JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

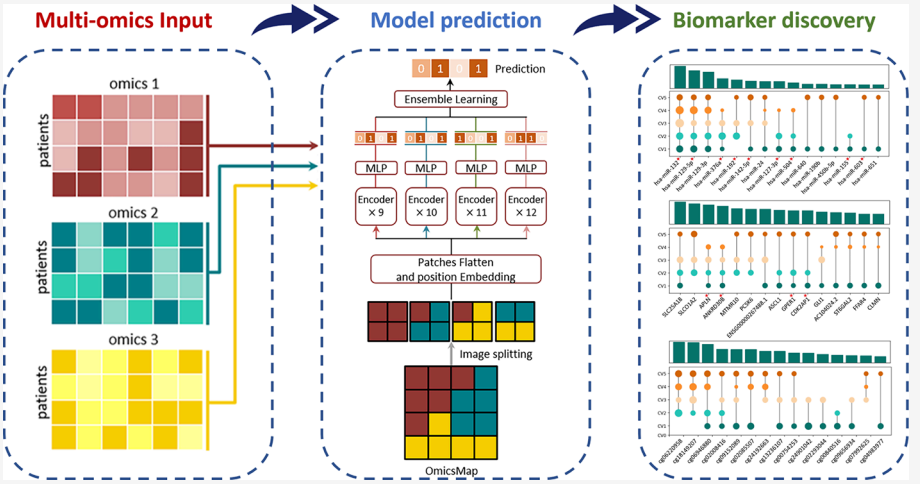
pubs.acs.org/jcim Article

MOINER: A Novel Multiomics Early Integration Framework for Biomedical Classification and Biomarker Discovery

Wei Zhang, Minjie Mou, Wei Hu, Mingkun Lu, Hanyu Zhang, Hongning Zhang, Yongchao Luo, Hongquan Xu, Lin Tao, Haibin Dai, Jianqing Gao,* and Feng Zhu*

Cite This: J. Chem. Inf. Model. 2024, 64, 2720−2732 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: In the context of precision medicine, multiomics data integration provides a comprehensive understanding of underlying biological processes and is critical for disease diagnosis and biomarker discovery. One commonly used integration method is early integration through concatenation of multiple dimensionally reduced omics matrices due to its simplicity and ease of implementation. However, this approach is seriously limited by information loss and lack of latent feature interaction. Herein, a novel multiomics early integration framework (MOINER) based on information enhancement and image representation learning is thus presented to address the challenges. MOINER employs the self-attention mechanism to capture the intrinsic correlations of omics-features, which make it significantly outperform the existing state-of-the-art methods for multiomics data integration. Moreover, visualizing the attention embedding and identifying potential biomarkers offer interpretable insights into the prediction results. All source codes and model for MOINER are freely available https://github.com/idrblab/MOINER.

Downloaded via ZHEJIANG UNIV on April 12, 2024 at 00:58:22 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Given the rapid progress in high-throughput biomedical sequencing methodologies, it has become increasingly easy to access multiple omics (multiomics) data (mRNA expression, DNA methylation, microRNA expression, protein expression, etc.) from national programs of genome research,[1−3] such as The Cancer Genome Atlas (TCGA)[4] and the International Cancer Genome Consortium (ICGC),[5] etc. While each omics data type is specific in revealing partial biological information, their integration cultivates a more comprehensive understanding of disease mechanisms[6−10] and facilitates the advancement of precision medicine.[11−13] However, improper integration approaches may introduce the complexity and computational cost of the problem.[6,14,15] Therefore, there is an urgent demand for methodologies to handle, standardize, and integrate heterogeneous multiomics data into a unified compendium. Such integration aims to capture complementary information, establishing a foundational platform for subsequent analysis and learning.[16−18]

Recently, a variety of strategies have been developed for unsupervised multiomics integration,[19] such as iCluster,[20] Similarity Network Fusion (SNF),[21] Multi-Omics Factor Analysis (MOFA),[22] SubtypeGAN,[23] DeepProg,[24] etc. These methods primarily address the tasks of subtype clustering and prognostic analysis; that is, they do not require prior

Special Issue: Machine Learning in Bio-cheminformatics

Received: January 3, 2024
Revised: February 8, 2024
Accepted: February 8, 2024
Published: February 19, 2024

© 2024 American Chemical Society

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

Table 1. Summary of Datasets[a]

	Data set	Categories	Types of Multiomics Data
Binary-class	PRAD[4]	Early stage: 319, Late stage: 206	mRNA: 60,483, meth: 22,185, miRNA: 1880
	ROSMAP[25]	NC: 169, AD: 182	mRNA: 55,889, meth: 23,788, miRNA: 309
	COVID-19[38]	COVID: 102, Non-COVID: 26	lipidomics: 3357, metabolomics: 150, protein: 517, mRNA: 13,263
Multiclass	BRCA[25]	Normal-like: 115, Basal-like: 131, HER2-enriched: 46, Luminal A: 436, Luminal B: 147	mRNA: 20,531, meth: 20,106, miRNA: 503

[a]The ROSMAP dataset is for the classification of Alzheimer's disease (AD) patients and normal control (NC). The PRAD dataset is for stage classification in prostate cancer (PRAD). The COVID-19 dataset is for the classification of COVID patients and non-COVID patients. The LUSC dataset is for stage classification in lung squamous cell carcinoma (LUSC). The BRCA dataset is for breast invasive carcinoma (BRCA) subtype classification with normal-like, basal-like, human epidermal growth factor receptor 2 (HER2)-enriched, Luminal A, and Luminal B subtypes.

knowledge of sample phenotypes. With the availability of data sets containing detailed sample phenotype annotations on the rise, there is a growing interest in supervised approaches for integrating multiomics data, enabling accurate predictions on uncharacterized cases.[25,26] So far, supervised integration methods include: (1) early integration methods that concatenate matrices of different omics data types, such as RDFS,[27] Stetson et al.,[28] and Fu et al.,[29] (2) intermediate integration methods that map diverse omics data into a shared space, such as MoGCN,[30] and (3) late integration methods that combine predictions from different omics data types using ensemble learning, such as MOGONET[25] and MOMA.[31] Compared to other integration methods, early integration has become the most commonly used method[6,32] for the reasons that it preserves the attributes of biometric measurements and is easy to implement.

However, early integration encounters two primary challenges in its application: (1) The raw high-dimensional data generated by concatenating all omics data is intricate, noisy, and redundant, leading to challenging learning processes and suboptimal model performance.[6,33] Existing methods[27,34] often employ feature selection algorithms to reduce the complexity of the composite matrix, which results in information loss as certain useful information is filtered out during the selection process.[21] (2) Another challenge lies in the fact that sequential high-dimensional multiomics vectors can hardly reflect the intrinsic correlations of omics-features from the representational level.[35] This limitation hinders the application of such data in advanced deep learning models, including 2D-CNN and Vision Transformer.[36,37]

To address these challenges, we propose MOINER, a novel multi-omics early integration framework based on information enhancement and image representation learning strategies. Specifically, all feature variables within the raw high-dimensional multiomics data are designated as a global feature set (GFS), while the feature subsets resulting from feature selection are designated as a local feature set (LFS). MOINER constructs a sample similarity network utilizing the GFS, wherein features within the LFS undergo information enhancement through neighborhood aggregation and message passing in this sample similarity network. Subsequently, the LFS is mapped to a regular 2D-map (omicsMap) by calculating the feature cosine similarity. Finally, an ensemble model of Vision Transformer (ViT) with different number of encoders (En-ViT) is employed for capturing intrinsic correlations of omics-variables in the omicsMap and conducting accurate label prediction for novel cases. To validate MOINER's effectiveness and adaptability, we performed a comprehensive performance comparison with other methods for integrating multiomics data on four biomedical categorization tasks: Alzheimer's disease (AD) patient categorization, breast carcinoma (BRCA) subtype categorization, prostate cancer (PRAD) grading categorization, and COVID-19 patient categorization. Our results indicate the superiority of MOINER over other state-of-the-art (SOTA) approaches while providing interpretable insights into prediction results through latent visualizing and biomarker discovery.

MATERIALS AND METHODS

Data Sets Collection. The superiority of MOINER was substantiated on four distinct biomedical classification tasks: PRAD for tumor grade classification in prostate cancer, ROSMAP for AD patients vs normal control, BRCA for breast invasive carcinoma PAM50 subtype classification, and COVID-19 for corona virus disease patients vs normal control. Specifically, preprocessed data sets of ROSMAP and BRCA were derived from a prior study,[25] each encompassing mRNA data, DNA methylation data, and miRNA data. For the PRAD data set, preprocessed mRNA data, DNA methylation data, miRNA data, and clinical annotation were sourced from the GDC TCGA PRAD on Xena. Patients with both mRNA data, DNA methylation data, and miRNA data were included. For the COVID-19 data set, mRNA data, proteins data, lipids data, and metabolites data were retrieved from MassIVE Summary (ID: MSV000085703). This data set was part of a cohort study conducted by Overmyer et al.[38] and encompasses 128 patients with and without a COVID-19 diagnosis. It facilitated a thorough and systematic analysis of blood samples from individuals affected by COVID-19. Table 1 provides detailed information on the four data sets.

Data Preprocessing. Features with zero mean values or low variances are filtered out first.[39,40] Then, chi-square (χ^2) feature selection is a supervised feature selection method that is commonly used in the feild of statistics and biomedical science. Specifically, it assesses the correlation between the feature and the real label by a chi-square test and then determines whether to select it. In order to make the selected features match the 2D grid map, which hold the same length of width and height, the number of features of each omics will be computed before feature selection. Similar to the study by Wang et al.,[25] the ROSMAP data set used 200 mRNA, 200 meth, and 200 miRNA, respectively, while the BRCA, PRAD, and COVID-19 data sets used 1000 mRNA, 1000 meth, and 500 miRNA, respectively. Finally, each feature is scaled to [0, 1] through linear transformations by using the sklearn package.

MOINER Construction. MOINER is proposed for multiomics integration and classification. This framework is composed of three main modules: (1) information enhancement module for reducing information loss of omics-features after feature selection, (2) image representation module for

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

capturing intrinsic correlations between omics-features, and (3) classification module for performing sample classification tasks.

Module 1: Information Enhancement Module. Information enhancement is performed through neighborhood aggregation and message passing in the sample similarity network (SSN). This process can be abstractly understood as updating data using a single-layer Graph Neural Network, with detailed explanations provided in the Supporting Information Figure S2. SNF algorithm[21] was utilized to build individual sample networks for each accessible omics, followed by their efficient fusion into a unified network (SSN). This integrated network encapsulates the complete information of raw data. Suppose that given n samples and m omics data types, for the v-th omics type, an n × n scaled sample similarity network $\mathbf{W}^{(v)}$ is calculated

$$\mathbf{W}^{(v)}(i, j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu\varepsilon_{i,j}}\right)$$

where x is a vector represented by the v-th omics type and $\rho(x_i,x_j)$ is the Euclidean distance between sample i and sample j. μ is a hyper-parameter that can be empirically set and $\varepsilon_{i,j}$ is used to eliminate the scaling problem. Then, a normalized sample weight matrix $\mathbf{P}^{(v)}$ and a K-nearest neighbors local affinity matrix $\mathbf{K}^{(v)}$ of the v-th omics type will be calculated. The detailed can refer to the original calculation steps of the SNF algorithm.[21]

In the case of there are two types of omics, the similarity matrix for each data type will be iteratively updated as follows

$$\mathbf{P}^{(1)}_{t+1} = \mathbf{K}^{(1)} \times \mathbf{P}^{(2)}_{t} \times (\mathbf{K}^{(1)})^{\mathrm{T}}$$

$$\mathbf{P}^{(2)}_{t+1} = \mathbf{K}^{(2)} \times \mathbf{P}^{(1)}_{t} \times (\mathbf{K}^{(2)})^{\mathrm{T}}$$

where $\mathbf{P}^{(1)}_{t+1}$ is the status matrix of the first omics type after t iterations and $\mathbf{P}^{(2)}_{t+1}$ is the status matrix of the second omics type. After t steps, the overall status matrix can be calculated as

$$\mathbf{P}^{(c)} = \frac{\mathbf{P}^{(1)}_{t} + \mathbf{P}^{(2)}_{t}}{2}$$

Given a sample matrix $\mathbf{S}\in\mathrm{R}^{\mathrm{n}\times\mathrm{m}}$ (n samples and m features), a new sample matrix $\mathbf{S}'$ will be calculated for fusing this SSN ($\mathbf{P}^{(c)}$) into the sample matrix $\mathbf{S}$.

$$\mathbf{S}' = \mathbf{P}^{(c)} \times \mathbf{S}$$

Module 2: Image Representation Module. A sample matrix $\mathbf{S}'\in\mathrm{R}^{\mathrm{n}\times\mathrm{m}}$ is generated from the information enhancement module; therefore, each feature is represented by an n-dimension vector $f\in\mathrm{R}^{\mathrm{n}}$. Then the sklearn package is applied to calculate feature similarity network $\mathbf{D} \in \mathrm{R}^{\mathrm{m}\times\mathrm{m}}$. The similarity between feature i and feature j is indicated by $\mathbf{D}$ (i,j) as follows.

$$\mathbf{D}(i, j) = 1 - \frac{f_i \cdot f_j}{\|f_i\|\|f_j\|}$$

Then, the UMAP or tSNE algorithm is used to reduce the matrix $\mathbf{D}$ to 2D space. The omics-features in this 2D space are further rearranged to a regular 2D-grid map using the J-V algorithm for linear assignment (Figure S3). The J-V algorithm optimally determines the solution by minimizing distance between the 2D scatter and the regular grid and generates a prelearned map reflecting the intrinsic correlations between omics-features. Finally, the raw multiomics data is transformed into an image representation by rearranging each feature from different omics layers to a specific position according to this prelearned map (OmicsMap).

Module 3: Classification Module. The combination of image and deep learning models has made significant advancements in the field of biomedical research.[41,42] In this study, ViT[43] is used as the default model for conducting classification tasks utilizing multiomics images. We demonstrate that ViT achieves better performance compared to other image classifiers. After obtaining the OmicsMap ($\mathbf{X}\in\mathrm{R}^{H\times W}$) of multiomics image representation, it is divided into a series of flattened patches $\mathrm{X_p} \in \mathrm{R}^{N\times P^2}$, where $H \times W = N \times P^2$. N is the number of patches, H and W are the shape of the OmicsMap, and P is the shape of each patch. The patches are flattened and mapped to D dimensions with a trainable linear projection. Then position embedding is added to these patches while a class token is concatenated to the first patch; that is, the i-th 2D image is newly represented as follows.

$$\mathbf{z}_i = [\mathbf{x}_{\mathrm{class}}; \mathrm{X}^1_p\mathrm{E}; \mathrm{X}^2_p\mathrm{E}; \cdots; \mathrm{X}^N_p\mathrm{E}] + \mathbf{E}_{pos},\ \mathrm{E} \in \mathrm{R}^{(P^2)\times D},\ \mathbf{E}_{pos} \in \mathrm{R}^{(N+1)\times D}$$

The ViT encoder is structured with alternating layers of Transformer Encoder (TE)

$$\mathbf{z}_l = \mathrm{TE}(\mathbf{z}_{l-1}),\ l = 1, 2, 3, \ldots, L$$

where L is the number of TE blocks, $\mathbf{z}_{l-1}$ is the output of the $(l - 1)$-th TE block. The class token $\mathbf{z}_{\mathrm{L}}$ $[\mathbf{x}_{\mathrm{class}}]$ of the output from the last encoder block will be transferred into an MLP Head for the final prediction.

$$\hat{Y} = \mathrm{MLP}_{\mathrm{Head}(\mathbf{z}_{\mathrm{L}}[\mathbf{x}_{\mathrm{class}}])}$$

Ultimately, we deployed an ensemble model, En-ViT, comprised of ViT models with 9, 10, 11, and 12 encoding layers. This ensemble model facilitated robust and effective class predictions for new samples through a voting approach. Additional hyperparameters, including random state, learning rate, and num_mlp, were set to 0, 5e-5, and 2048, respectively.

Interpretability Assessment of MOINER. The capability of a deep learning model to identify potential biomarkers is critical to interpreting results and comprehending the intrinsic biology in biomedical contexts.[44] In this study, the significance of input features can be assessed through an importance score calculated using the mask strategy and the mean squared error (MSE). Specifically, the performance decrease after masking the features reflects the importance of these input features. Suppose that given a valid data set $\mathrm{S} \in \mathrm{R}^{\mathrm{n}\times\mathrm{m}}$, sample's label $\mathrm{Y}\in \mathrm{R}^{1\times\mathrm{n}}$ and a trained model ViT. For the feature $\mathrm{m_i}$, its importance can be computed as follows

$$\mathrm{Importance}_{\mathrm{m_i}} = \mathrm{MSE}(\mathrm{Y}, \mathrm{ViT}(\mathrm{S})) - \mathrm{MSE}(\mathrm{Y}, \mathrm{ViT}(\mathrm{S}^{\mathrm{m}}))$$

where $\mathbf{S}^{\mathrm{m}}$ represents the masked matrix after the i-th feature is replaced by 0 value.

Adjusted Rand index (ARI)[45] is used to evaluate the clustering performance of latent vectors, which reflects the degree of overlap between clustering results and actual labels. Specifically, clustering label K is generated for latent vectors using K-means clustering, and then we calculate RI based on actual labels

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

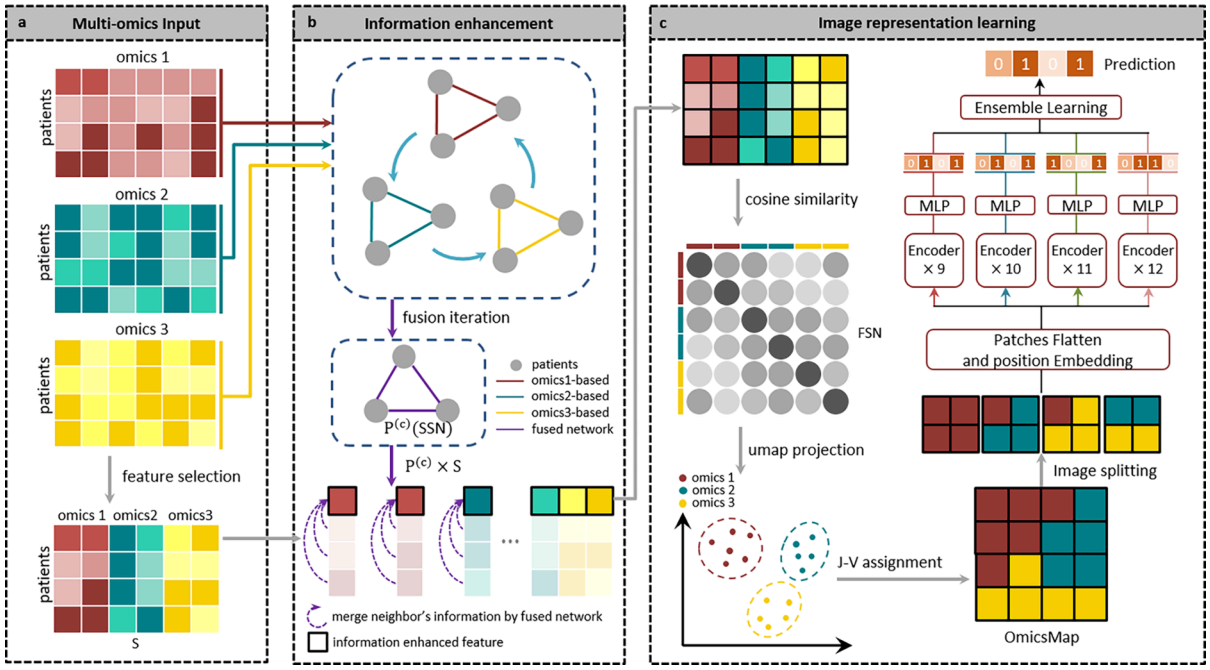


Figure 1. Overview of the MOINER. (a) Input processing during the application phase: MOINER necessitates that each sample possesses multiomics features concurrently. Dimensionality reduction is achieved through the application of feature selection to each omics data type. A single matrix is generated by concatenating all omics data. (b) MOINER employs neighborhood aggregation and message passing in a sample similarity network to minimize information loss. SNF constructs networks of patients for each omics type and then efficiently fuses these into a fused network. This fused network incorporates all features of a given input and provides a comprehensive representation of a patient cohort. The value of each feature is recalculated based on the weights in the fused network. (c) Image representation learning: A feature similarity network is constructed using cosine similarity in the concatenated multiomics matrix and projected into 2D-space. Each feature is then rearranged to a regular image using the J-V algorithm. In En-ViT learning, an image is divided into a sequence of flattened 2D patches and serves as input to multiple ViT models. The labels generated by these models are integrated through a voting mechanism to produce the final label prediction.

$$\mathrm{RI} = \frac{a + b}{\binom{n}{2}}$$

where a is defined as the count of instance pairs assigned to the same class in C and to the same cluster in K. b is defined as the count of instance pairs that are assigned to different classes in C and different clusters in K. The ARI is then calculated using the following formula.

$$\mathrm{ARI} = \frac{\mathrm{RI} - E(\mathrm{RI})}{\max(\mathrm{RI}) - E(\mathrm{RI})}$$

RESULTS AND DISCUSSION

Architecture of MOINER. Here, we propose the MOINER, a novel multiomics early integration framework for biomedical classification tasks and biomarker discovery (illustrated in Figure 1). Given a preprocessed multiomics data set, our approach initiates by utilizing SNF[21] to construct omics-specific sample similarity networks (SSNs) for each distinct omics layer. These SSNs are subsequently iteratively fused to formulate the ultimate fusion network. Concurrently, a feature selection method is applied for the raw multiomics input, effectively filtering out redundant and noisy features. These features are further enhanced by performing neighborhood aggregation and message passing in the SSN (as illustrated in Figures 1b, S2). Then, the matrix containing information-enhanced features is employed to construct a feature similarity network (FSN) by calculating the pairwise cosine similarity. FSN is projected into 2D space using the dimensionality reduction algorism (UMAP[46] or tSNE[47]), which is further assigned to a regular 2D-grid map (OmicsMap) by using the J-V algorithm.[48] As a result, all the features from different omics types will be rearranged to a specific position according to this prelearned OmicsMap. After image representation, the OmicsMap is split into a sequence of flattened 2D patches and forwarded to an ensemble learning framework (En-ViT), where ViT models use 9, 10, 11, and 12 encoding layers, respectively (as illustrated in Figure 1c). The detailed structure of ViT is presented in Figure S1. En-ViT effectively detects the variation in patches through a powerful self-attention mechanism and makes robust label prediction.

MOINER Demonstrates Superior Performance Compared to Established Supervised Multiomics Integration Methods across Diverse Classification Tasks. The classification performance of MOINER was compared with four SOTA supervised multiomics integration (SMI) methods and four traditional supervised machine learning (TML) methods: (1) MOGONET.[25] MOGONET employs a graph convolutional network (GCN) and a view correlation discovery network (VCDN) to investigate correlations across omics in the label space, facilitating efficient multiomics integration. (2) MoGCN.[30] MoGCN is an SMI method based on auto encoder and GCN. (3) RDFS.[27] RDFS is an SMI model that uses RF and deep neural network (DNN). (4) MOMA.[31] MOMA is a multiomics integration algorithm employing attention learning, demonstrating superior performance in categorizing phenotypes related to diseases. (5) K-nearest neighbor (KNN). (6) Random forest (RF). (7) Support vector machine (SVM). (8) Extreme gradient

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

Table 2. Summary of Comparison of Our Work with Other State-of-the-Art Multi-Omics Integration Methods

Model Name	Method	Category	Code availability	Reference
MoGCN	Graph convolutional networks	Supervised	https://github.com/Lifoof/MoGCN	Li et al. (2022)[30]
MOGONET	Graph convolutional networks	Supervised	https://github.com/txWang/MOGONET	Wang et al. (2021)[25]
RDFS	Feedforward neural networks	Supervised	https://github.com/huyy96/RDFS	Hu et al. (2022)[27]
MOMA	Feedforward neural networks	Supervised	https://github.com/dmcb-gist/MOMA	Moon et al. (2022)[31]
MOFA	Matrix factorization	Unsupervised	https://github.com/bioFAM/MOFA	Argelaguet et al. (2018)[22]
SNF	Network fusion	Unsupervised	https://github.com/maxconway/SNFtool	Wang et al. (2014)[21]
SubtypeGAN	Generative Adversarial Network	Unsupervised	https://github.com/haiyang1986/Subtype-GAN	Yang et al. (2021)[23]

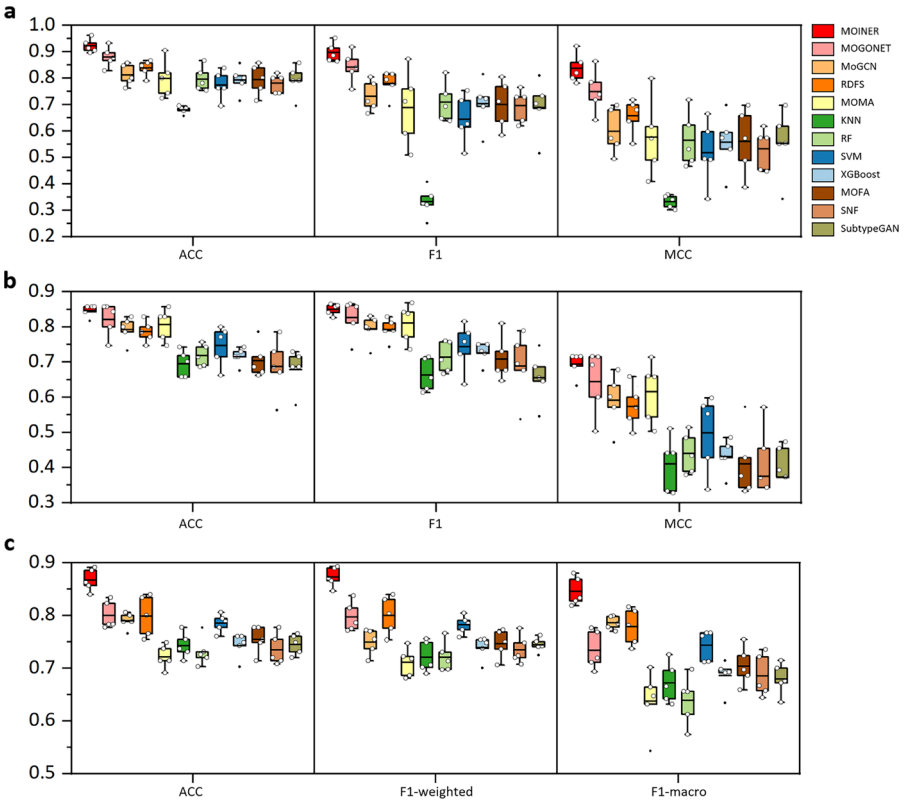


Figure 2. Performance comparison of multiomics integration methods by 5-fold cross-validation. (a) Results of the ROSMAP data set. (b) Results of the PRAD data set. (c) Results of the BRCA data set. ACC, F1, MCC for binary classification. ACC, F1-weighted, F1-macro for multiclass classification. Box plots show the median (center lines), interquartile range (hinges), and 1.5-times the interquartile range (whiskers), and black dots represent outliers. ACC: accuracy, MCC: Matthews correlation coefficient.

boosting (XGBoost). The details of these methods were listed in Table 2. To ensure comparability, MOGONET, MoGCN, RDFS, and MOMA underwent retraining utilizing multiomics data sets, as per the input formats specified in the original literature. KNN, RF, SVM, and XGBoost were trained with the concatenated matrix of the multiomics data. We used the above methods to perform stratified 5-fold cross validation (CV) on four data sets (ROSMAP, BRCA, PRAD, and COVID-19). The average accuracy (ACC), F1-score, and Matthews correlation coefficient (MCC) of 5-fold CV were used as the evaluation metrics for binary classification, while ACC, F1_weight, and F1_macro were used for multiclass classification. The subsequent findings highlight the superiority of the proposed MOINER when compared to alternative methods for supervised integration of multiomics data, particularly in terms of "effectiveness and robustness" and "extensibility and practicability".

The Effectiveness and Robustness of MOINER. As illustrated in Figure 2 and Supporting Information Tables S1, S2, and S3, MOINER demonstrated superior performance across all metrics in three multiomics data sets. Specifically, the ACC values of MOINER were 0.8405, 0.8674, and 0.9219 for ROSMAP, BRCA, and PRAD data sets, respectively, surpassing all other supervised integration methods. Among the SMI methods, there was no consistent superiority of one method over others. For instance, MORONET displayed better performance on ROSMAP and PRAD while MoGCN performed better on BRCA compared with other integration methods. In the binary classification tasks (ROSMAP and PRAD), MOINER gets 4.86% and 10.24% higher in the F1 metric, respectively, compared with MORONET. In the multiclass classification task (BRCA), MOINER was 4.99% and 5.81% higher than MoGCN in F1_weighted and F1_macro, respectively. The results indicated the superior robustness of MOINER across multiple biomedical classification tasks. It was noteworthy that both methods (MOGONET and MoGCN) were grounded on GCN, which suggested that integration methods utilizing neighborhood aggregation and message passing could more effectively glean insights from multiomics data compared to feedforward

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

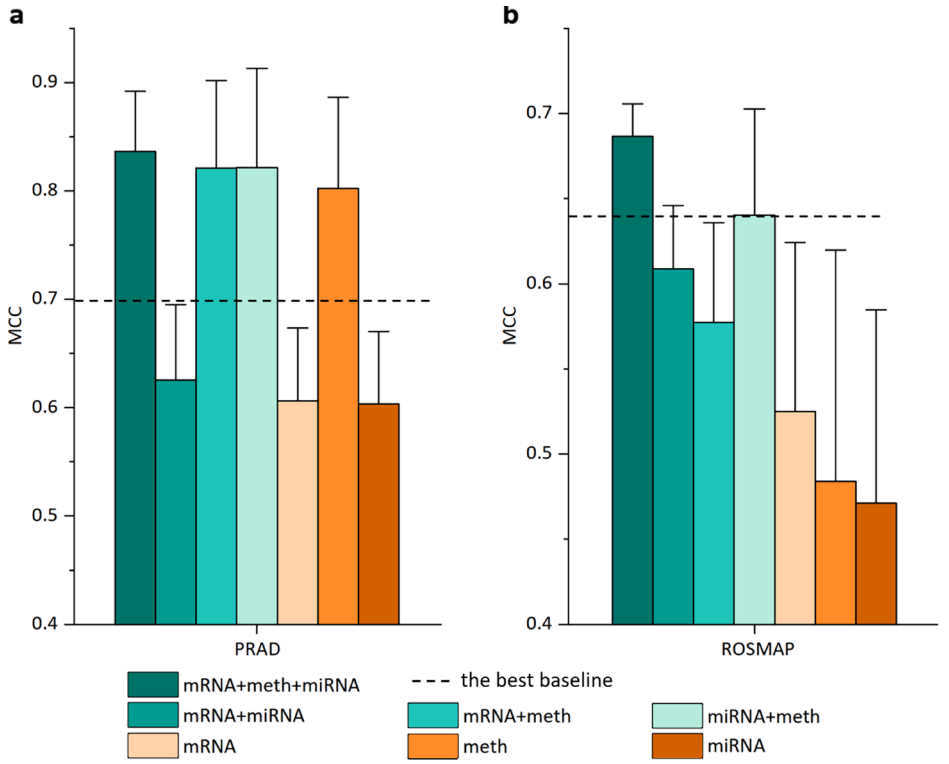


Figure 3. Performance comparison between single-omics and multiomics via MOINER. mRNA, meth, and miRNA refer to single-omics data classification with mRNA expression data, DNA methylation data, and miRNA expression data, respectively. mRNA + meth, mRNA + miRNA, and meth + miRNA refer to classification with two types of omics data. mRNA + meth + miRNA refers to classification with three types of omics data. Box plots show the mean and standard deviation (whiskers). MCC: Matthews correlation coefficient.

neural networks (RDFS and MOMA). Interestingly, the SSN module in MOINER could function as the GCN for neighborhood aggregation and message passing (Figure S2), thus improving the performance of MOINER. MOMA exhibited suboptimal performance compared to other SMI methods on two of three data sets (PRAD and BRCA). This result might be attributed to the fact that MOMA utilizes raw high-dimensional multiomics data as model input, which contains much noise. Compared to the most effective among the four TML methods on three data sets, MOINER gets 19.53%, 27.9% higher in MCC on ROSMAP and PRAD, respectively, and 6.5% higher in F1_macro on BRCA. These TML methods, trained using the concatenated multiomics data set, are associated with the early integration approach, which inherently falls short of fully leveraging the potential of multiomics data. This observation serves to underscore the effectiveness and robustness of our multiomics integration strategy with information enhancement and image representation learning.

The extensibility and Practicability of MOINER. The COVID-19 data set is a binary classification task with four types of multiomics data. As shown in Table S4, MOINER (0.9840), MOGONET (0.9840), and RDFS (0.9600) achieved the comparable performance in ACC on COVID-19 data set. However, it was worth noting that all the four SMI methods were initially developed for dealing with multiomics data containing three or fewer omics types. MOMA and MoGCN, in particular, were difficult to apply to the COVID-19 data set due to their poor extensibility. Therefore, we focused on comparing the proposed MOINER with MOGONET and RDFS on the COVID-19 data set. The source codes of MOGONET and RDFS were manually modified to cope with the COVID-19 data set. In contrast, MOINER stands out as an end-to-end integrated framework. It simplifies the user experience by necessitating only the input of multiomics data without the need for source code modification, and it exhibits flexibility by not imposing restrictions on the number of multiomics data types. In conclusion, MOINER excels in both extensibility and practicality, offering the ability to seamlessly perform multiomics data integration and various classification tasks automatically.

MOINER Demonstrates Superior Performance Compared to Established Unsupervised Multiomics Integration Methods across Diverse Classification Tasks. MOINER was also compared with three unsupervised multiomics integration (UMI) methods: (1) MOFA.[22] MOFA, as a Bayesian model, facilitates the unsupervised integration of multiomics data. It deduces a collection of latent factors aimed at capturing both the biological and technical origins of variability. (2) SNF.[21] SNF is an unsupervised method that creates a comprehensive view of a disease by computing and fusing patient similarity networks. (3) SubtypeGAN.[23] SubtypeGAN employs a deep adversarial framework to integrate multiomics data in an unsupervised manner. To facilitate a comparative analysis of these UMI methods, the combination strategy proposed by Sehwan et al.,[31] denoted as "unsupervised_method + supervised_classifier", was employed in this study. This strategy leveraged "unsupervised_method" for the latent encoding of multiomics data and "supervised_classifier" for subsequent classification. A total of 12 methods were derived and evaluated by pairing the aforementioned three UMI methods with four commonly used classifiers (KNN, RF, SVM, XGBoost). For each UMI method,

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

Table 3. Ablation Study on the BRCA Dataset (5-Fold Cross Validation)[a]

Ablation studies	Model Name	Accuracy	F1_weighted	F1_macro
Study 1	$MOINER_{R2DE+ViT}$	0.8526 ± 0.0178	0.8589 ± 0.0164	0.8306 ± 0.0251
Study 2	$MOINER_{AlexNet}$	0.8354 ± 0.0169	0.8397 ± 0.0158	0.8022 ± 0.2530
	$MOINER_{GoogLeNet}$	0.7897 ± 0.0017	0.7945 ± 0.0018	0.7427 ± 0.0052
	$MOINER_{ResNet}$	0.7703 ± 0.0036	0.7677 ± 0.0039	0.7039 ± 0.0099
	$MOINER_{VGGNet}$	0.8469 ± 0.0045	0.8517 ± 0.0045	0.8225 ± 0.0057
Final model	MOINER	**0.8674 ± 0.0212**	**0.8732 ± 0.019**	**0.8455 ± 0.0271**

[a]The results are presented as mean ± standard deviation. The best result is marked in bold. Study 1: original concatenated matrix of multi-omics data was randomly embedded into 2D space (R2DE) and then transferred to the ViT model for the classification tasks. Study 2: the OmicsMap transformation part in the image representation learning module is retained, and other CNN-based image classifiers (AlexNet, GoogLeNet, ResNetNet, and VGGNet) are used to replace En-ViT and perform classification tasks.

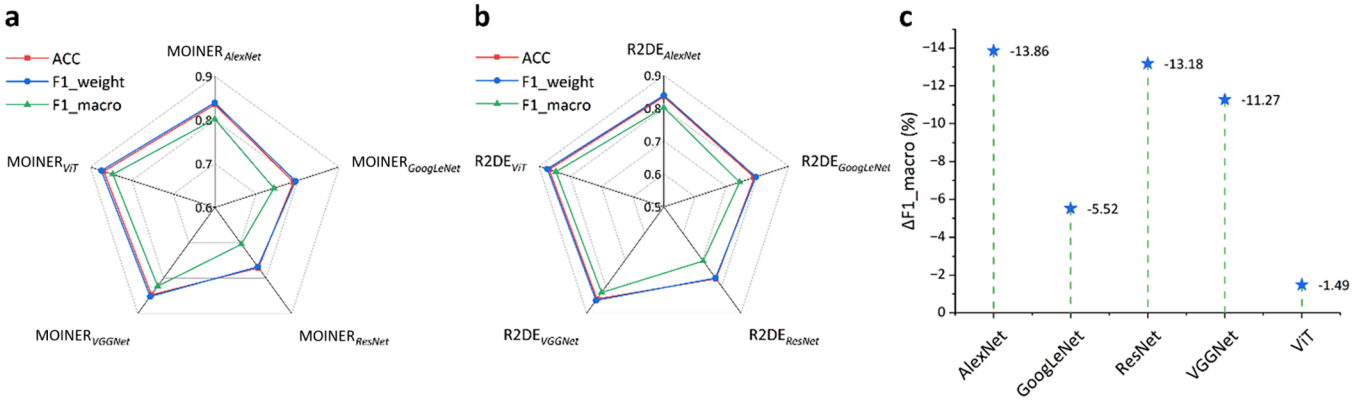


Figure 4. Performance comparison of randomly 2D embedding strategy (R2DE) on different image classifiers. (a) MOINER with different image classifier. (b) MOINER with randomly 2D embedding strategy (R2DE). (c) Performance decrease of image classifier under R2DE strategy.

the optimal combination was selected as the final model, and the results are presented in Supporting Information Tables S5−S7.

As shown in Figure 2, the combination methods with "unsupervised_method + supervised_classifier" strategy were significantly worse than our MOINER. Specifically, compared to the best-performing UMI method, MOINER was able to get 27.55%, 27.56% higher in MCC on ROSMAP and PRAD, respectively, and 11.9% higher in F1_macro on BRCA. Most of the UMI methods were worse than SMI methods, and certain UMI methods displayed inferior performance compared to TML methods (e.g., MOFA worse than SVM on ROSMAP and SNF worse than XGBoost). These results indicate that typical unsupervised integration methods do not work effectively on current biomedical classification tasks, though they are popular in sample clustering and prognostic analysis. This also explains the reason for the emergence of novel supervised multiomics integration approaches.

Performance of MOINER under Different Omics Data Type. In order to demonstrate the effectiveness of multiomics integration in improving the performance of classification task, we performed a comparative analysis to assess the classification effectiveness of MOINER using a combination of three omics data types ("mRNA + meth + miRNA"), MOINER utilizing a combination of two omics data types ("mRNA + meth", "mRNA + miRNA", and "meth + miRNA"), and MOINER using single-omics data type (mRNA, meth, and miRNA). To achieve this goal, the integrated OmicsMap was partitioned into a multichannel map where each channel represented an individual omics layer and was used as the omics-specific map. The maps for three combinations of any two omics types were obtained by pairing different channels within the multichannel map.

As shown in Figure 3, the MOINER models utilizing three types of omics data consistently achieved optimal performance across the two binary classification tasks (ROSMAP and PRAD), which illustrated the necessity of multiomics data integration in biomedical application. Furthermore, the MOINER models utilizing two types of omics data presented superior performance compared to the models employing corresponding single-omics data (e.g., "mRNA + miRNA" outperforms either mRNA or miRNA). Interestingly, certain MOINER models utilizing two omics data types and single-omics data exhibited better performance compared to the best baseline model utilizing three types of omics data (e.g., "mRNA + meth" and "miRNA" in the PRAD data set). This further substantiates that MOINER can effectively capture the intrinsic correlations of omics-features during the early integration of multiomics data through SSN for feature enhancement and FSN for image representation.

Ablation Studies. In the workflow of the MOINER, it represents multiomics data in an image-like format (OmicsMap) using information enhancement and JV algorithms, which are then fed into the ViT model for classification tasks. Therefore, two ablation studies were conducted to systematically investigate the influences of the 2D-embedding strategy (***Study 1***) and classification model (***Study 2***). Specifically, in the ***Study 1***, an original concatenated matrix of multiomics data was randomly embedded into 2D space (R2DE) and then transferred to the En-ViT model for the classification tasks. In the ***Study 2***, the influence of image classifier was comprehensively evaluated. We retained the OmicsMap transformation part and tested four classical CNN-based

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

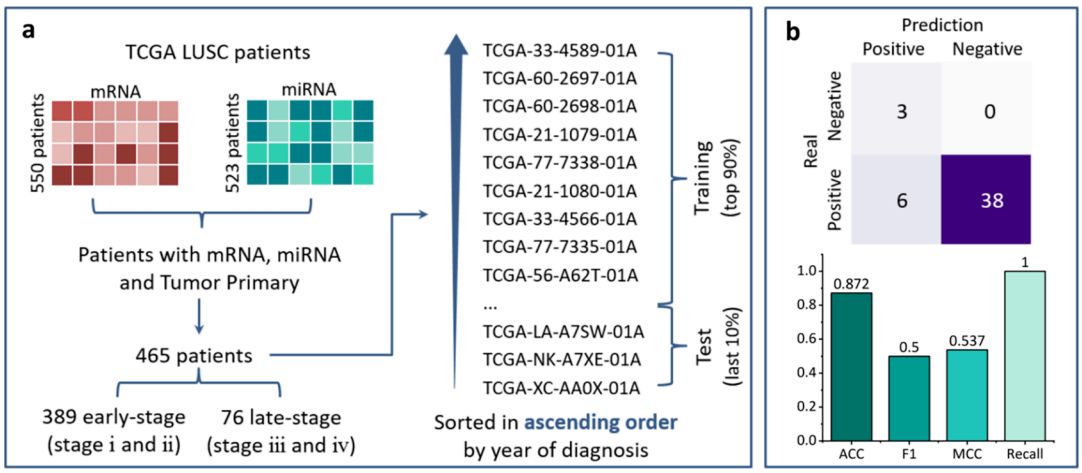


Figure 5. A case study for lung squamous cell carcinoma (LUSC) diagnosis. (a) LUCS data set processing. Patients with primary tumor, stage information, and both types of omics data are included, and they are divided into early (stage i and stage ii) and late (stage iii and stage iv) stages based on tumor stage. These samples are sorted by diagnosis year in ascending order, and the top 90% samples are used as training data. The last 10% samples are used as an independent test data. (b) MOINER prediction result on Test set. ACC: accuracy, MCC: Matthews correlation coefficient.

image classifiers (AlexNet,[49] GoogLeNet,[50] ResNet,[51] and VGGNet[52]). These image classifiers were implemented using the torchvision package. For convenience, the model names for different ablation studies were indicated in Table 3.

As shown in Table 3, removing any module from MOINER or replacing the image classifier resulted in the decreased classification performance on BRCA. Specifically, MOINER outperformed $\text{MOINER}_{\text{R2DE+ViT}}$ by 1.49% in F1_macro. MOINER gets 4.33%, 10.28%, 14.16%, and 2.30% higher in F1_macro compared to $\text{MOINER}_{\text{AlexNet}}$, $\text{MOINER}_{\text{GoogLeNet}}$, $\text{MOINER}_{\text{ResNet}}$, and $\text{MOINER}_{\text{VGGNet}}$, respectively. Furthermore, the influence of R2DE strategy for these CNN-based image classifiers was also evaluated, as shown in Figure 4 and Table S8 & S9; "R2DE + CNN model" under randomly 2D embedding strategy decreases by 5.5% to 13.9% in F1_macro. The results indicate that the ViT model based on the attention mechanism has a natural advantage in multimodal data integration compared to other CNN models. These results also indicate that the combination of all the proposed modules collectively contributes to the overall superiority of MOINER and effectively compensates for the shortcomings of simply early integration methods in multiomics data.

A Case Study for Lung Squamous Cell Carcinoma (LUSC) Diagnosis. The application prospects of MOINER in disease diagnosis was validated by using the LUSC data set. To be specific, a multiomics data set of mRNA and miRNA for LUSC was obtained from GDC TCGA. As shown in Figure 5a, patients with primary tumor, stage information, and both types of omics data were included, and they were divided into early (stage i and stage ii) and late (stage iii and stage iv) stages based on tumor stage. In total, 465 samples (389 early stage and 76 late-stage) were obtained. These samples were sorted by diagnosis year in ascending order, and the top 90% samples were used as training data for 5-fold CV, which included 345 early stage and 73 late-stage patients. The last 10% samples were used as an independent test data, which included 44 early stage and 3 late-stage patients. After model training on 5-fold CV, the best model was then evaluated on the independent test set. As shown in Figure 5b, MOINER achieved an ACC of 0.872, F1-score of 0.5, and MCC of 0.537 on the independent test set, and all three positive samples were well-identified. It was worth noting that the LUSC data set is highly imbalanced with a large discrepancy between positive and negative patients. We mainly focued on the recall metric, which was the proportion of positive samples correctly predicted by the model. The recall produced by MOINER was 1.0, demonstrating the power of MOINER in identifying the ground-truth positive patients in clinical practice.

Investigating the Interpretability of MOINER. To visualize the latent representation of multiomics samples, the attention embedding of class token was extracted from MOINER and the clustering performance of attention embedding was compared to that of raw multiomics data. As shown in Figure 6, the MOINER embedding was more distinguishable for sample clustering than raw data and achieved better ARI scores,[45] indicating the power of MOINER in multiomics data analysis. Furthermore, a main advantage of MOINER was its ability in giving crucial feature-level insights and interpretation into potential biomarker discovery. The capability of MOINER for potential biomarker discovery was evaluated on ROSMAP data set. Important biomarkers were identified based on their importance score (described in Materials and Methods). Figure 7 depicted the top 15 features identified by MOINER from each CV. The ranking of features was determined through a comprehensive analysis of 5-fold CV, where features identified more frequently across the folds received higher rankings.

Based on the comprehensive consideration of 5-fold CV results, MOINER identified several crucial mRNA features, including *APLN*, *ANKRD30B*, *SLC25A18*, *GPER1*, and *CDK2AP1* et al. *APLN* encodes Apelin, a bioactive neuropeptide[53] widely distributed within neuronal cell bodies and fibers across the neuraxis.[54] Numerous investigations suggest that apelin may exert a crucial influence on the pathophysiology of AD by regulating Tau and amyloid-β,[55−57] and it has been suggested as a potential focus for neurodegenerative diseases beyond AD.[55,58] Additionally, Semick et al. initially indicated a significant downregulation of ANKRD30B in AD patients when compared to the control group in entorhinal cortex brain regions and hippocampus, suggesting that it is a promising AD-related gene.[59] Other genes identified by MOINER, such as *GPER1*[60] and *CDK2AP1*,[61] had also been

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

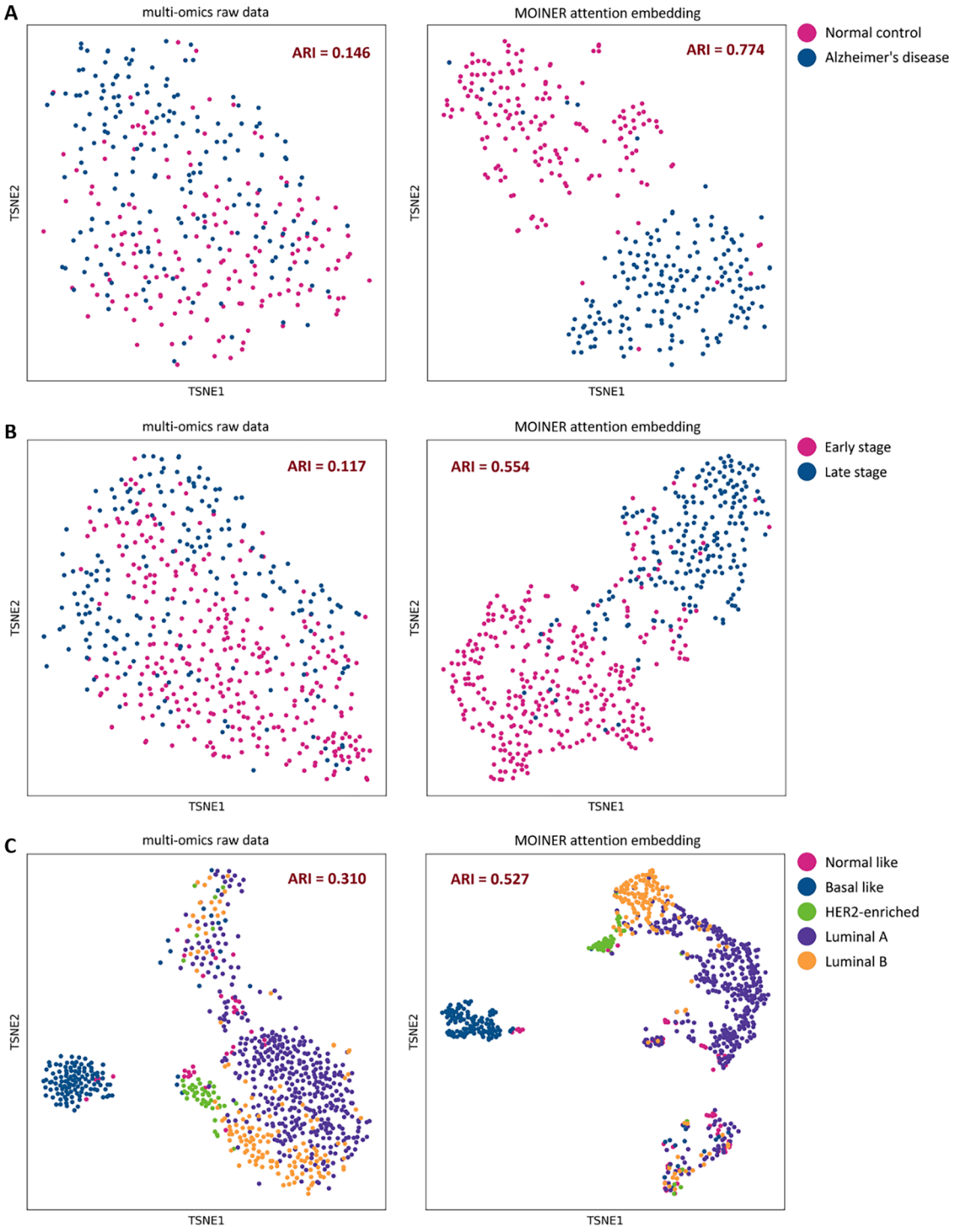


Figure 6. TSNE visualization of patients based on the MOINER attention embedding (right) and the initial raw expression (left). (a) Visualization of the ROSMAP data set (Alzheimer's disease and normal control). (b) Visualization of the PRAD data set (early stage and late stage). (c) Visualization of the BRCA data set (normal-like, basal-like, human epidermal growth factor receptor 2 (HER2)-enriched, Luminal A, and Luminal B subtypes). The adjusted Rand index (ARI) score is calculated and shown in the plot.

proved to be associated with AD. Moreover, highly ranking miRNAs identified by MOINER, such as *has-mir-129-5p*,[62] *has-mir-132*,[63,64] *has-mir-376a*,[65] and *has-mir-127−3p*[66] et al., had also been reported to be associated with AD.[67] For instance, there is an association between the expression of *miR-129-5p* in serum and the levels of cognitive function markers in AD patients.[62] The validation of important features identified by MOINER against existing experimental literature underscores the promising applications of MOINER in the discovery of potential biomarkers for disease diagnosis in clinical practice.

CONCLUSION

In this study, a novel multiomics early integration framework (MOINER) was constructed by (1) information enhancement and (2) image representation learning for biomedical

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

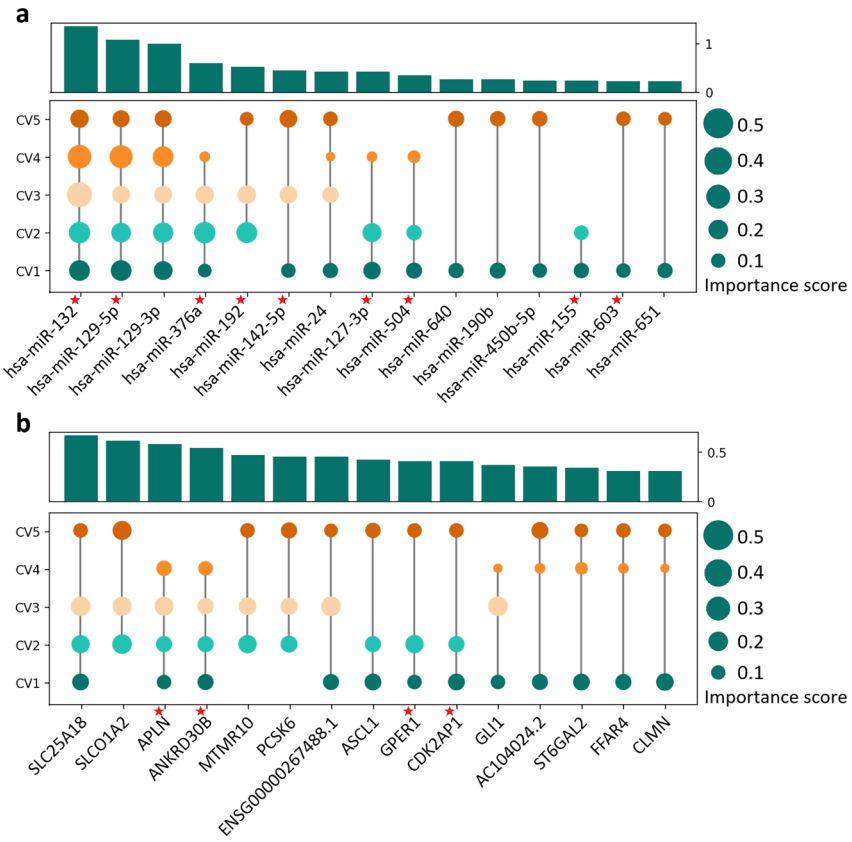


Figure 7. Important input features identified by MOINER on the ROSMAP data set. (a) miRNA level. (b) mRNA level. The circle represents whether this feature is identified in this fold. The height of the bar represents the sum of the scores for this feature in 5-CV, while the size of the circle represents the importance score of the feature in a certain fold. The red pentagram in the upper right corner of the feature represents that this feature has been reported in the literature.

classification and biomarker discovery. Based on a comprehensive comparison with SOTA multiomics integration methods and traditional machine learning models, our proposed method consistently achieves superior performance and holds good interpretability. The effectiveness of each key module in MOINER is demonstrated by systematic ablation studies. All in all, this work enables better use of multiomics data and would become an essential tool for omics research, disease diagnosis, and biomarker discovery.

■ ASSOCIATED CONTENT

Data Availability Statement

ROSMAP and BRCA are derived from MOGONET (https://github.com/txWang/MOGONET). PRAD is sourced from the GDC TCGA Prostate Cancer on Xena (https://xenabrowser.net/). COVID-19 data set is retrieved from the MassIVE Data set Summary (accession = MSV000085703). The source code of this study can be found in GitHub: https://github.com/idrblab/MOINER.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00013.

The comparative performance of multiomics integration methods on benchmark data sets using 5-fold cross-validation (Tables S1−S4), classifier testing of unsupervised multiomics integration methods (Tables S5−S7), ablation studies of image classifier and R2DE strategy in the MOINER (Tables S8 and S9), ViT model structure (Figure S1), information enhancement module and image representation module in the MOINER (Figures S2 and S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Feng Zhu − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China; Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China;* orcid.org/0000-0001-8069-0053; Email: zhufeng@zju.edu.cn

Jianqing Gao − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0003-1052-7060; Email: gaojianqing@zju.edu.cn

Authors

Wei Zhang − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China; Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China;* orcid.org/0009-0007-8335-8663

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732

Minjie Mou − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0001-7619-2975
Wei Hu − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China*
Mingkun Lu − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0003-1522-6320
Hanyu Zhang − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China*
Hongning Zhang − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0002-7818-7915
Yongchao Luo − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0002-4140-5392
Hongquan Xu − *Key Laboratory of Elemene Class Anti-Cancer Chinese Medicines, School of Pharmacy, Hangzhou Normal University, Hangzhou 311121, China*
Lin Tao − *Key Laboratory of Elemene Class Anti-Cancer Chinese Medicines, School of Pharmacy, Hangzhou Normal University, Hangzhou 311121, China;* orcid.org/0000-0001-7494-4758
Haibin Dai − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00013

Author Contributions

Wei Zhang: Investigation, Methodology, Software, Writing - Original Draft, Writing - Review & Editing. Minjie Mou: Writing - Original Draft. Wei Hu: Investigation, Validation. Mingkun Lu: Formal analysis, Validation. Hanyu Zhang: Resources, Data Curation. Hongning Zhang: Investigation. Hongquan Xu: Investigation. Yongchao Luo: Visualization. Lin Tao: Software. Haibin Dai: Writing - Review & Editing. Jianqing Gao: Supervision, Writing - Review & Editing. Feng Zhu: Conceptualization, Supervision, Writing - Review & Editing, Project administration.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by Natural Science Foundation of Zhejiang Province (LR21H300001); National Key R&D Program of China Synthetic Biology Research (2019YFA0905900); National Natural Science Foundation of China (22220102001, U1909208, 81872798); Scientific Research Grant of Ningbo University (215-432000282); Ningbo Top Talent Project (215-432094250); National Key R&D Program of China (2022YFC3400501); Leading Talent of "Ten Thousand Plan" National High-Level Talents Special Support Plan of China; Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine); Fundamental Research Fund for Central Universities (2018QNA7023); "Double Top-Class" University Project (181201*194232101); Key R&D Program of Zhejiang Province (2020C03010); Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud; The information technology center of Zhejiang University. Funds for open access charge: Natural Science Foundation of Zhejiang Province (LR21H300001).

■ REFERENCES

(1) Li, F.; Yin, J.; Lu, M.; Mou, M.; Li, Z.; Zeng, Z.; Tan, Y.; Wang, S.; Chu, X.; Dai, H.; Hou, T.; Zeng, S.; Chen, Y.; Zhu, F. DrugMAP: molecular atlas and pharma-information of all drugs. *Nucleic acids research* **2023**, *51*, D1288−D1299.
(2) Fu, J.; Yang, Q.; Luo, Y.; Zhang, S.; Tang, J.; Zhang, Y.; Zhang, H.; Xu, H.; Zhu, F. Label-free proteome quantification and evaluation. *Briefings in bioinformatics* **2023**, *24*, bbac477.
(3) Yang, Q.; Li, B.; Wang, P.; Xie, J.; Feng, Y.; Liu, Z.; Zhu, F. LargeMetabo: an out-of-the-box tool for processing and analyzing large-scale metabolomic data. *Briefings in bioinformatics* **2022**, *23*, bbac455.
(4) Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113−20.
(5) Alexandrov, L. B.; Kim, J.; Haradhvala, N. J.; Huang, M. N.; Tian Ng, A. W.; Wu, Y.; Boot, A.; Covington, K. R.; Gordenin, D. A.; Bergstrom, E. N.; Islam, S. M. A.; Lopez-Bigas, N.; Klimczak, L. J.; McPherson, J. R.; Morganella, S.; Sabarinathan, R.; Wheeler, D. A.; Mustonen, V.; Getz, G.; Rozen, S. G.; Stratton, M. R.; et al. The repertoire of mutational signatures in human cancer. *Nature* **2020**, *578*, 94−101.
(6) Picard, M.; Scott-Boyer, M. P.; Bodein, A.; Périn, O.; Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct Biotechnol J.* **2021**, *19*, 3735−3746.
(7) Nicora, G.; Vitali, F.; Dagliati, A.; Geifman, N.; Bellazzi, R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Front. Oncol.* **2020**, *10*, 1030.
(8) Arjmand, B.; Hamidpour, S. K.; Tayanloo-Beik, A.; Goodarzi, P.; Aghayan, H. R.; Adibi, H.; Larijani, B. Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer. *Front Genet* **2022**, *13*, 824451.
(9) Sammut, S. J.; Crispin-Ortuzar, M.; Chin, S. F.; Provenzano, E.; Bardwell, H. A.; Ma, W.; Cope, W.; Dariush, A.; Dawson, S. J.; Abraham, J. E.; Dunn, J.; Hiller, L.; Thomas, J.; Cameron, D. A.; Bartlett, J. M. S.; Hayward, L.; Pharoah, P. D.; Markowetz, F.; Rueda, O. M.; Earl, H. M.; Caldas, C. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **2022**, *601*, 623−629.
(10) Yang, J.; Chen, Y.; Jing, Y.; Green, M. R.; Han, L. Advancing CAR T cell therapy through the use of multidimensional omics data. *Nat. Rev. Clin Oncol* **2023**, *20*, 211−228.
(11) Sandhu, C.; Qureshi, A.; Emili, A. Panomics for Precision Medicine. *Trends Mol. Med.* **2018**, *24*, 85−101.
(12) Friedman, A. A.; Letai, A.; Fisher, D. E.; Flaherty, K. T. Precision medicine for cancer with next-generation functional diagnostics. *Nat. Rev. Cancer* **2015**, *15*, 747−56.
(13) Biswas, N.; Chakrabarti, S. Artificial Intelligence (AI)-Based Systems Biology Approaches in Multi-Omics Data Analysis of Cancer. *Front Oncol* **2020**, *10*, 588221.
(14) Misra, B. B.; Langefeld, C. D.; Olivier, M.; Cox, L. A. Integrated Omics: Tools, Advances, and Future Approaches. *J. Mol. Endocrinol* **2018**, *62*, JME-18-0055.
(15) Mirza, B.; Wang, W.; Wang, J.; Choi, H.; Chung, N. C.; Ping, P. Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes (Basel)* **2019**, *10*, 87.
(16) Kim, M.; Tagkopoulos, I. Data integration and predictive modeling methods for multi-omics datasets. *Mol. Omics* **2018**, *14*, 8−25.

(17) Momeni, Z.; Hassanzadeh, E.; Saniee Abadeh, M.; Bellazzi, R. A survey on single and multi omics data mining methods in cancer data classification. *J. Biomed Inform* **2020**, *107*, 103466.
(18) Zhan, Y.; Liu, J.; Ou-Yang, L. scMIC: A Deep Multi-level Information Fusion Framework for Clustering Single-cell Multi-omics Data. *IEEE J. Biomed. Health Inform.* **2023**, *27* (12), 6121−6132.
(19) Stanojevic, S.; Li, Y.; Ristivojevic, A.; Garmire, L. X. Computational Methods for Single-cell Multi-omics Integration and Alignment. *Genomics Proteomics Bioinformatics* **2022**, *20*, 836−849.
(20) Shen, R.; Olshen, A. B.; Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **2009**, *25*, 2906−12.
(21) Wang, B.; Mezlini, A. M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **2014**, *11*, 333−7.
(22) Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J. C.; Buettner, F.; Huber, W.; Stegle, O. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **2018**, *14*, No. e8124.
(23) Yang, H.; Chen, R.; Li, D.; Wang, Z. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics* **2021**, *37*, 2231−2237.
(24) Poirion, O. B.; Jing, Z.; Chaudhary, K.; Huang, S.; Garmire, L. X. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* **2021**, *13*, 112.
(25) Wang, T.; Shao, W.; Huang, Z.; Tang, H.; Zhang, J.; Ding, Z.; Huang, K. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* **2021**, *12*, 3445.
(26) Singh, A.; Shannon, C. P; Gautier, B.; Rohart, F.; Vacher, M.; Tebbutt, S. J; Le Cao, K.-A. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **2019**, *35*, 3055−3062.
(27) Hu, Y.; Zhao, L.; Li, Z.; Dong, X.; Xu, T.; Zhao, Y. Classifying the multi-omics data of gastric cancer using a deep feature selection method. *Expert Systems with Applications* **2022**, *200*, 116813.
(28) Fraser, M.; Rouette, A. Prostate Cancer Genomic Subtypes. *Adv. Exp. Med. Biol.* **2019**, *1210*, 87−110.
(29) Fu, Y.; Xu, J.; Tang, Z.; Wang, L.; Yin, D.; Fan, Y.; Zhang, D.; Deng, F.; Zhang, Y.; Zhang, H.; Wang, H.; Xing, W.; Yin, L.; Zhu, S.; Zhu, M.; Yu, M.; Li, X.; Liu, X.; Yuan, X.; Zhao, S. A gene prioritization method based on a swine multi-omics knowledgebase and a deep learning model. *Commun. Biol.* **2020**, *3*, 502.
(30) Li, X.; Ma, J.; Leng, L.; Han, M.; Li, M.; He, F.; Zhu, Y. MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. *Front Genet* **2022**, *13*, 806842.
(31) Moon, S.; Lee, H. MOMA: a multi-task attention learning algorithm for multi-omics data interpretation and classification. *Bioinformatics* **2022**, *38*, 2287−2296.
(32) Zitnik, M.; Nguyen, F.; Wang, B.; Leskovec, J.; Goldenberg, A.; Hoffman, M. M. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *Inf Fusion* **2019**, *50*, 71−91.
(33) Zhang, Y.; Sun, H.; Lian, X.; Tang, J.; Zhu, F. ANPELA: significantly enhanced quantification tool for cytometry-based single-cell proteomics. *Advanced science* **2023**, *10*, No. e2207061.
(34) Yoosuf, N.; Maciejewski, M.; Ziemek, D.; Jelinsky, S. A.; Folkersen, L.; Müller, M.; Sahlström, P.; Vivar, N.; Catrina, A.; Berg, L.; Klareskog, L.; Padyukov, L.; Brynedal, B. Early prediction of clinical response to anti-TNF treatment using multi-omics and machine learning in rheumatoid arthritis. *Rheumatology (Oxford)* **2022**, *61*, 1680−1689.
(35) Shen, W. X.; Zeng, X.; Zhu, F.; Wang, Y. l.; Qin, C.; Tan, Y.; Jiang, Y. Y.; Chen, Y. Z. Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nature Machine Intelligence* **2021**, *3*, 334−343.
(36) Gokhale, M.; Mohanty, S. K.; Ojha, A. GeneViT: Gene Vision Transformer with Improved DeepInsight for cancer classification. *Comput. Biol. Med.* **2023**, *155*, 106643.
(37) Mou, M.; Pan, Z.; Zhou, Z.; Zheng, L.; Zhang, H.; Shi, S.; Li, F.; Sun, X.; Zhu, F. A Transformer-Based Ensemble Framework for the Prediction of Protein-Protein Interaction Sites. *Research (Wash D C)* **2023**, *6*, 0240.
(38) Overmyer, K. A.; Shishkova, E.; Miller, I. J.; Balnis, J.; Bernstein, M. N.; Peters-Clarke, T. M.; Meyer, J. G.; Quan, Q.; Muehlbauer, L. K.; Trujillo, E. A.; He, Y.; Chopra, A.; Chieng, H. C.; Tiwari, A.; Judson, M. A.; Paulson, B.; Brademan, D. R.; Zhu, Y.; Serrano, L. R.; Linke, V.; Drake, L. A.; Adam, A. P.; Schwartz, B. S.; Singer, H. A.; Swanson, S.; Mosher, D. F.; Stewart, R.; Coon, J. J.; Jaitovich, A. Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Syst* **2021**, *12*, 23−40 e7.
(39) Fu, J.; Zhang, Y.; Wang, Y.; Zhang, H.; Liu, J.; Tang, J.; Yang, Q.; Sun, H.; Qiu, W.; Ma, Y.; Li, Z.; Zheng, M.; Zhu, F. Optimization of metabolomic data processing using NOREVA. *Nature protocols* **2022**, *17*, 129−151.
(40) Fu, J.; Zhang, Y.; Liu, J.; Lian, X.; Tang, J.; Zhu, F. Pharmacometabonomics: data processing and statistical analysis. *Briefings in bioinformatics* **2021**, *22*, bbab138.
(41) Xu, J.; Zhou, D.; Deng, D.; Li, J.; Chen, C.; Liao, X.; Chen, G.; Heng, P. A. Deep Learning in Cell Image Analysis. *Intelligent Computing* **2022**, *2022*. DOI: 10.34133/2022/9861263
(42) Liu, X.; Gao, K.; Liu, B.; Pan, C.; Liang, K.; Yan, L.; Ma, J.; He, F.; Zhang, S.; Pan, S.; Yu, Y. Advances in Deep Learning-Based Medical Image Analysis. *Health Data Science* **2021**, *2021*, 8786793.
(43) Sharir, G.; Noy, A.; Zelnik-Manor, L. An image is worth 16 × 16 words, what is a video worth? *arXiv* 2021. DOI: 10.48550/arXiv.2103.13915
(44) Zhang, S. W.; Xu, J. Y.; Zhang, T. DGMP: Identifying Cancer Driver Genes by Jointing DGCN and MLP from Multi-omics Genomic Data. *Genomics Proteomics Bioinformatics* **2022**, *20*, 928−938.
(45) Yang, F.; Wang, W.; Wang, F.; Fang, Y.; Tang, D.; Huang, J.; Lu, H.; Yao, J. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **2022**, *4*, 852−866.
(46) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C. A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **2019**, *37*, 38−44.
(47) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579−2605.
(48) Jonker, R.; Volgenant, A. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* **1987**, *38*, 325−340.
(49) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems;* Curran Associates, Inc., 2012, *25*.
(50) Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015;* **2015**, 1−9.
(51) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016;* **2016**, 770−778.
(52) Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014. DOI: 10.48550/arXiv.1409.1556
(53) Lee, D. K.; Cheng, R.; Nguyen, T.; Fan, T.; Kariyawasam, A. P.; Liu, Y.; Osmond, D. H.; George, S. R.; O'Dowd, B. F. Characterization of apelin, the ligand for the APJ receptor. *J. Neurochem* **2000**, *74*, 34−41.

(54) Cheng, B.; Chen, J.; Bai, B.; Xin, Q. Neuroprotection of apelin and its signaling pathway. *Peptides* **2012**, *37*, 171−3.
(55) Luo, H.; Han, L.; Xu, J. Apelin/APJ system: A novel promising target for neurodegenerative diseases. *J. Cell Physiol* **2020**, *235*, 638−657.
(56) Luo, H.; Xiang, Y.; Qu, X.; Liu, H.; Liu, C.; Li, G.; Han, L.; Qin, X. Apelin-13 Suppresses Neuroinflammation Against Cognitive Deficit in a Streptozotocin-Induced Rat Model of Alzheimer's Disease Through Activation of BDNF-TrkB Signaling Pathway. *Front Pharmacol* **2019**, *10*, 395.
(57) Dragomir, A.; Vrahatis, A. G.; Bezerianos, A. A network-based perspective in Alzheimer's disease: Current state and an integrative framework. *IEEE journal of biomedical and health informatics* **2019**, *23*, 14−25.
(58) Masoumi, J.; Abbasloui, M.; Parvan, R.; Mohammadnejad, D.; Pavon-Djavid, G.; Barzegari, A.; Abdolalizadeh, J. Apelin, a promising target for Alzheimer disease prevention and treatment. *Neuropeptides* **2018**, *70*, 76−86.
(59) Semick, S. A.; Bharadwaj, R. A.; Collado-Torres, L.; Tao, R.; Shin, J. H.; Deep-Soboslay, A.; Weiss, J. R.; Weinberger, D. R.; Hyde, T. M.; Kleinman, J. E.; Jaffe, A. E.; Mattay, V. S. Integrated DNA methylation and gene expression profiling across multiple brain regions implicate novel genes in Alzheimer's disease. *Acta Neuropathol* **2019**, *137*, 557−569.
(60) Neuner, S. M.; Heuer, S. E.; Zhang, J. G.; Philip, V. M.; Kaczorowski, C. C. Identification of Pre-symptomatic Gene Signatures That Predict Resilience to Cognitive Decline in the Genetically Diverse AD-BXD Model. *Front Genet* **2019**, *10*, 35.
(61) Li, Q. S.; De Muynck, L. Differentially expressed genes in Alzheimer's disease highlighting the roles of microglia genes including OLR1 and astrocyte gene CDK2AP1. *Brain Behav Immun Health* **2021**, *13*, 100227.
(62) Li, Z.; Chen, Q.; Liu, J.; Du, Y. Physical Exercise Ameliorates the Cognitive Function and Attenuates the Neuroinflammation of Alzheimer's Disease via miR-129−5p. *Dement Geriatr Cogn Disord* **2020**, *49*, 163−169.
(63) Qian, Y.; Song, J.; Ouyang, Y.; Han, Q.; Chen, W.; Zhao, X.; Xie, Y.; Chen, Y.; Yuan, W.; Fan, C. Advances in Roles of miR-132 in the Nervous System. *Front Pharmacol* **2017**, *8*, 770.
(64) Cong, L.; Cong, Y.; Feng, N.; Liang, W.; Wu, Y. Up-regulated microRNA-132 reduces the cognition-damaging effect of sevoflurane on Alzheimer's disease rats by inhibiting FOXA1. *Genomics* **2021**, *113*, 3644−3652.
(65) Mun, S. K.; Chae, H.; Piao, X. Y.; Lee, H. J.; Kim, Y. K.; Oh, S. H.; Chang, M. MicroRNAs Related to Cognitive Impairment After Hearing Loss. *Clin Exp Otorhinolaryngol* **2021**, *14*, 76−81.
(66) Piscopo, P.; Grasso, M.; Puopolo, M.; D'Acunto, E.; Talarico, G.; Crestini, A.; Gasparini, M.; Campopiano, R.; Gambardella, S.; Castellano, A. E.; Bruno, G.; Denti, M. A.; Confaloni, A. Circulating miR-127−3p as a Potential Biomarker for Differential Diagnosis in Frontotemporal Dementia. *J. Alzheimers Dis* **2018**, *65*, 455−464.
(67) Zhang, H.; Wang, Y.; Pan, Z.; Sun, X.; Mou, M.; Zhang, B.; Li, Z.; Li, H.; Zhu, F. ncRNAInter: a novel strategy based on graph neural network to discover interactions between lncRNA and miRNA. *Briefings in bioinformatics* **2022**, *23*, bbac411.

https://doi.org/10.1021/acs.jcim.4c00013
J. Chem. Inf. Model. 2024, 64, 2720−2732