better capture the inherent structure of the data.

2.1.2. Joint deep semi-nonnegative matrix factorization with nonlinear feature correlation extraction for collaborative module

Based on the above adjusted multi-omics data, Joint Deep Semi-Nonnegative Matrix Factorization (JDSNMF) is then carried out according to the principle of multilayer NMF and nonlinear manifolds [16].

JDSNMF decomposes different cleaned omics data into a common sample latent matrix and multiple feature latent matrices, where the nonlinear feature association analysis is achieved by layer-by-layer dimensionality reduction of the feature latent matrix and nonlinear transformation during dimensionality reduction by a layer-by-layer activation function. The objective function of JDSNMF is shown as follows:

$$\begin{aligned}
&\min \sum_{i=1}^{2} \left\| f(X_i) - U H_{i0} \right\|_F^2 + \lambda \|S\|_F \\
&s.t. H_{io} = s\left(Z_{i_n} s\left(Z_{i_2} \ldots H_{i_{N-1}}\right)\right) \\
&H_{i_{n-1}} = s\left(Z_{i_n} H_{i_n}\right), H_{i_{N-1}} \geq 0, \\
&S \in \left\{ U, Z_{i=1} \left\| f(X_i) - U H_{i0} \right\|_F^2 + \lambda \|S\|_F \right\}
\end{aligned} \quad (5)$$

Among them, $U \in \mathbb{R}^{n \times k_0}$ is the sample latent matrix; $H_{i_0} \in \mathbb{R}^{k_0 \times p_i}$ is the feature latent matrix produced at the first layer; $H_{i_n} \in \mathbb{R}^{k_n \times p_i}$ is the feature latent matrix produced at the $(n+1)_{th}$ sub-layer; $Z_n \in \mathbb{R}^{k_{n-1} \times k_n}$ is the junction latent matrix; N is the number of layers of the network. S is the set of decomposition terms. $\| \cdot \|_F$ is the Frobenius norm. In the JDSNMF, $k_0 < k_i < k_{i+1} < k_n < \min\{n, p_i\}$ need to be satisfied. And the nonlinear decomposition of H_{i_0} adopts the sigmoid activation function as follows:

$$s(x) = \frac{1}{1+e^{-x}} \quad (6)$$

Through JDSNMF, the final U, H_{10}, and H_{20} can be obtained. The cleaned data matrices (e.g., gene expression matrix $f(X_1)$ and DNA methylation matrix $f(X_2)$) might share a sample latent matrix U, which can be regarded as the common feature basis matrix for samples and each feature basis indicates a collaborative module across a group of samples. Meanwhile, H_{10} and H_{20} would represent the decomposed feature coefficient matrices respectively, indicating the potential relation between feature basis (e.g. hidden sample representation) and original features (e.g. genes or methylation loci).

To further determine the salient (original) features corresponding to each feature base of U, a z-score is applied to extract the coefficients of each feature coefficient vector of each feature coefficient matrix. It is defined as $Z_{ij} = \frac{h_{ij}-\mu_j}{\sigma_j}$, where h_{ij} refers to the feature coefficient element, μ_j is the mean of feature coefficients of feature base j, and σ_j refers to the standard deviation of these feature coefficients. For each original feature, if its z-score is greater than a threshold T (e.g., 2), it is considered to be a salient feature for one feature base. The whole salient features for each feature base consist of a collaborative module.

2.1.3. Deep orthogonal constrained canonical correlation analysis with salient feature ranking for diagnostic model

For each collaborative module, a Deep Orthogonal Constrained Canonical Correlation Analysis (DOCCA) algorithm was developed to reduce the influence of the colinearity of salient features on the feature ranking and selection for building diagnostic model, by the orthogonal constraint on weight vectors of CCA. The input data of DOCCA should be the cleaned data corresponding to one collaborative module, and the objective function of DOCCA is as follows:

$$\begin{aligned}
&\min_{u,v} - u^T f\left(X_1^{[K]}\right)^T \left(X_2^{[k]}\right) v + \lambda_1 \|uu^T - I\|_2^2 + \lambda_2 \|VV^T - I\|_2^2 \\
&\quad s.t. \|f\left(X_1^{[K]}\right) u\|_2^2 = 1, \|f\left(X_2^{[k]}\right) v\|_2^2 = 1
\end{aligned} \quad (7)$$

Among them, $f(X_1^{[k]}) \in \mathbb{R}^{n \times p^{[k]}}$ is the cleaned gene expression submatrix for collaborative module k, $p^{[k]}$ represents the gene in module k quantity. $f(X_2^{[k]}) \in \mathbb{R}^{n \times q_k}$ is the cleaned methylation expression submatrix for the same module, and q_k represents the number of methylation loci in module k. $u \in \mathbb{R}^{p_k \times 1}$ and $v \in \mathbb{R}^{q_k \times 1}$ represent CCA weight vectors for genes and methylation, respectively. I is the identity matrix, and λ_1 and λ_2 are hyperparameters used to control the strength of u and v sparsity constraints, respectively. Especially, above objective function can be re-presented as the following formula:

$$\begin{aligned}
&\min_{u,v} \ \|f\left(X_1^{[K]}\right) u - f\left(X_2^{[k]}\right) v\|_2^2 + \lambda_1 \|uu^T - I\|_2^2 + \lambda_2 \|vv^T - I\|_2^2 \\
&\quad s.t. \|f\left(X_1^{[K]}\right) u\|_2^2 = 1, \|f\left(X_2^{[k]}\right) v\|_2^2
\end{aligned} \quad (8)$$

Obviously, this objective function can be optimized by alternately iteratively updating u and v using the Lagrange operator:

$$L(u,v) = \left\| f\left(X_1^k\right)u - f\left(X_2^{[k]}\right)v \right\|_2^2 + \lambda_1 \left\| uu^T - I \right\|_2^2 + \lambda_2 \left\| vv^T - I \right\|_2^2 + \gamma_1 \left(\left\| f\left(X_1^{[k]}\right)u \right\|_2^2 - 1 \right) + \gamma_2 \left(\left\| f\left(X_2^{[k]}\right)v \right\|_2^2 - 1 \right) \quad (9)$$

First considering v as a constant term, fix v to solve u, and L(u, v) takes the derivative of u and sets it as 0.

$$-\mathrm{f}\left(\mathrm{X}_1^{[\mathrm{k}]}\right)^{\mathrm{T}} \mathrm{f}\left(\mathrm{X}_2^{[\mathrm{k}]}\right)\mathrm{v} + 2\lambda_1(\mathrm{uu}^{\mathrm{T}} - \mathrm{I})\mathrm{u} + (1 + \gamma_{\mathrm{u}})\mathrm{f}\left(\mathrm{X}_1^{[\mathrm{k}]}\right)^{\mathrm{T}} \mathrm{f}\left(\mathrm{X}_1^{[\mathrm{k}]}\right)\mathrm{u} = 0 \quad (10)$$

Then the iterative formula of u can be obtained:

$$\mathrm{u} = \left(2\lambda_1(\mathrm{uu}^{\mathrm{T}} - \mathrm{I}) + (1 + \gamma_1)\mathrm{f}\left(\mathrm{X}_1^{[\mathrm{k}]}\right)^{\mathrm{T}} \mathrm{f}\left(\mathrm{X}_1^{[\mathrm{k}]}\right)\right)^{-1} \mathrm{f}\left(\mathrm{X}_1^{[\mathrm{k}]}\right)^{\mathrm{T}} \mathrm{f}\left(\mathrm{X}_2^{[\mathrm{k}]}\right)\mathrm{v} \quad (11)$$

In the same way, the iterative formula of v can be obtained:

$$\mathrm{v} = \left(2\lambda_2(\mathrm{vv}^{\mathrm{T}} - \mathrm{I}) + (1 + \gamma_2)\mathrm{f}\left(\mathrm{X}_2^{[\mathrm{k}]}\right)^{\mathrm{T}} \mathrm{f}\left(\mathrm{X}_2^{[\mathrm{k}]}\right)\right)^{-1} \mathrm{f}\left(\mathrm{X}_2^{[\mathrm{k}]}\right)^{\mathrm{T}} \mathrm{f}\left(\mathrm{X}_1^{[\mathrm{k}]}\right)\mathrm{u} \quad (12)$$

2.2. Evaluation and case study

2.2.1. Dataset

In this work, the childhood asthma datasets were used for model and algorithm evaluation and deep case study, which were downloaded from the GEO database [22]. The gene expression profiling dataset GSE40732 and DNA methylation expression profiling dataset GSE40576 [23] of the same samples from children with asthma were collected. The samples of the two datasets were DNA/RNA of peripheral blood mononuclear cells (PBMCs) from inner-city 6–12-year-old children, which were used to compare gene expression and methylation patterns in children with persistent atop-rankedic asthma and healthy controls. There was total 194 samples, including 97 normal samples and 97 correspondingly matched diseased samples. The data platform of GSE40732 is GPL16025 (NimbleGen Homo sapiens Expression Array). The data platform of GSE40576 is GPL13534 (Illumina HumanMethylation450 BeadChip). Dataset probe name annotations all use the chip GPL platform file. We have also collected two gene expression profile datasets (GSE27011 and GSE40888) as external test sets. The GSE27011 dataset comprises samples of DNA/RNA from white blood cells of children in the asthma and healthy control groups, including 18 healthy control samples and 36 asthma samples. The GSE40888 dataset includes samples of DNA/RNA from children's PBMCs recruited by the Munich Clinical Asthma Research Association during visits to asthma clinics from January 2009 to July 2014, consisting of 40 healthy control samples and 65 asthma samples. The platforms for both datasets are GPL6244 (Affymetrix Human Gene 1.0 ST Array). The GSE109446 dataset comprises samples from the nasal epithelial cells of children at the Cincinnati Children's Hospital Medical Center, including 29 healthy control samples and 29 asthma samples. The platform for this dataset is GPL13534 (Illumina HumanMethylation450 BeadChip). All data samples in datasets were included in subsequent analysis after pre-procession.

2.2.2. Setting of model and algorithm evaluation

The original datasets were divided as training/discovery and test/validation datasets independently. The training and test samples in a ratio of about 8:2, and the proportion of asthma and control samples were kept the same in the training and test datasets.

(i) The Limma package was used to perform differential expression analysis of gene expression and DNA methylation profiles in the training dataset, respectively.

(ii) The DSR was applied to obtain cleaned training data matrices by incorporating the sample diagnostic information.

(iii) The JDSNMF was carried out to extract collaborative modules and corresponding salient gene or methylation features, and two performance measurements were adopted in the parameter selection. One is the Pearson correlation coefficient (PCC) between the original and reconstructed matrices of the two datasets as a recovery indicator:

$$PCC\left(X_i, UH_{i_0}\right) = \frac{E\left(\left(X_i - \mu_{X_i}\right)\left(UH_{i_0} - \mu_{UH_{i_0}}\right)\right)}{\sqrt{E\left(\left(X_i - \mu_{X_i}\right)^2\right)}\sqrt{E\left(\left(UH_{i_0} - \mu_{UH_{i_0}}\right)^2\right)}} \quad (13)$$

Among them, E represents the expectation, and μ_{X_i} represents the mean values. Two is the reconstruction error for measuring the algorithm's performance, as follows:

$$relative-error = \frac{(1/N * p)\sum_{ij}\left|(X_1)_{ij}\left(UH_{10}\right)_{ij}\right|}{1/N * p\sum_{ij}\left(UH_{10}\right)_{ij}} + \frac{1/N * q\sum_{ij}\left|(X_2)_{ij} - \left(UH_{20}\right)_{ij}\right|}{1/N * q\sum_{ij}(X_2)_{ij}}a \quad (14)$$

(iv) The escape rate is also used as an evaluation indicator for collaborative module selection. Suppose there are m candidate modules in total. When calculating the escape rate for the n_{th} $(0 < n \leq m)$ module, the escape rate is defined as the ratio of elements in the n_{th} module that do not overlap with other $m - 1$ modules to the total number of elements in the n_{th} module. The smaller the escape rate, the more the intersection between the representative modules, and the more representative the common elements contained.

(v) Ten-fold cross-validation was performed on the training dataset with feature ranking by a grid search strategy to select the best parameters for the classifier. Such validation was repeated ten times with different random seeds fixed, and the area under the receiver operating characteristic curve (AUC) was calculated for each classification model. The average AUC of ten times was taken as the final classification performance, and the standard deviation of AUC of ten times was also calculated for assessing the stability of classification modeling.

Then, four main compared algorithms as baselines were briefly introduced, with their objective functions and parameter settings.

(i) The JNMF algorithm is a common multi-omics data integration algorithm, and its objective function is as follows:

$$\sum_{i=1}^{2} ||f(X_i) - WH_i||_F^2 S.t. W \leq 0, H_i \leq 0 \quad (15)$$

(ii) The JCB-SNMF algorithm made some improvements based on the JNMF algorithm, by adding the Frobenius norm constraints on W and H_i and the Laplacian constraint terms on the coefficient matrix, whose objective function is as follows:

$$\sum_{i=1}^{2} ||f(X_i) - WH_i||_F^2 - \lambda_1 Tr\left(H_1 A_1 H_1^T\right) - \lambda_2 Tr\left(H_2 A_2 H_2^T\right) + \left(H_1 B H_2^T\right) + \gamma_1 ||W||_F^2 + \gamma_2\left(\sum_{i=1}^{2} H_{iF}^2\right) s.t. \cdot W \leq 0, H_i \leq 0 \quad (16)$$

Among them, A_1 or A_2 are matrices composed of the absolute values of the PCC of X_1 or X_2, B is a matrix composed of the absolute values of the PCC between X_1 and X_2, and γ_2 is used to control the degree of constraint of the orthogonal constraint.

$$B = \begin{bmatrix} |r_{11}| & \cdots & |r_{1p}| \\ \vdots & \ddots & \vdots \\ |r_{q1}| & \cdots & |r_{pq}| \end{bmatrix} \quad (17)$$

Among them, $|\cdot|$ represents the absolute value of $\cdot$. r_{ij} represents the vector corresponding to the i_{th} feature of X_{1i}. X_{1i} can be expressed as $(x_{1i}, x_{1_{2,i}}, \cdots, x_{1_{ni}})$ and the vector corresponding to the j_{th} feature of X_{2i}. X_{2i} can be expressed as $(x_{1i}, x_{1_{2,i}}, \cdots, x_{2_{ni}})$, then r_{ij} can be expressed as $PCC(X_{1,i}, X_{2,j})$. In addition, A_1 and A_2 represent the connectivity matrices of X_1 and X_2, respectively, which can be used to explore the structural correlation between features within the same data. Taking A_1 as an example, $A_1 = D_1 - C_1$. C_1 can be obtained by calculating the Pearson correlation between any two features of X_1. D_1 is a diagonal matrix, and its i_{th} diagonal element represents the sum of the i_{th} row elements of the connectivity matrix C_1.

(iii) In addition to using the Pearson correlation coefficient matrix as prior information, the MDJNMF algorithm also adds orthogonal constraints to the coefficient matrix to prevent redundant features from their negative influence on the analysis results. The objective function of the MDJNMF algorithm is shown below:

$$\sum_{i=1}^{2}\left\|f(X_i)-WH_1\right\|_F^2-\lambda_1\left\|H_1H_1^T-I\right\|_F^2-\lambda_2\left\|H_2H_2^T-I\right\|_F^2 +\beta\mathrm{Tr}\left(H_1BH_2^T\right)+\gamma_1\|W\|_F^2+\gamma_2\left(\sum_{i=1}^{2}\|H_i\|_F^2\right)s.t.W\geq 0, H_i\geq 0\, s.t.\, W\geq 0, H_i\geq 0 \tag{18}$$

where α is used to control the strength of the orthogonal constraint.

(iv) Based on the MDJNMF algorithm, the NSOJNMF algorithm simultaneously embeds the absolute values of PCC between different modal data and the same modality. The NSOJNMF algorithm was used to construct the ceRNA network of liver cancer patients, and its constraint term needed to use the interaction information between different RNAs. Therefore, the Laplace matrix and Pearson correlation coefficient matrix mentioned in the MDJNMF and JCB-SNMF algorithms could be used instead. The objective function of the NSOJNMF algorithm is as follows:

$$\min\sum_{i=1}^{2}\left\|f(X_i)-WH_i\right\|_F^2+\alpha\sum_{i=1}^{2}\left\|H_iH_i^T-I\right\|_F^2-\lambda_1\sum_{i=1}^{2}Tr\left(H_iA_iH_i^T\right) -\lambda_2Tr\left(H_1BH_2^T\right)+\gamma\sum_{i=1}^{2}\|H_i\|_1 s.t.w\geq 0 \tag{19}$$

Among them, λ_1 and λ_2 respectively control the degree of the constraint of the absolute value of PCC in the same mode and different modes. A_i is the PCC matrix of co-modal data.

For selecting the number K of co-expression modules, all algorithms were adopted the consistent method with the JDSNMF algorithm under given parameter setting: for the JCB-SNMF algorithm, $\lambda_1 = \lambda_2 = \gamma_1 = \gamma_2 = \beta = 0.01$; for the MDJNMF algorithm, $\lambda_1 = \lambda_2 = \gamma_1 = \beta = 0.01, \gamma_2 = 10$; for the NSOJNMF algorithm, $\alpha = 10, \lambda_1 = \lambda_2 = \gamma = 0.01$.

Finally, the DOCCA was applied to assign feature weights for ranking the salient features in given collaborative module, which was also compared with typical CCA, KCCA, and gradKCCA, whose input was the same DSR-reconstructed data for given collaborative module and output the corresponding canonical correlation coefficients.

(i) Again, the objective function of typical CCA algorithm is as follows:

$$\min_{u,v}-u^Tf\left(X_1^k\right)^Tf\left(X_2^{[k]}\right)vs.t.\left\|f\left(X_1^{31}\right)u\right\|_2^2=1,\left\|f\left(X_2^{31}\right)v\right\|_2^2=1 \tag{20}$$

And let $f(X_1^{[k]})$ be mapped into a Hilbert space F through a nonlinear mapping

$$\Phi:\mathbf{R}^{n_x}\to\mathbf{F},\ f\left(X_1^{[k]}\right)\to\Phi\left(f\left(X_1^{[k]}\right)\right) \tag{21}$$

(ii) The objective function of the KCCA algorithm is as follows:

$$\min_{u,v}-u^T\Phi\left(f\left(X_1^{[k]}\right)\right)\Phi\left(f\left(X_2^{[k]}\right)\right)^Tv \quad s,t.\left\|\Phi\left(f\left(X_1^{[k]}\right)\right)u\right\|_2^2=1,\left\|\Phi\left(f\left(X_2^{[k]}\right)\right)v\right\|_2^2=1 \tag{22}$$

(iii) And the gradKCCA model is a more direct way to find the underlying relationship, which will optimize the KCCA with respect to the coefficients of the data space. In this setting, the data space coefficient vectors u and v can be regarded as pre-images. There are constraints on the pre-images of $\phi_x(u)$ and $\phi_x(v)$ to solve the KCCA problem in the feature space, and the objective function is as follows:

$$\rho_{\mathrm{gradKCCA}}=\frac{\frac{1}{n}\sum_{i=1}^{n}\phi_x\left(f(X_1^{[k](i)}\right),\phi_x(u)\phi_y\left(f(X_2^{[k](i)}\right),\phi_y(v)}{\frac{1}{n}\sum_{i=1}^{n}\phi_x\left(f(X_1^{[k](i)}\right),\phi_x(u)\phi_x(f\left(X_1^{[k](i)}\right),\phi_x(u)\frac{1}{n}\sum_{k=1}^{n}\phi_y\left(f(X_2^{[k](i)}\right),\phi_y(v)\phi_y(f\left(X_2^{[k](i)}\right),\phi_y(v)} \tag{23}$$

where a, b represents the inner product between a and b. Further, replace the inner product with a kernel function$k^x(f(X_1^{[k]},\mathbf{z})=\phi_x(f(X_1^{[k]}),\phi_x(\mathbf{z})$, $k^y(f(X_2^{[k]},\mathbf{z})=\phi_x(f(X_2^{[k]}),\phi_y(\mathbf{z})$ and denoting the resulting score vectors as $\mathbf{k}^x(\mathbf{u})=(k^x(f(X_1^{[k](i)},\mathbf{u}))_{i=1}^n$ and $\mathbf{k}^y(\mathbf{v})=(k^y(f(X_2^{[k](i)},\mathbf{v}))_{i=1}^n$.

$$\rho_{gradKCCA}=\max_{u,v}\frac{k^x(u)^TK^Y(v)}{\|k^x(u)\|_2\|k^y(v)\|_2} \tag{24}$$

In addition, the performance evaluation criteria of all the above algorithms are canonical correlation coefficients (CCCs), whose formula is as follows:

$$CCCs=PCC\left(f\left(X_1^{[k]}\right)u,f\left(X_1^{[k]}\right)v\right) \tag{25}$$

2.2.3. Setting of case study on childhood asthma

As mentioned above, the whole dataset contained of 97 asthma and 97 normal samples; 80% of the samples were random selected as the training dataset and remaining 20% sample were the test dataset; e.g., 154 training samples (including 77 diseased and 77 normal samples) were used for collaborative module and biomarker discovery and other 40 test samples (including 20 diseased and 20 normal samples) were prepared for case study.

Limma package was applied to identify differentially expressed genes and methylation sites between asthma and normal samples in the training set, ensuring the independence of the training set and the test set. There were 449 differentially expressed genes and 856 differentially methylated loci identified with $p < 0.05$.

The DSR was used to reconstruct the two kinds of data, and change the data sample distribution by introducing the diagnostic information of the samples, and use the t-distributed Stochastic Neighbor Embedding (tSNE) algorithm to reduce the data dimension and visualize the sample distribution in reconstructed/cleaned data [24]. Then, the JDSNMF decomposed the reconstructed data normalized by DSR and performed various downstream analyses on the extracted collaborative modules.

The collaborative modules were screened using typical methods [10,11]. And the escape rate and reconstruction error values of different modules were also calculated to evaluate corresponding modules' biological salient and reconstruction quality. The salient features in selected/targeted module with the slightest reconstruction error were selected for further analysis, including protein-protein interaction network (PPI) analysis, functional enrichment analysis, and diagnostic model construction. In PPI network analysis, the gene pairs were filtered according to the significance of the gene expression correlation and only the significant gene pairs ($p < 0.05$) were retained for diseased and normal samples respectively, e.g., differential network analysis [25]. The functional enrichment analysis was carried out as following steps: the modules containing no salient elements were filtered; the remaining modules were considered as effective modules when they were enriched in at least one GO term [26] or one KEGG pathway [27] ($p < 0.05$).

3. Results

3.1. Discriminative effect of DSR for adjusting data matrices

The parameter settings of the DSR were detailed in section 1.3 of the supplementary material. To confirm that DSR can effectively incorporate diagnostic information into the original data matrix, the t-sne dimensionality reduction of original and adjusted data matrices were visualized in Fig. 1a, where C_1 and C_2 respectively represent the self-expression coefficient matrix of genes after DSR. Obviously, the groups of samples had disorganized distribution in original data space; meanwhile the same groups of samples displayed a distinguishing distribution in the new adjusted data space after incorporating the diagnostic labels through DSR.

3.2. Deep matrix decomposition of JDSNMF for extracting collaborative modules

The parameters used in JDSNMF were detailed in section 1.4 of the supplementary material. The loss of JDSNMF and the PCC between the input and output data matrices were adopted for parameter selection.

In Fig. S2a-d, these two performance indicators of JDSNMF under different parameter combinations were summarized and compared. Indeed, different hyperparameter combinations would have remarkable impact on the JDSNMF, e.g., the larger the neural network dimension, the better the reconstruction performance, which indicated that the nonlinearity of the data would be fully captured and help improve analysis performance. Finally, a set of parameter combinations maximizing the average recovery PCC were determined, and the reconstruction process of two data matrices by the JDSNMF algorithm was visualized in Fig. S3a-l.

3.3. Ablation analysis of deep association model

To verify the effectiveness of the DAM, it is necessary to evaluate the effect of DSR on the loss of following JDSNMF. As an ablation analysis for this evaluation, one is to directly input the original data after Min-Max normalization into the JDSNMF; the other is to input the cleaned data from DSR with Min-Max normalization into the same JDSNMF. As shown in Fig. 2i, DSR indeed enables JDSNMF to have a faster convergence rate and effective matrix decomposition.

3.4. Ranking and selection of collaborative modules

By DAM analysis, 151 collaborative modules were obtained. To verify the biological significance of all modules, this paper takes the union of genes and methylation site genes in all modules. Further, this paper uses the Dose package [28] to perform disease ontology enrichment analysis on these genes ($p < 0.05$) (Fig. S4). The circle size in the figure represents the number of genes enriched in the disease, and the line color corresponds to different diseases. These genes can be enriched for allergic asthma.

Then, 78 modules were removed because they did not contain any salient features; 9 modules were retained, which included exceeded 2% of the total number of genes and methylation loci. Since too few elements in the module will be unfavorable for subsequent analysis, we calculate the mean of the number of elements contained in these modules separately and retain four modules with the number of elements greater than the mean.

Actually, DAM has shown its effective reconstruction of the original data on the module level again. The correlation scatter plot between the two kinds of data matrices and their reconstructed matrices (Fig. 2a-h) indicated that these co-expression modules have good reconstruction performance, and Venn diagrams revealed many overlapping elements in these modules too (Fig. S5), which both confirmed the effectiveness of DAM. The element intersection of four essential collaborative modules were shown in Fig. S6a-b, and the escape rates of module 12, module 25, module 31, and module 75 were 23.8%, 86.67%, 16.67%, and 100% on gene expression level; meanwhile, they were 79.31%, 14.29%, 81.48% and 4.17% on methylation level, respectively. These findings disclosed the element preference of collaborative modules on gene expression or methylation, which supported the necessity of integrative analysis on omics data, such as done by DAM.

On the functional level, there were six pathways observed in many modules (Table 1). Three pathways have been confirmed to be closely related to the occurrence of a Tachykinins: receptor to effector asthma. Zhu et al. demonstrated that macrophage migration inhibitory factor is involved in the pathogenesis of asthma [29]. Corinna Braun et al. revealed the existence of a vinculin-binding sequence in CPn0572, a TarP family member of Chlamydia pneumoniae closely related to asthma [30]. And Tachykinins related to intra-species interaction have also been shown to be involved in developing asthma-like drugs [31]. Thus, essential collaborative modules have remarkable biological significances, indicating the biological interpretability of DAM (Table 2).

Besides, the features contained in the collaborative modules with small reconstruction error should be more accurate in representing the original data, thus, the reconstruction performance of selected essential collaborative modules were comprehensively evaluated as shown in Fig. 1b-1i and Fig. 1j-1q. On one hand, the DSR effect of genes in module 25 and module 75 is better; while the DSR effect of methylation loci in module 12 and module 31 is better. On the other hand, there are better JDSNMF effect of genes in module 25, module 31, and module 75; and better effect of methylation loci in four modules. According to the total sum of reconstruction errors for both genes and methylation loci displayed in Fig. S7, module 31 tended to have the least global reconstruction error, and was a key collaborative module for following diagnostic model analysis.

3.5. Ranking of salient features in key collaborative module based on DOCCA for diagnostic model analysis

For the features in the key collaborative module 31, our proposed DOCCA algorithm was applied to assign feature weights considering the group-wide association in the CCA manner. The hyperparameter selection of DOCCA and the comparison with CCA, KCCA and gradKCCA on the test dataset were shown in Table 3 and the highest CCCs indicated the effectiveness of DOCCA. After taking the absolute value of the feature weights from DOCCA, they were sorted from high to low (Fig. 5h).

3.6. Comparison of the results of single-omics and multi-omics integrated analysis

To validate the advantage of integrating multiple omics data over using a single modality alone, we compared the performance of the DAM model when integrating gene and methylation data versus using each data type separately. In detail, the Pearson correlation coefficients before and after matrix factorization, and the AUC of the selected top markers, serve as baseline performance metrics for both multi-omics and single-omics analysis. Deep semi-supervised matrix factorization was applied to reconstruct the training sets for both gene and methylation loci. The Pearson correlation coefficients for gene and methylation loci before and after factorization were 0.9988 and 0.9971, respectively. By integrating the two data types, the corresponding correlation coefficients increased to 0.9995 for genes and 0.9982 for methylation loci. Cooperative modules were then set up and errors were calculated for each module (Fig. S8a-b). The errors for genes in Module 3 and methylation loci in Module 136 were minimal, with values of 1.36 and 10.52, respectively. For comparison, we evaluated the AUC of the diagnostic model by integrating both omics data types and analyzing the top 10 weighted markers of each omics data type separately. ROC curves for diagnostic models based on these genes and methylation loci are shown in Fig. S8c-d, with AUC values of 0.723 and 0.705, respectively. Thus, integrating

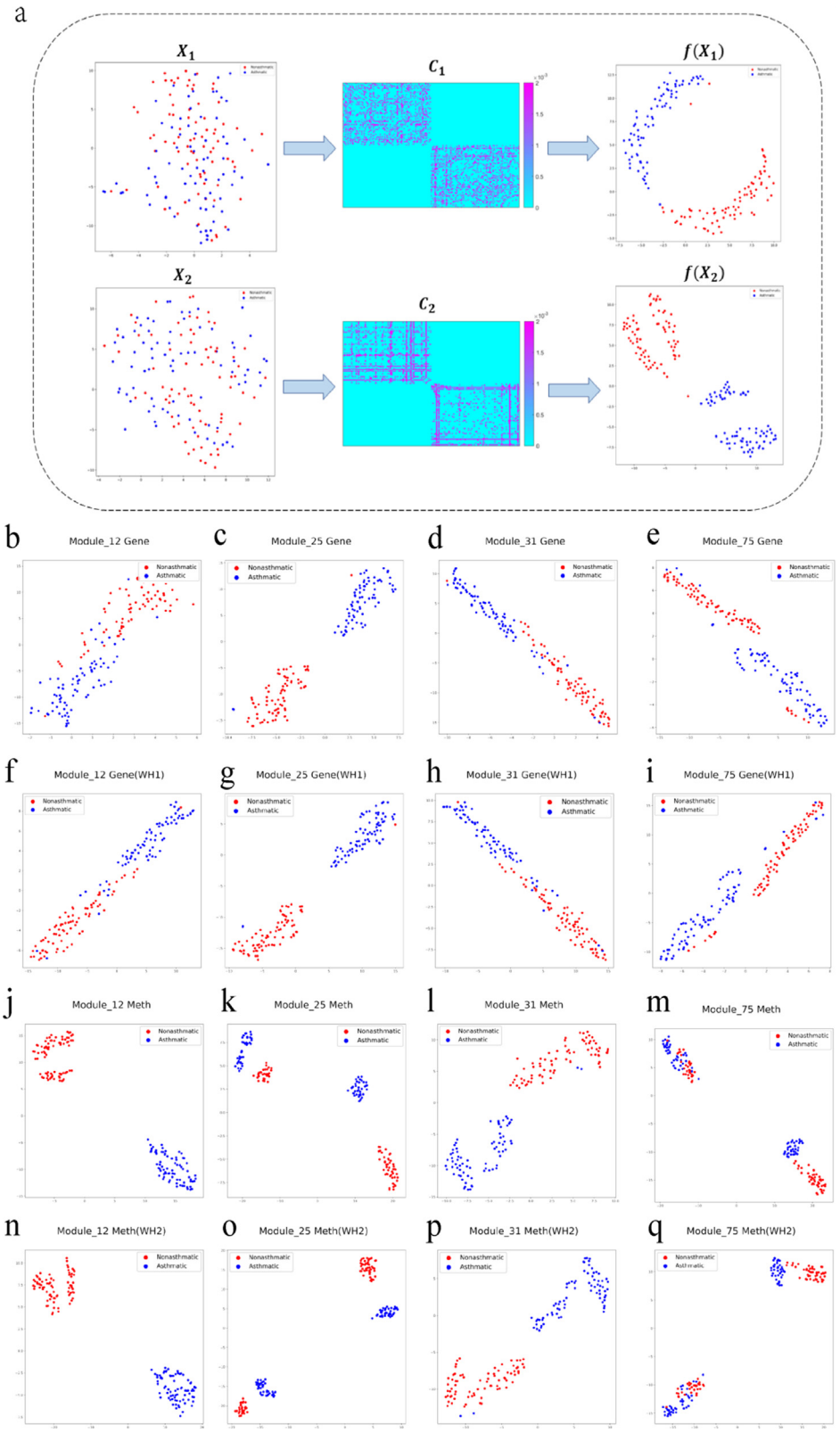

Fig. 1. Subspace reconstruction visualization. (a) Visualization of the original data and cleaned data by DSR using t-sne. X_1 and $f(X_1)$ are the original and cleaned differential gene expression matrices, respectively. X_2 and $f(X_2)$ are the original and cleaned differential methylation matrices, respectively. C_1 and C_1 are expression heatmaps of the self-expression coefficient matrix for genes and methylation, respectively. (b-i) The t-sne visualization of sample distribution based on cleaned gene expression data from DSR and the reconstructed data from JDSNMF in four collaborative modules. (b), (d), (f), and (h) are t-sne visualization of gene expression data from DSR. (c), (e), (g), and (i) are the t-sne visualization of the reconstructed gene expression data from JDSNMF. (j-q) The t-sne visualization of sample distributions based on cleaned DNA methylation data from DSR and reconstructed data from JDSNMF in four collaborative modules. (j), (l), (n), and (p) are t-sne visualization of cleaned DNA methylation data from DSR. (k), (m), (o), and (q) are the t-sne visualizations of the reconstructed DNA methylation data from JDSNMF.

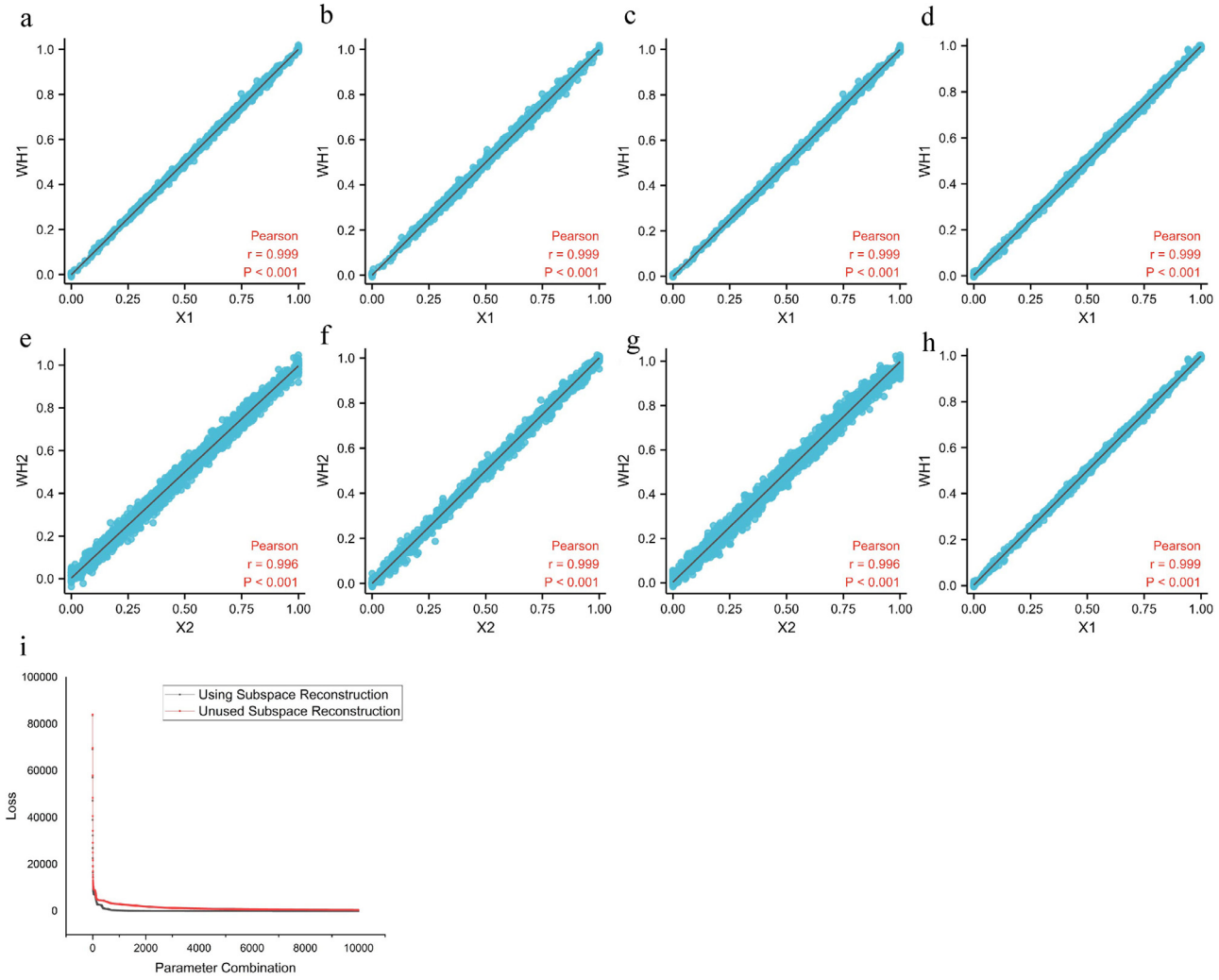

Fig. 2. Analysis of reconstruction performance of DSR and JDSNMF. (a-h) Scatter plots of the data reconstructed by subspace and the reconstructed data decomposed by the JDSNMF algorithm in the four modules. (a), (c), (e) and (g) are the scatter plots between the subspace-reconstructed data of genes and the data reconstructed by JDSNMF algorithm in module 12, module 25, module 31 and module 75, respectively. (b), (d), (f) and (h) are scatter plots between the subspace-reconstructed data of DNA methylation and the data reconstructed by JDSNMF algorithm in module 12, module 25, module 31 and module 75, respectively. (I) Influence of whether subspace reconstruction is used on the reconstruction error generated by the algorithm.

Table 1
Common pathways for the four modules.

Numbers and names of common pathways of modules 12, 31, and 75		Numbers and names of common pathways of modules 12, 25, and 31	
GO:0035176	social behavior	GO:0017166	vinculin binding
GO:0051703	intraspecies interaction between organisms	GO:0016289	CoA hydrolase activity
GO:1905517	macrophage migration	GO:0016790	thiolester hydrolase activity

both data types can achieve the best performance (Fig. S9q-r), i.e. the integration of multiple omics data outperforms single-omics analysis.

3.7. Evaluation of dam model performance with comparable multi-omics integration algorithms

To further validate the performance of the DAM model, in accordance with the literature review of predictive models for childhood asthma based on clinical indicators rather than omics data [32], we compared the performance of two machine learning models with superior prediction capabilities, namely, the Least Squares Support Vector Machine and the Multi-Layer Perceptron, with that of the DAM model (Fig. S9a-d). The AUC of the diagnostic models built on the DAM model for two data types is higher than that of the above two algorithms. In addition, three multi-omics integration analysis algorithms based on different technologies (CIMLR, MOGONET, and MTSCCALR) are presented in this study. The CIMLR algorithm learns the similarity between each pair of samples in multi-omics datasets by combining multiple Gaussian kernels for each omics data type, corresponding to different complementary representations of the data [33]. It enforces a block structure in the generated similarity matrix, which is then utilized for dimensionality reduction, k-means clustering, and feature selection. The MOGONET algorithm employs graph convolutional networks for omics-specific learning [34]. MOGONET not only directly connects the label distributions of each omics data type, but also utilizes the View Correlation Discovery Network to explore cross-omics correlations in the label space, achieving effective multi-omics integration. MTSCCALR is a multi-task learning-based correlation analysis method that combines the advan-

Table 2
Top-ranked 10 gene/methylation loci based on weights.

Genes	Weight	Methylation loci	Weight
LARS1	0.8307	cg00456348	0.1496
WBP1L	0.5583	cg00053393	0.1175
NLRP12	0.4887	cg00313876	0.1058
HIBCH	0.4273	cg00322319	0.0932
RPUSD3	0.3967	cg00483304	0.0774
ASRGL1	0.3759	cg00266865	0.0757
MGST2	0.3130	cg00240732	0.0640
OAS2	0.2996	cg00146676	0.0638
CORO2B	0.2985	cg00610021	0.0533
TRIM5	0.1770	cg00594129	0.0504

Table 3
CCCs comparison of four CCA-based algorithms on the test set.

Algorithm	CCCs
CCA	0.2287
DOCCA	**0.2681**
KCCA	0.1067
gradKCCA	0.1307

tages of SCCA and logistic regression to jointly learn the correlation between two omics data types for multiple tasks [17]. Each task focuses on identifying a diagnosis-specific pattern. Since all three algorithms only provide feature weights, for the purpose of performance comparison, we present the impact of these three algorithms and the DAM model in building diagnostic models using top features (Fig. S9e-p). The DAM model achieves a slightly lower AUC than the MTSCCALR model when constructing a diagnostic model using gene expression data (Supplementary Material, Fig. S9q). However, the highest AUC is achieved when building a diagnostic model using DNA methylation data (Supplementary Material, Fig. S9r).

4. Discussion

4.1. Improvement of association analysis during multi-omics data integration by deep association model compared with typical methods

To verify the association analysis ability of DAM on multi-omics data, the reconstruction error and recovery PCC were evaluated on the test dataset under the same experimental conditions as the other compared methods. As shown in Table 4, DAM can achieve least reconstruction error and highest recovery PCC. Especially, DSR and JDSNMF have shown their great contributions in performance promotion of DAM. The nonlinear feature association and extraction strategy in DAM can fully integrate prior information, improving the reconstruction performance of DAM. As shown in Fig. 3a-b, DMA outperforms several other competing algorithms in reconstructing gene and methylation data.

In addition, according to the functional enrichments of modules detected by different methods illustrated in Fig. 3c-d, DAM can effectively find many collaborative modules with significant functional enrichments with GO and KEGG, indicating the improved biological interpretability by DAM considering nonlinear association in omics data.

4.2. Diverse biological significance and disease relevance of the key collaborative module

For the key collaborative module 31 linking gene and methylation signatures to asthma, the correlation heatmap of collaborative genes and DNA methylation loci was shown in Fig. 4a-b. There are strong correlations between the heterogeneous elements, and these features especially had more significant correlation pairs in the asthmatic group than those in the non-asthmatic group, confirming the effectiveness of DAM considering diagnostic information fusion.

Next, the gene and methylation expression pattern of module 31 could remarkably distinguish the asthmatic and non-asthmatic groups on the test dataset as shown in Fig. 4f. The expression levels of six genes in module 31 were significantly different between two groups. Indeed, the proportion of differentially expressed genes in module 31 was 58.33%, while such proportion is 16.26% (73 genes) in all genes, confirming again that JDSNMF in DAM can efficiently identify discriminative patterns of the two groups by multi-layer nonlinear transformation.

Then, the differential PPI network among those key module genes and genes where the methylation sites are located were extracted for asthmatic and non-asthmatic group respectively. As shown in Fig. 4c-e, the significantly co-expressed gene pairs of asthmatic groups were different from those of non-asthmatic group, and PVT1, MACF1, LIMA1, MGST2, TRIM5, NIPSNAP1, WBP1L, CORO2B, OAS2, WBSCR17 and DCC were only presented in the PPI network of asthmatic group, indicating their relevant functional roles in asthmatic condition.

Furthermore, the functional enrichment analysis showed many asthmas associated pathways of genes and methylation loci respectively from module 31, as seen in Fig. 4g-h.

On one hand, key genes related to asthma would involve in excessive airway hyperresponsiveness and inflammation, and lipopolysaccharide exposure is associated with disease severity and steroid resistance [35]. Airway hyperresponsiveness is independent of various Th2 cytokines and their signaling pathways but is dependent on interferon-γ [36]. Neurotransmitter-triggered calcium signaling induces actomyosin-mediated contraction of airway smooth muscle, and the resulting shortening of cells leads to airway narrowing, which induces asthma [37]. Yin et al. confirmed that Transgelin-2, an actin-binding protein, can relax the myosin cytoskeleton of airway smooth muscle cells by acting as a receptor for extracellular metallothionein-2, which may be used as a treatment for asthma [38]. PD-L1, an immune checkpoint molecule associated with viral escape from the host immune system, is an immune checkpoint molecule in which double-stranded RNA from viruses induces host immune responses and plays a role in a persistent viral infection leading to exacerbation of asthma or chronic obstructive pulmonary disease [39]. Jayalatha et al. has confirmed that the receptor interleukin-1 receptor-like-1 (IL-1RL1) is a susceptibility gene for childhood asthma, and the IL-1RL1 gene transcript encodes different isoforms generated by alternative splicing, whose soluble isoforms IL-1RL1-a inhibit IL1RL1-b/IL-33 signaling by sequestering IL-33 as a decoy receptor [40].

On the other hand, the genes where the methylation sites are located were also involved in many pathways linked to the occurrence of asthma. Recent research has shown that microRNAs play multiple roles in regulating airway smooth muscle phenotypes, including cell proliferation and size, that play a key role in asthma pathogenesis [41]. Bianco et al. reviewed the studies on asthma patients' inhaled transmembrane ion transport modulators, which confirmed that some mechanisms of ion transmembrane transport are involved in regulating airway responses to various stimuli. Highly conductive calcium-sensitive potassium channels (BK+Ca) and ATP-sensitive potassium channels (K+ATP) play essential roles in airway smooth muscle cell, goblet cell function, and cytokine production [42]. The BK+Ca channel is also a promising new drug for treating airway allergic inflammation [43]. Thymic stromal lymphopoietin, and innate cytokine that plays a key pathogenic role in asthma, activates dendritic cells after its release from airway epithelial cells [44].

4.3. Efficient diagnostic model based on key collaborative module for distinguishing childhood asthma

The logistic regression model in IBM SPSS Statistics 2016 selected the top 10 genes/methylation sites from the key modules to construct a diagnostic model of childhood asthma. Using the top 12 genes, the AUC in the inner test set can reach 0.875 (Fig. 5a-b), and the AUC in the two outer test sets (GSE27011 and GSE40888) can reach 0.906 and 0.832,

Table 4
Omics data reconstruction performance comparison for DAM and other comparable methods.

Algorithm	Pearson correlation coefficient between X_1 and WH_1	Pearson correlation coefficient between X_2 and WH_2
Subspace reconstruction +JDSNMF (DAM)	**0.9995**	**0.9982**
JDSNMF	0.9899	0.9592
Subspace reconstruction +MDJNMF	0.9689	0.9714
MDJNMF	0.9350	0.9146
Subspace reconstruction +JCB-SNMF	0.9810	0.9800
JCB-SNMF	0.9436	0.9230
Subspace reconstruction +NSOJNMF	0.9812	0.9820
NSOJNMF	0.9432	0.9240
Subspace reconstruction +JNMF	0.9818	0.9824
JNMF	0.9432	0.9239

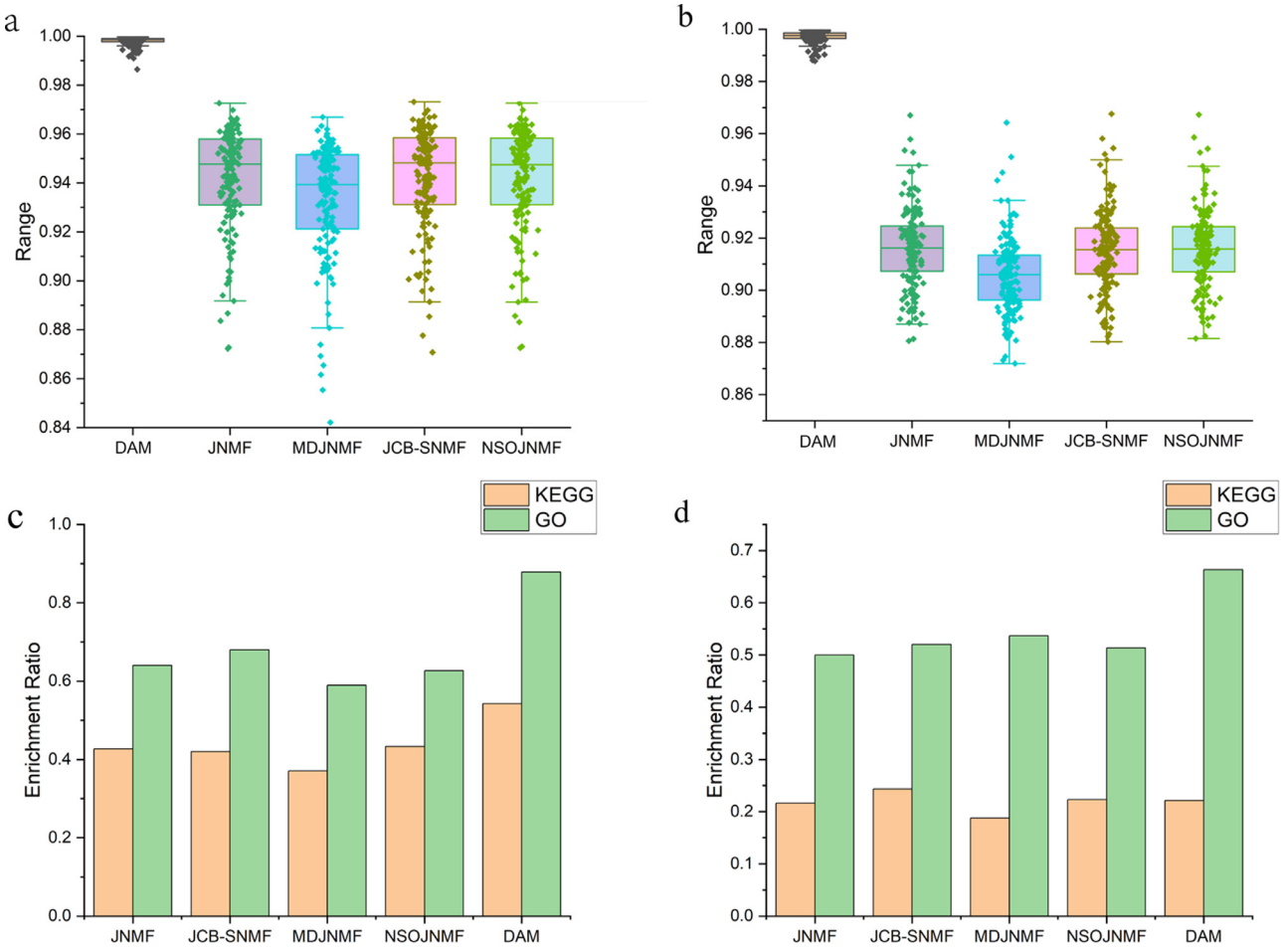

Fig. 3. Comparison of reconstruction performance and enrichment ratio with other algorithms. (a) and (b) are the reconstructed boxplots of genes and methylation, respectively. (c) and (d) are the histogram of KEGG and GO enrichment ratio of genes and methylation in all co-expression modules of several algorithms.

respectively (Fig. 5c-d). We observed obvious differences in model performance between two external test sets. This phenomenon may be due to biological disparities or data heterogeneity. Specifically, the cell types sampled in the two datasets are distinct, and there is no standardized criterion for disease severity. Batch effects may also be present, leading to variations in the distribution of dataset features and consequently affecting the generalization ability of the model. In future research, efforts will be made to use samples derived from the same cell type as the training set, and additional methods like transfer learning will be implemented to overcome possible batch effects, and ensure that the trained model has both generalization and applicability. The ROC of the model using the top 22 methylation sites can reach AUCs of 0.912 in the internal test sets, respectively (Fig. 5e-f), and 0.818 in the external validation set GSE109446 (Fig. 5g).

Especially, several top-ranked genes used in diagnostic model are strongly associated with asthma. Airway remodeling in asthma is characterized by thickening of the reticular basement membrane, which may be associated with altered epithelial structure and function. Among the genes associated with increased reticular basement membrane (RBM) thickness, OAS2 is one of the most critical genes in cell activation, proliferation, and growth [45]. TRIM5 [46], LIMA1 [47], NLRP12 [48] and CX3CR1 [49] have also been confirmed to be involved in multiple pathways related to inflammation .

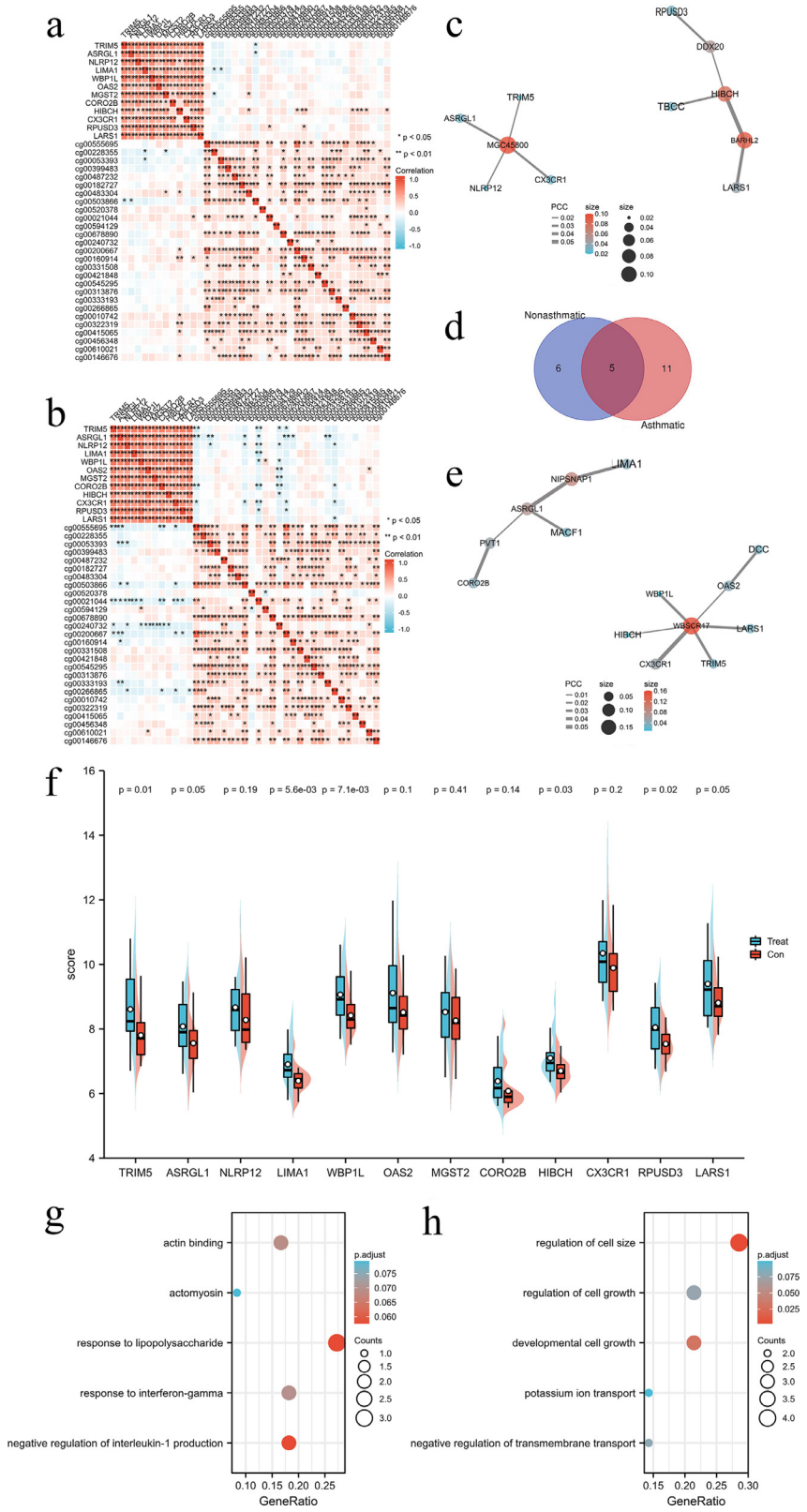

Fig. 4. Comprehensive Analysis of Module 31. (a) and (b) are the correlation heatmaps of genes and methylation loci in module 31 for the non-asthmatic group and the asthmatic group, respectively. (c) PPI network for module 31 in the non-asthmatic group. (d) Venn diagram of overlapping genes. (e) PPI network for module 31 in the asthmatic group. (f) Boxplot of gene expression in module 31. (g) and (h) are the GO enrichment results of genes and methylation-driven genes.

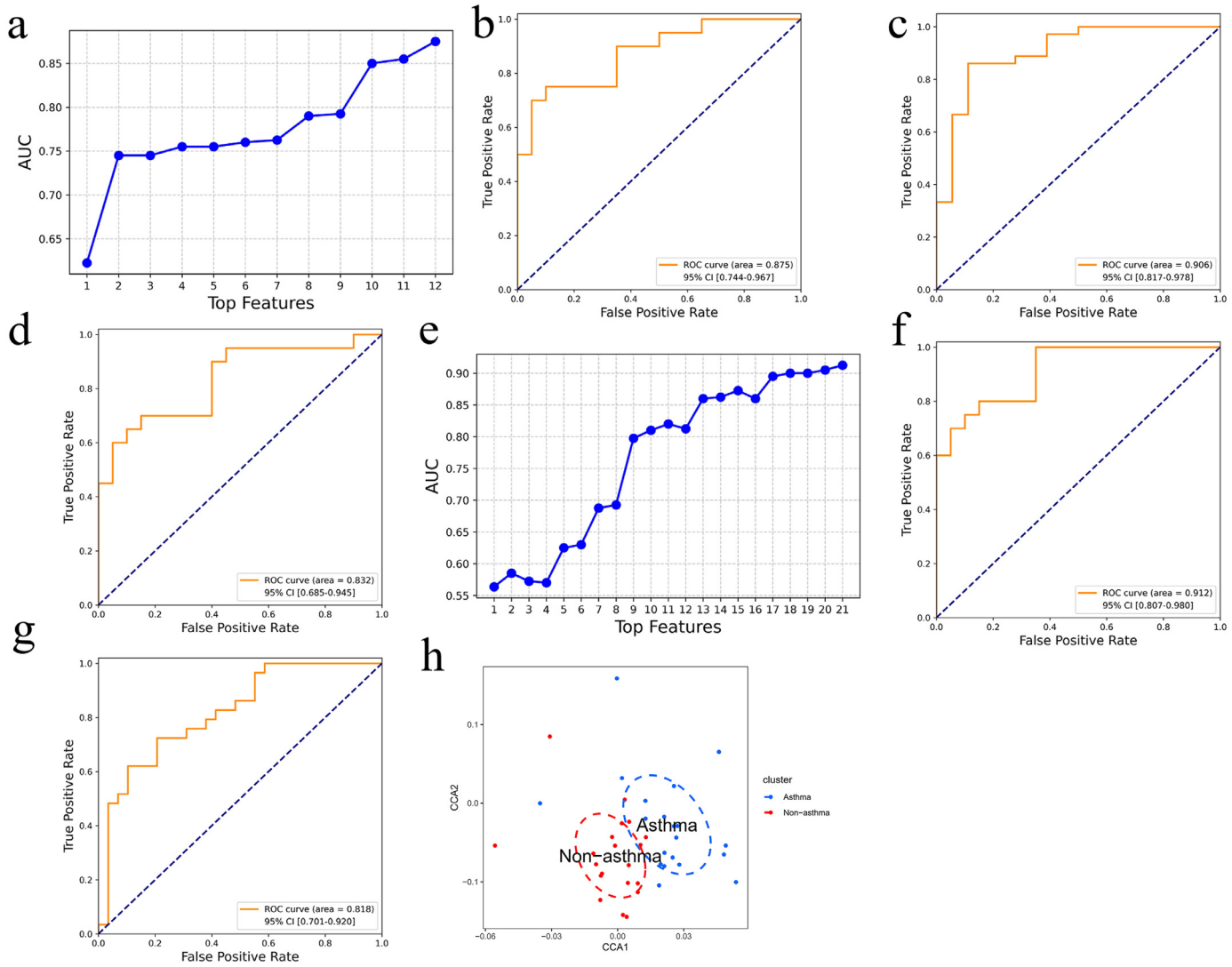

Fig. 5. The **features selected by the DOCCA algorithm are applied to the construction of the diagnostic model.** The diagnostic model constructed using top-ranked genes is shown in (a-d). In (a), the line graph illustrates the AUC variation in the internal test set corresponding to different numbers of top genes. (b-d) present the AUC values for the internal test set and two external test sets (GSE27011 and GSE40888) based on genes. On the other hand, (e-g) represent the diagnostic model built using the highest ranked methylation sites. In (e) and (f), the AUC values for the internal training and test set are shown, respectively. (g) displays the AUC in the external validation set (GSE109446). Specifically, the abscissa of (h) represents $f(X_1^{[31]})u$, and the ordinate represents $f(X_2^{[31]})v$. The two colored dots represent two types of samples, respectively. Two kinds of sample points are distinguished using ellipses.

4.4. Feasibility of DAM for multi-omics data analysis

Besides, the proposed DAM has great feasibility for multi-omics data analysis in the new single-cell fields [50–52]. Taking the integration of gene expression profiles (scRNA-seq) and epigenome profiles (scATAC-seq) as an example, DAM would also be able to identify cell types and signatures, together with the associations between scRNA and scATAC elements/features.

Let $X_1 \in \mathbb{R}^{n \times p}$ be the normalized expression profile of scRNA-seq data, where the columns correspond to p genes, and the rows represent n cells. Similarly, let $X_2 \in \mathbb{R}^{n \times q}$ be the indicator matrix of q ATAC peaks (regions) in n cells. Given DAM, the cell-related information can be introduced at the first stage of DSR, such as calculating the score of each cell involved in the pathway by the single-sample gene set enrichment analysis (ssGSEA) and setting a threshold to discretize the score to be new input data for following analysis. Then in the second stage, cell clustering can be achieved by JDSNMF, where the objective function can be formulated as bellows:

$$\begin{aligned} &\min \sum_{i=1}^{2} \|X_i - U_0 H_i\|_F^2 + \lambda \|S\|_F \\ &s.t.\ U_0 = s\left(s\left(U_n Z_n\right) \ldots Z_2\right) Z_1\right) \\ &U_{n-1} = s\left(U_n Z_n\right), U_0 \ldots U_n \geq 0, \\ &S \in \left\{U_n, Z_{i_1}, \ldots, Z_{i_n}, H_i\right\}, i = 1, 2; n = 1, 2, \ldots \end{aligned} \tag{26}$$

Among them, n represents the number of layers of decomposition, $U_0 \in \mathbb{R}^{n \times k_0}$ is the cell (sample) latent matrix, $H_1 \in \mathbb{R}^{k_0 \times p}$ and $H_2 \in \mathbb{R}^{k_0 \times q}$ are the feature latent matrix of the first layer of the neural network, $Z_n \in \mathbb{R}^{k_n \times k_{n-1}}$ is the junction latent matrix, and $U_0 \in \mathbb{R}^{k_n \times n}$ can be used for cell clustering and cell type identification, and k_n is the number of clusters.

4.5. Scalability of the DAM model

Multidimensional omics data is often at risk of overfitting due to small sample sizes and high feature dimensions. In situations with extremely high feature dimensions, feature selection becomes imperative. The DAM model, based on omics features that have differential expression between two groups, has shown excellent performance. When the dimensionality of differentially expressed features remains high, various feature selection strategies can be used. For instance, the variance threshold method filters out features whose variance is below a predefined threshold, thereby excluding features that have minimal variation across the entire dataset [53]. Recursive feature elimination can recursive remove features that are considered least important by the model until the specified number of features is reached [54]. To validate the scalability and computational efficiency of the DAM model, we also generated simulated datasets for different omics types. The study calculated the running time and performance variations of the DAM model under different sample sizes and feature dimensions (Supplementary Material 1.6). The results indicate that the DAM model has fast operation efficiency and high performance even with a sample size of 3000 and a feature count of 7000 for four omics data types.

5. Conclusion

Asthma is a common respiratory disease in children, and its pathogenesis is closely related to airway inflammation. Based on the multi-omics data of childhood asthma, a deep association model was designed and implemented explore the collaborative module and biomarkers, which help building efficient diagnostic model for asthma. Different omics datasets can carry complementary molecular information and specific prior clinical information [55–57]. First, both gene expression and DNA methylation provide insights into various aspects of gene functionality and regulation. By synthesizing information from these two molecular levels, a more comprehensive understanding of the active states of genes and the biological mechanisms associated with the onset of asthma can be achieved. Second, pediatric asthma is a complex disease involving the regulation of multiple genes and biological processes. Relying solely on gene expression or DNA methylation may not capture this complexity. The Integration of information from both levels enables a more holistic understanding of the molecular mechanisms associated with the development of asthma. Finally, integrating gene expression and DNA methylation data helps consider information at multiple molecular levels, reducing the impact of noise or variation resulting from a single data source and increasing the reliability of diagnosis. The algorithms involved in DAM are scalable and can be flexibly extended to the integration analysis of various omics data. It can also solve the multi-classification problem by cooperating with the classifier [58]. A future work should integrate new types and domains of multi-omics data [59] to conduct a more comprehensive and systematic nonlinear association analysis of biological system involved in childhood asthma or other complex diseases.

Availability

DAM is publicly available at https://github.com/babykai12345/DAM.

Authors' contributions

Conception and design of the research: Kai Wei and Tao Zeng. Acquisition, analysis, and interpretation of data: Kai Wei, Fang Qian and Tao Huang. Statistical analysis: Fang Qian. Molecular biological analysis: Kai Wei and Fang Qian. Drafting the manuscript: Tao Huang and Tao Zeng. Manuscript revision for important intellectual content: Yixue Li. All authors have read and approved the manuscript.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

This paper was supported by the Self-supporting Program of Guangzhou Laboratory (SRPG22-007); R&D Program of Guangzhou National Laboratory (GZNL2024A01002); National Natural Science Foundation of China (12371485,11871456); II Phase External Project of Guoke Ningbo Life Science and Health Industry Research Institute (2020YJY0217); Science and Technology Project of Yunnan Province (202103AQ100002); National Key R&D Program of China (2022YFF1202100); The Strategic Priority Research Program of the Chinese Academy of Sciences (XDB38050200, XDB38040202, XDA26040304).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2024.03.022.

References

[1] K. Golebski, M. Kabesch, E. Melén, et al., Childhood asthma in the new omics era: Challenges and perspectives, Curr. Opin. Allergy Clin. Immunol. 20 (2020) 155–161.
[2] M.M. Soliai, A. Kato, B.A. Helling, et al., Multi-omics colocalization with genome-wide association studies reveals a context-specific genetic mechanism at a childhood onset asthma risk locus, Genome Med. 13 (2021) 157.
[3] E. Forno, T. Wang, Q. Yan, et al., A multiomics approach to identify genes associated with childhood asthma risk and morbidity, Am. J. Respir. Cell Mol. Biol. 57 (2017) 439–447.
[4] L. Zhang, S.A. Zhang, Unified joint matrix factorization framework for data integration, arXiv (2017). https://doi.org/10.48550/arXiv.1707.08183.
[5] R.S. Kelly, B.L. Chawes, K. Blighe, et al., An integrative transcriptomic and metabolomic study of lung function in children with asthma, Chest 154 (2018) 335–348.
[6] X.T. Yu, T. Zeng, Integrative analysis of omics big data, Methods Mol. Biol. 1754 (2018) 109–135.
[7] J. Hu, T. Zeng, Q. Xia, et al., Identification of key genes for the ultrahigh yield of rice using dynamic cross-tissue network analysis, Genomics. Proteomics. Bioinformatics. 18 (2020) 256–270.
[8] C. Zhang, Y. Chen, T. Zeng, et al., Deep latent space fusion for adaptive representation of heterogeneous multi-omics data, Brief. Bioinform. 23 (2022) bbab600.
[9] S. Zhang, C.C. Liu, W. Li, et al., Discovery of multi-dimensional modules by integrative analysis of cancer genomic data, Nucleic Acids Res. 40 (2012) 9379–9391.
[10] J. Deng, W. Zeng, W. Kong, et al., Multi-constrained joint non-negative matrix factorization with application to imaging genomic study of lung metastasis in soft tissue sarcomas, IEEE Trans. Biomed. Eng. 67 (2020) 2110–2118.
[11] J. Deng, W. Zeng, S. Luo, et al., Integrating multiple genomic imaging data for the study of lung metastasis in sarcomas using multi-dimensional constrained joint non-negative matrix factorization, Inf. Sci. 576 (2021) 24–36.
[12] J. Deng, W. Kong, S. Wang, et al., Prior knowledge driven joint NMF algorithm for ceRNA co-module identification, Int. J. Biol. Sci. 14 (2018) 1822–1833.
[13] Y. Wang, G. Zhou, T. Guan, et al., A network-based matrix factorization framework for ceRNA co-modules recognition of cancer genomic data, Brief. Bioinform. 23 (2022) bbac154.
[14] K. Wei, W. Kong, S. Wang, Integration of imaging genomics data for the study of Alzheimer's disease using joint-connectivity-based sparse nonnegative matrix factorization, J. Mol. Neurosci. 72 (2022) 255–272.
[15] D. Salazar, J. Rios, S. Aceros, et al., Kernel joint non-negative matrix factorization for genomic data, IEEE Access. 9 (2021) 101863–101875.
[16] H. Lee, JDSNMF: Joint deep semi-non-negative matrix factorization for learning integrative representation of molecular signals in Alzheimer's disease, J. Pers. Med. 11 (2021) 686.
[17] L. Du, F. Liu, K. Liu, et al., Identifying diagnosis-specific genotype-phenotype associations via joint multitask sparse canonical correlation analysis and classification, Bioinformatics. 36 (2020) i371–i379.
[18] L. Du, K. Liu, X. Yao, et al., Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach, Med. Image Anal. 61 (2020) 101656.
[19] T. Melzer, M. Reiter, H. Bischof, Kernel Canonical Correlation Analysis, John Wiley & Sons, Inc, 2001.
[20] V. Uurtio, S. Bhadra, and J. Rousu (2019), *36th International conference on machine learning*.
[21] M. Wang, W. Shao, X. Hao, et al., Identify connectome between genotypes and brain network phenotypes via deep self-reconstruction sparse canonical correlation analysis, Bioinformatics 38 (2022) 2323–2332.

[22] T. Barrett, D.B. Troup, S.E. Wilhite, et al., NCBI GEO: Mining tens of millions of expression profiles–database and tools update, Nucleic Acids Res. 35 (2007) D760–D765.
[23] I.V. Yang, B.S. Pedersen, A. Liu, et al., DNA methylation and childhood asthma in the inner city, J. Allergy Clin. Immunol. 136 (2015) 69–80.
[24] N. Pezzotti, B.P.F. Lelieveldt, L. Van Der Maaten, et al., Approximated and user steerable tSNE for progressive visual analytics, IEEe Trans. Vis. Comput. Graph. 23 (2017) 1739–1752.
[25] T. Zeng, S.Y. Sun, Y. Wang, et al., Network biomarkers reveal dysfunctional gene regulations during disease progression, FEBS. J. 280 (2013) 5682–5695.
[26] G. Yu, Gene ontology semantic similarity analysis using GOSemSim, Methods Mol. Biol. 2117 (2020) 207–215.
[27] M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30.
[28] G. Yu, L.G. Wang, G.R. Yan, et al., DOSE: An R/Bioconductor package for disease ontology semantic and enrichment analysis, Bioinformatics. 31 (2015) 608–609.
[29] H. Zhu, S. Yan, J. Wu, et al., Serum macrophage migration inhibitory factor as a potential biomarker to evaluate therapeutic response in patients with allergic asthma: an exploratory study, J. Zhejiang. Univ. Sci. B 22 (2021) 512–520.
[30] C. Braun, A.R. Alcázar-Román, A. Laska, et al., CPn0572, the C. pneumoniae ortholog of TarP, reorganizes the actin cytoskeleton via a newly identified F-actin binding domain and recruitment of vinculin, PLoS One 14 (2019) e0210403.
[31] A.M. Khawaja, D.F. Rogers, Tachykinins: receptor to effector, Int. J. Biochem. Cell Biol. 28 (1996) 721–738.
[32] D.M. Kothalawala, L. Kadalayil, V.B.N. Weiss, et al., Prediction models for childhood asthma: a systematic review, Pediatr. Allergy Immunol. 31 (2020) 616–627.
[33] D. Ramazzotti, A. Lal, B. Wang, et al., Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival, Nat. Commun. 9 (2018) 4453.
[34] T. Wang, W. Shao, Z. Huang, et al., MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification, Nat. Commun. 12 (2021) 3445.
[35] L. Wang, K.G. Netto, L. Zhou, et al., Single-cell transcriptomic analysis reveals the immune landscape of lung in steroid-resistant asthma exacerbation, Proc. Natl. Acad. Sci. U.S.A. 118 (2021) e2005590118.
[36] R.K. Kumar, M. Yang, C. Herbert, et al., Interferon-γ, pulmonary macrophages and airway responsiveness in asthma, Inflamm. Allergy Drug Targets 11 (2012) 292–297.
[37] M.H. Jo, B.C. Kim, K. Sung, et al., Molecular nanomechanical mapping of histamine-induced smooth muscle cell contraction and shortening, ACS Nano 15 (2021) 11585–11596.
[38] L.M. Yin, L. Ulloa, Y.Q. Yang, Transgelin-2: biochemical and clinical implications in cancer and asthma, Trends Biochem. Sci. 44 (2019) 885–896.
[39] N. Seki, O.K. Kan, K. Matsumoto, et al., Interleukin-22 attenuates double-stranded RNA-induced upregulation of PD-L1 in airway epithelial cells via a STAT3-dependent mechanism, Biochem. Biophys. Res. Commun. 494 (2017) 242–248.
[40] A.K. Saikumar Jayalatha, L. Hesse, M.E. Ketelaar, et al., The central role of IL-33/IL-1RL1 pathway in asthma: From pathogenesis to intervention, Pharmacol. Ther. 225 (2021) 107847.
[41] M. Sun, Q. Lu, MicroRNA regulation of airway smooth muscle function, Biol. Chem. 397 (2016) 507–511.
[42] S. Bianco, M. Robuschi, A. Vaghi, et al., Inhaled transmembrane ion transport modulators and non-steroidal anti-inflammatory drugs in asthma, Thorax 55 (2) (2000) S48–S50 **Suppl**.
[43] M. Kocmalova, M. Oravec, M. Adamkov, et al., Potassium ion channels and allergic asthma, Adv. Exp. Med. Biol. 838 (2015) 35–45.
[44] C. Pelaia, G. Pelaia, C. Crimi, et al., Tezepelumab: A potential new biological therapy for severe refractory asthma, Int. J. Mol. Sci. 22 (2021) 4369.
[45] S. Bazan-Socha, S. Buregwa-Czuma, B. Jakiela, et al., Reticular basement membrane thickness is associated with growth- and fibrosis-promoting airway transcriptome profile-study in asthma patients, Int. J. Mol. Sci. 22 (2021) 998.
[46] B. Saha, M.A. Mandell, The retroviral restriction factor TRIM5/TRIM5α regulates mitochondrial quality control, Autophagy 19 (2023) 372–373.
[47] M.W. Su, C.K. Chang, C.W. Lin, et al., Blood multiomics reveal insights into population clusters with low prevalence of diabetes, dyslipidemia and hypertension, PLoS One 15 (2020) e0229922.
[48] Q. Li, K.J. Baines, P.G. Gibson, et al., Changes in expression of genes regulating airway inflammation following a high-fat mixed meal in asthmatics, Nutrients 8 (2016) 30.
[49] S.A. Doggrell, CX3CR1 as a target for airways inflammation, Expert. Opin. Ther. Targets 15 (2011) 1139–1142.
[50] T. Zeng, H. Dai, Single-cell RNA sequencing-based computational analysis to describe disease heterogeneity, Front. Genet. 10 (2019) 629.
[51] H. Tang, X. Yu, R. Liu, et al., Vec2image: An explainable artificial intelligence model for the feature representation and classification of high-dimensional biological data by vector-to-image conversion, Brief. Bioinform. 23 (2022) bbab584.
[52] K. Yuan, T. Zeng, L. Chen, Interpreting functional impact of genetic variations by network QTL for genotype-phenotype association study, Front. Cell Dev. Biol. 9 (2021) 720321.
[53] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[54] I.M. Guyon, J. Weston, S.D. Barnhill, et al., Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.
[55] W.F. Guo, X. Yu, Q.Q. Shi, et al., Performance assessment of sample-specific network control methods for bulk and single-cell biological data analysis, PLoS Comput. Biol. 17 (2021) e1008962.
[56] W.F. Guo, S.W. Zhang, Y.H. Feng, et al., Network controllability-based algorithm to target personalized driver genes for discovering combinatorial drugs of individual patients, Nucleic Acids Res. 49 (2021) e37.
[57] X. Yu, J. Zhang, S. Sun, et al., Individual-specific edge-network analysis for disease prediction, Nucleic Acids Res. 45 (2017) e170.
[58] J. Liang, Z.W. Li, Z.N. Sun, et al., Latent space search based multimodal optimization with personalized edge-network biomarker for multi-purpose early disease prediction, Brief. Bioinformatics 24 (2023) bbad364.
[59] Y. Liu, Y. Li, T. Zeng, Multi-omics of extracellular vesicles: an integrative representation of functional mediators and perspectives on lung disease study, Front. Bioinform. 3 (2023) 111727.

Kai Wei graduated from Shanghai DianJi University with a bachelor's degree in 2019. He graduated from Shanghai Maritime University with a master's degree in engineering in 2022. He currently works as an assistant engineer at the Biomedical Big Data Center of the Institute of Nutrition and Health, Chinese Academy of Sciences. His main research interests are in the field of bioinformatics, including transcriptomic data, imaging data, and single-cell multi-omics data analysis

Tao Huang (BRID: 05238.00.11652) is a professor at Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences. He completed his post-doctoral research at Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York City, USA. His research interests include bioinformatics, computational biology, systems genetics and big data research. He has published over 200 articles. His works have been cited for 14,877 times with an h-index of 54. He has edited books of Computational Systems Biology: Methods and Protocols, Precision Medicine, and Liquid Biopsies: Methods and Protocols for Methods in Molecular Biology series. He has been Editors or Guest Editors for over 30 journals and Reviewer for 270 journals. He is Highly Cited Chinese Researcher and World's Top 2% Scientist (2020, 2021, 2022, 2023).