Supplementary Material

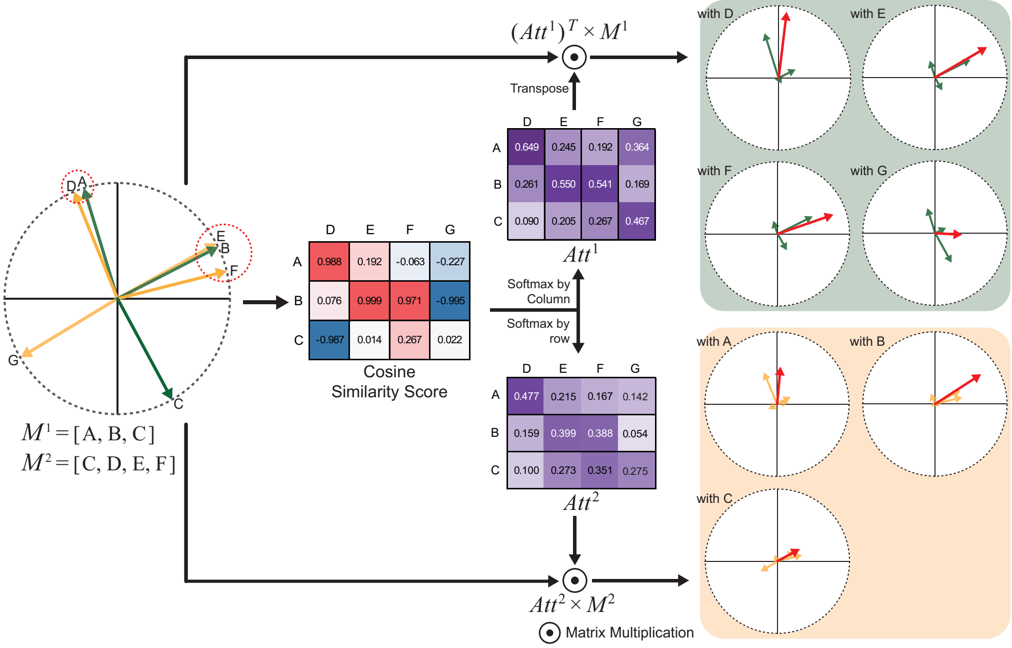
MOMA: A Multi-task Attention Learning Algorithm for Multi-omics Data Interpretation and Classification

Sehwan Moon, Hyunju Lee

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea.

1. A toy example for the module attention mechanism

An example of module attention mechanism is illustrated in the toy example shown in Figure S1, demonstrates how attention mechanism works and why it improves on learning. In the left two-dimension Cartesian coordinate system, green vectors A, B, and C are gene expression module vectors, and yellow vectors D, E, F, and G are DNA methylation module vectors. In this example, gene expression module A and DNA methylation module D, and gene expression module B and DNA methylation modules E and F have high similarities. The similarity score represents cosine similarity between the gene expression and DNA methylation modules. Each gene expression module vector was computed as the weighted sum according to the similarity with each DNA methylation module. Also, the weight sum of each DNA methylation module vector was calculated according to the similarity to the gene expression module. The red vectors in the two-dimensional coordinate system on the green background represent the modified gene expression module vectors, and the red vectors on the orange background represent the modified DNA methylation vectors. Through the module attention mechanism, we could pay attention to modules with high similarity between the two data modules.



	D	E	F	G
A	0.988	0.192	-0.063	-0.227
B	0.076	0.999	0.971	-0.995
C	-0.987	0.014	0.267	0.022

Cosine Similarity Score

	D	E	F	G
A	0.649	0.245	0.192	0.364
B	0.261	0.550	0.541	0.169
C	0.090	0.205	0.267	0.467

Att^1

	D	E	F	G
A	0.477	0.215	0.167	0.142
B	0.159	0.399	0.388	0.054
C	0.100	0.273	0.351	0.275

Att^2

Fig. S1. Toy example to illustrate the module attention. In a two-dimensional Cartesian coordinate system, green vectors A, B, and C are gene expression module vectors, yellow vectors D, E, F, and G are DNA methylation vectors. In this case, the similarity is high between vectors A and D and between vectors E, B and F, respectively. After calculating cosine similarity scores between the gene expression and DNA methylation modules, attention is given to the modules with the high similarity.

2. Our Proposed Multi-task Attention Learning Algorithm for gene expression and DNA methylation data

In the main script, we applied MOMA with gene expression and DNA methylation data.

A. MOMA algorithm on two different datasets

Given a training sample $\{x^1, x^2, y\}$, where x^1 and x^2 denote the sample under the gene expression and DNA methylation profile, and y is the corresponding label. Let f_{module} denotes the module encoder. The module vector of each omics data is defined as follows:

$$M^1(x^1) = f^1_{module}\left(X^1;\theta^1_{module}\right) \in \mathbb{R}^{N^1\times D},$$
$$M^2(x^2) = f^2_{module}\left(X^2;\theta^2_{module}\right) \in \mathbb{R}^{N^2\times D},$$

where f_{module} consists of the fully connected layer and unit vector normalization, θ_{module} denotes the weights of f_{module}, D is the dimension of the module vector, N^1 and N^2 are the number of gene expression modules and number of DNA methylation modules, respectively, and M^1 and M^2 are the module vectors of gene expression and DNA methylation, respectively.

We devised a module attention mechanism that focuses on modules with high similarity between the gene expression module and the DNA methylation module. We used cosine similarity to measure relevance. We defined the attention matrix Att^1 and Att^2 according to the softmax axis as follows:

$$Att^1_{lk}(x) = \frac{exp\left(\cos\left(M^1_l, M^2_k\right)\right)}{\sum_{k=1}^{N^2} exp\left(\cos\left(M^1_l, M^2_k\right)\right)}, \quad Att^2_{lk}(x) = \frac{exp\left(\cos\left(M^2_l, M^1_k\right)\right)}{\sum_{k=1}^{N^1} exp\left(\cos\left(M^2_l, M^1_k\right)\right)},$$

where $M = M(x)$ for short, M_l denotes l-th module of the module vector. Each element of Att_{lk} stores the relation information with possible dependence between the l-th module from one omics data and the k-th module from another omics data module.

To highlight the important modules, the module vectors are multiplied by the attention matrix. The fully connected layers are then applied, which flattens the multi-dimensional vectors and yields the final probabilities for each label. Loss L is set to the cross-entropy error between the gold label and task-specific outputs:

$$L = -\sum_{c=1}^{C}\left(y_c \cdot log\left(f^1_{fc}\left(\left(Att^1(x)\right)^T M^1;\theta^1_{fc}\right)\right) + y_c \cdot log\left(f^2_{fc}\left(\left(Att^2(x)\right)^T M^2;\theta^2_{fc}\right)\right)\right)$$
$$+\ \lambda \sum W^2, \quad .s.t.W \in \{\theta_{module}, \theta_{fc}\},$$

where $M = M(x)$ for short, C represents the total number of classes, y_c denotes a labeling of c, f^1_{fc} and f^2_{fc} consist of multiple fully connected layers, θ_{fc} denotes the weights of f_{fc}, and W denotes the weight. A L2-norm penalty was used to the optimization to avoid overfitting of the module encoders and the multiple connected layers.

B. Hyperparameters list for nested lidation

We proceeded to performance measurements under the 5-fold outer cross-validation (CV) (Figure S2). Given data were split into training and test data with a 4:1 ratio. In the training data, the optimal hyperparameters of the model were determined using grid search by inner 3-fold CV in training data. Table S1 and Table S2 show the grid search results for MOMA on ROSMAP datasets.

Hyperparameters are as follows for each model. For XGBoost, the parameters of 'the max depth' from the set {3, 5, 7, 9}, 'the regularization lambda' from the set {100, 10, 1, 0.1, 0.01}, and 'the learning rate' from the set {0.3, 0.2, 0.1, 0.01} were optimized. For DNN, the parameters of 'the number of layers' from the set {3, 5, 7}, 'the learning rate' from the set $\{5\times10^{-5}, 5\times10^{-6}, 5\times10^{-7}\}$, 'the weight decay' from the set $\{10^{-2}, 10^{-4}, 10^{-6}, 0\}$, and 'the early stopping patience' from the set {10, 30, 50, 100} were optimized. All hidden layers are equipped with ReLU activation and the final layer is with sigmoid or softmax functions. For MORONET[1], the parameters of 'the threshold of affinity values' from the set {2, 4, 6, 8, 10}, 'the learning rate for pretraining' from

[1] Wang, Tongxin, et al. "MORONET: Multi-omics Integration via Graph Convolutional Networks for Biomedical Data Classification." bioRxiv (2020).

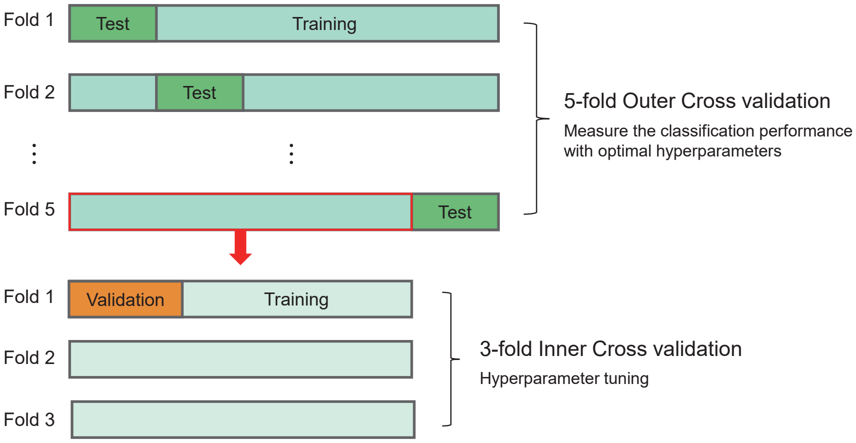

Fold 1
Test
Training
Fold 2
Test
Fold 5
Test
5-fold Outer Cross validation
Measure the classification performance with optimal hyperparameters
Fold 1
Validation
Training
Fold 2
Fold 3
3-fold Inner Cross validation
Hyperparameter tuning

Fig. S2. Illustration of the nested cross-validation used in this study.

the set $\{5 \times 10^{-3}, 5 \times 10^{-4}\}$, 'the learning rate for graph convolutional network' from the set $\{5 \times 10^{-3}, 5 \times 10^{-4}\}$, 'the learning rate for classifier' from the set $\{5 \times 10^{-3}, 5 \times 10^{-4}\}$, and 'the number of significant features for each omics data type' from the set {200, 400} were used. MOFA[2] was used for integrating multi-omics data, and the performance was measured with support vector machine (SVM) based on the extracted factors. The parameters of 'the number of factors' from the set {32, 64, 128, 256} (with less than the sample size) was optimized. For SVM, the parameters of 'Regularization parameter' from set $\{0.001, 0.01, 0.1, 1, 10\}$, and the parameters of 'kernel' from the set {linear and radial basis function} were used. For SMSPL[3], the parameters of 'the parameter for adjusting influence from other modalities' from the set {0.66,0.1,0.01}, 'the age parameter' from the set $\{(0.66, 0.66), (0.1, 0.1), (0.01, 0.01)\}$, 'the size of increasing the age parameter with each iteration' from the set $\{0.01, 0.02, 0.04, 0.08\}$, and 'The size to increase the selected sample for each iteration' from the set {2, 4} were optimized. In TCGA 34 class classification, due to the enormous computation cost ($5 \times 3 \times 72 \times 3$ days) of adopting the inner CV strategy, the parameters were directly adjusted. For P-NET[4], the network structure made up of biological entities was used, and 'the drop rate', 'the drop interval', and 'the epoch' required for learning were tuned from the set $\{(0.1, 100, 300), (0.1, 200, 500), (0.25, 100, 500), (0.25, 200, 1000), (0.5, 100, 500), (0.5, 200, 1000)\}$. In TCGA 34 class classification, we used cross entropy instead of binary cross entropy and applied a softmax function to the last layer. 'the drop rate', 'the drop interval', and 'the epoch' were optimized from the set $\{(0.1, 100, 300), (0.1, 200, 500), (0.25, 100, 500), (0.25, 200, 1000), (0.5, 100, 500), (0.5, 200, 1000), (0.75, 200, 1000), (0.75, 200, 2000), (0.75, 1000, 5000), (0.75, 1500, 7500)\}$. And, it was selected whether to combine or intersect each feature group in multi-omics data for optimization.

C. Analysis

Table S3 shows performance comparison in terms of average precision (AP) in addition to other metrics in Table 2 in the main manuscript. MOMA demonstrated the best AP in the ROSMAP cohort. In the TCGA early- and late-stage classification, MOMA showed the best AP performance in 8 of 18 data sets. In addition, we performed a two-tailed paired *t*-test between other methods and MOMA for 5 s × 20 Tasks (including ROSMAP normal (NL)/Alzheimer's disease (AD) classification, TCGA 34 classes classification, and Early- and late-stage classification of 18 cancer types) (Table S4) to show the significance of performance improvement.

We compared the computational time, the number of parameters, memory usage, and GPU memory usage of XGBoost, DNN, MORONET, MOFA, SMSPL, P-NET, and MOMA (Table S5) on a 10-core Intel i9-10900X CPU and an NVIDIA TITAN RTX GPU. We counted the number of

[2] Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." Molecular systems biology 14.6 (2018): e8124.

[3] Yang, Ziyi, et al. "SMSPL: Robust Multimodal Approach to Integrative Analysis of Multiomics Data." IEEE Transactions on Cybernetics (2020).

[4] Elmarakeby, Haitham A., et al. "Biologically informed deep neural network for prostate cancer discovery." Nature 598.7880 (2021): 348-352.

parameters for the neural network-based models. XGBoost and SMSPL only used the CPU. In MOFA, the resources were estimated for the MOFA model, not including the resources used for SVM for classification.

3. Our Proposed Multi-task Attention Learning Algorithm for MRI, PET, and gene expression data

We applied MOMA with new three datasets to predict Alzheimer's disease and control on the ADNI-TADPOLE[5] data: two bioimaging datasets (MRI ROIs and AV45 PET ROIs) and gene expression data.

A. MOMA algorithm on three different datasets

Given a training sample $\{x^1, x^2, x^3, y\}$, where x^1, x^2, and x^3 denote the sample under the PET profile, the MRI profile, and the gene expression profile, and y is the corresponding label. Let f_{module} denotes the module encoder. The module vector of each datasets is defined as follows:

$$
\begin{aligned}
M^1(x^1) &= f^1_{module}\left(X^1; \theta^1_{module}\right) \in \mathbb{R}^{N^1 \times D}, \\
M^2(x^2) &= f^2_{module}\left(X^2; \theta^2_{module}\right) \in \mathbb{R}^{N^2 \times D}, \\
M^3(x^3) &= f^3_{module}\left(X^3; \theta^3_{module}\right) \in \mathbb{R}^{N^3 \times D}.
\end{aligned}
$$

where f_{module} consists of the fully connected layer and unit vector normalization, θ_{module} denotes the weights of f_{module}, D is the dimension of the module vector, N^1, N^2, and N^3 are the number of PET modules, the number of MRI modules and number of gene expression modules, respectively, and M^1, M^2, and M^3 are the module vectors of PET, MRI, and gene expression, respectively. We devised a module attention mechanism that focuses the relationship between the three modules. We devised ${}_3P_2$ attention matrices.

$$
\begin{aligned}
Att^1_{lk}(x) &= \frac{exp\left(\cos\left(M^1_l, M^2_k\right)\right)}{\sum_{k=1}^{N^2} exp\left(\cos\left(M^1_l, M^2_k\right)\right)}, \ Att^2_{lk}(x) = \frac{exp\left(\cos\left(M^2_l, M^1_k\right)\right)}{\sum_{k=1}^{N^1} exp\left(\cos\left(M^2_l, M^1_k\right)\right)}, \\
Att^3_{lk}(x) &= \frac{exp\left(\cos\left(M^1_l, M^3_k\right)\right)}{\sum_{k=1}^{N^3} exp\left(\cos\left(M^1_l, M^3_k\right)\right)}, \ Att^4_{lk}(x) = \frac{exp\left(\cos\left(M^3_l, M^1_k\right)\right)}{\sum_{k=1}^{N^1} exp\left(\cos\left(M^3_l, M^1_k\right)\right)}, \\
Att^5_{lk}(x) &= \frac{exp\left(\cos\left(M^2_l, M^3_k\right)\right)}{\sum_{k=1}^{N^3} exp\left(\cos\left(M^2_l, M^3_k\right)\right)}, \ Att^6_{lk}(x) = \frac{exp\left(\cos\left(M^3_l, M^2_k\right)\right)}{\sum_{k=1}^{N^2} exp\left(\cos\left(M^3_l, M^2_k\right)\right)},
\end{aligned}
$$

where $M = M(x)$ for short, M_l denotes l-th module of the module vector. Each element of Att_{lk} stores the relation information with possible dependence between the l-th module from one dataset and the k-th module from another dataset module. Modules with high similarity to the other two data sets were focused and used for prediction. Loss L is set to the cross-entropy error between the gold label and task-specific outputs:

$$
\begin{aligned}
L = -\sum_{c=1}^{C} \Bigg(& y_c \cdot log\left(f^1_{fc}\left(\left(Att^1(x)\right)^T M^1, \left(Att^3(x)\right)^T M^1; \theta^1_{fc}\right)\right) \\
& + y_c \cdot log\left(f^2_{fc}\left(\left(Att^2(x)\right)^T M^2, \left(Att^5(x)\right)^T M^2; \theta^2_{fc}\right)\right) \\
& + y_c \cdot log\left(f^3_{fc}\left(\left(Att^4(x)\right)^T M^3, \left(Att^6(x)\right)^T M^3; \theta^3_{fc}\right)\right)\Bigg) \\
& + \ \lambda \sum W^2, \quad .s.t. W \in \{\theta_{module}, \theta_{fc}\},
\end{aligned}
$$

where $M = M(x)$ for short, C represents the total number of classes, y_c denotes a labeling of c, f^1_{fc}, f^2_{fc}, and f^3_{fc} consist of multiple fully connected layers, θ_{fc} denotes the weights of f_{fc}, and W denotes the weight. A L2-norm penalty with a regularization parameter λ was used to the optimization to avoid overfitting.

[5] Jack Jr, Clifford R., et al. "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods." Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 27.4 (2008): 685-691

Table S1. The results of the grid search for CV1, CV2, and CV3 on ROSMAP NL/AD classification task.

Outer CV	# of Module	Early Stopping Patience	Learning Rate	Weight Decay	AUC
1	32	50	5.E-07	0.E+00	0.78015
1	32	50	5.E-07	1.E-07	0.7912
1	32	50	5.E-07	1.E-05	0.77847
1	32	50	5.E-07	1.E-03	0.75049
1	32	50	5.E-06	0.E+00	0.80132
1	32	50	5.E-06	1.E-07	0.78764
1	32	50	5.E-06	1.E-05	0.79101
1	32	50	5.E-06	1.E-03	0.787
1	32	100	5.E-07	0.E+00	0.79516
1	32	100	5.E-07	1.E-07	0.78287
1	32	100	5.E-07	1.E-05	0.79167
1	32	100	5.E-07	1.E-03	0.77334
1	32	100	5.E-06	0.E+00	0.78366
1	32	100	5.E-06	1.E-07	0.79824
1	32	100	5.E-06	1.E-05	0.77645
1	32	100	5.E-06	1.E-03	0.77381
1	64	50	5.E-07	0.E+00	0.79463
1	64	50	5.E-07	1.E-07	0.76707
1	64	50	5.E-07	1.E-05	0.77448
1	64	50	5.E-07	1.E-03	0.81526
1	64	50	5.E-06	0.E+00	0.79809
1	64	50	5.E-06	1.E-07	0.781
1	64	50	5.E-06	1.E-05	0.78015
1	64	50	5.E-06	1.E-03	0.78873
1	64	100	5.E-07	0.E+00	0.80313
1	64	100	5.E-07	1.E-07	0.7817
1	64	100	5.E-07	1.E-05	0.7921
1	64	100	5.E-07	1.E-03	0.78599
1	64	100	5.E-06	0.E+00	0.80546
1	64	100	5.E-06	1.E-07	0.78343
1	64	100	5.E-06	1.E-05	0.77942
1	64	100	5.E-06	1.E-03	0.79872
1	128	50	5.E-07	0.E+00	0.78609
1	128	50	5.E-07	1.E-07	0.7888
1	128	50	5.E-07	1.E-05	0.77888
1	128	50	5.E-07	1.E-03	0.79576
1	128	50	5.E-06	0.E+00	0.80557
1	128	50	5.E-06	1.E-07	0.78851
1	128	50	5.E-06	1.E-05	0.79403
1	128	50	5.E-06	1.E-03	0.79167
1	128	100	5.E-07	0.E+00	0.78608
1	128	100	5.E-07	1.E-07	0.78847
1	128	100	5.E-07	1.E-05	0.79956
1	128	100	5.E-07	1.E-03	0.80263
1	128	100	5.E-06	0.E+00	0.7951
1	128	100	5.E-06	1.E-07	0.7909
1	128	100	5.E-06	1.E-05	0.79269
1	128	100	5.E-06	1.E-03	0.79343
2	32	50	5.E-07	0.E+00	0.73584
2	32	50	5.E-07	1.E-07	0.74689
2	32	50	5.E-07	1.E-05	0.76896
2	32	50	5.E-07	1.E-03	0.75362
2	32	50	5.E-06	0.E+00	0.7723
2	32	50	5.E-06	1.E-07	0.74739
2	32	50	5.E-06	1.E-05	0.75683
2	32	50	5.E-06	1.E-03	0.74124
2	32	100	5.E-07	0.E+00	0.75316
2	32	100	5.E-07	1.E-07	0.76737
2	32	100	5.E-07	1.E-05	0.75567
2	32	100	5.E-07	1.E-03	0.77456
2	32	100	5.E-06	0.E+00	0.76921
2	32	100	5.E-06	1.E-07	0.75129
2	32	100	5.E-06	1.E-05	0.75825
2	32	100	5.E-06	1.E-03	0.78148
2	64	50	5.E-07	0.E+00	0.75962
2	64	50	5.E-07	1.E-07	0.75712
2	64	50	5.E-07	1.E-05	0.76868
2	64	50	5.E-07	1.E-03	0.77296
2	64	50	5.E-06	0.E+00	0.76444
2	64	50	5.E-06	1.E-07	0.76441
2	64	50	5.E-06	1.E-05	0.76371
2	64	50	5.E-06	1.E-03	0.74992
2	64	100	5.E-07	0.E+00	0.75609
2	64	100	5.E-07	1.E-07	0.76176
2	64	100	5.E-07	1.E-05	0.75431
2	64	100	5.E-07	1.E-03	0.76369
2	64	100	5.E-06	0.E+00	0.76819
2	64	100	5.E-06	1.E-07	0.77012
2	64	100	5.E-06	1.E-05	0.769
2	64	100	5.E-06	1.E-03	0.77517
2	128	50	5.E-07	0.E+00	0.76187
2	128	50	5.E-07	1.E-07	0.76789
2	128	50	5.E-07	1.E-05	0.76013
2	128	50	5.E-07	1.E-03	0.75842
2	128	50	5.E-06	0.E+00	0.77455
2	128	50	5.E-06	1.E-07	0.7594
2	128	50	5.E-06	1.E-05	0.75971
2	128	50	5.E-06	1.E-03	0.77038
2	128	100	5.E-07	0.E+00	0.7739
2	128	100	5.E-07	1.E-07	0.76741
2	128	100	5.E-07	1.E-05	0.76746
2	128	100	5.E-07	1.E-03	0.77596
2	128	100	5.E-06	0.E+00	0.7529
2	128	100	5.E-06	1.E-07	0.75918
2	128	100	5.E-06	1.E-05	0.75844
2	128	100	5.E-06	1.E-03	0.76722
3	32	50	5.E-07	0.E+00	0.76411
3	32	50	5.E-07	1.E-07	0.76033
3	32	50	5.E-07	1.E-05	0.76643
3	32	50	5.E-07	1.E-03	0.7776
3	32	50	5.E-06	0.E+00	0.77533
3	32	50	5.E-06	1.E-07	0.71794
3	32	50	5.E-06	1.E-05	0.75641
3	32	50	5.E-06	1.E-03	0.76363
3	32	100	5.E-07	0.E+00	0.7609
3	32	100	5.E-07	1.E-07	0.77717
3	32	100	5.E-07	1.E-05	0.77276
3	32	100	5.E-07	1.E-03	0.76823
3	32	100	5.E-06	0.E+00	0.77851
3	32	100	5.E-06	1.E-07	0.78613
3	32	100	5.E-06	1.E-05	0.79264
3	32	100	5.E-06	1.E-03	0.77251
3	64	50	5.E-07	0.E+00	0.79161
3	64	50	5.E-07	1.E-07	0.78826
3	64	50	5.E-07	1.E-05	0.79433
3	64	50	5.E-07	1.E-03	0.79331
3	64	50	5.E-06	0.E+00	0.77278
3	64	50	5.E-06	1.E-07	0.75814
3	64	50	5.E-06	1.E-05	0.78425
3	64	50	5.E-06	1.E-03	0.77224
3	64	100	5.E-07	0.E+00	0.7729
3	64	100	5.E-07	1.E-07	0.77539
3	64	100	5.E-07	1.E-05	0.7701
3	64	100	5.E-07	1.E-03	0.77676
3	64	100	5.E-06	0.E+00	0.78651
3	64	100	5.E-06	1.E-07	0.77586
3	64	100	5.E-06	1.E-05	0.77232
3	64	100	5.E-06	1.E-03	0.78252
3	128	50	5.E-07	0.E+00	0.78048
3	128	50	5.E-07	1.E-07	0.7925
3	128	50	5.E-07	1.E-05	0.78025
3	128	50	5.E-07	1.E-03	0.76891
3	128	50	5.E-06	0.E+00	0.77706
3	128	50	5.E-06	1.E-07	0.78213
3	128	50	5.E-06	1.E-05	0.79041
3	128	50	5.E-06	1.E-03	0.79413
3	128	100	5.E-07	0.E+00	0.76442
3	128	100	5.E-07	1.E-07	0.78628
3	128	100	5.E-07	1.E-05	0.78573
3	128	100	5.E-07	1.E-03	0.77989
3	128	100	5.E-06	0.E+00	0.78387
3	128	100	5.E-06	1.E-07	0.76381
3	128	100	5.E-06	1.E-05	0.7844
3	128	100	5.E-06	1.E-03	0.78252

Table S2. The results of the grid search for CV4 and CV5 on ROSMAP NL/AD classification task.

Outer CV	# of Module	Early Stopping Patience	Learning Rate	Weight Decay	AUC
4	32	50	5.E-07	0.E+00	0.6945
4	32	50	5.E-07	1.E-07	0.7151
4	32	50	5.E-07	1.E-05	0.7135
4	32	50	5.E-07	1.E-03	0.69261
4	32	50	5.E-06	0.E+00	0.6882
4	32	50	5.E-06	1.E-07	0.72246
4	32	50	5.E-06	1.E-05	0.71966
4	32	50	5.E-06	1.E-03	0.73126
4	32	100	5.E-07	0.E+00	0.71821
4	32	100	5.E-07	1.E-07	0.72505
4	32	100	5.E-07	1.E-05	0.74258
4	32	100	5.E-07	1.E-03	0.72475
4	32	100	5.E-06	0.E+00	0.71271
4	32	100	5.E-06	1.E-07	0.73879
4	32	100	5.E-06	1.E-05	0.72145
4	32	100	5.E-06	1.E-03	0.73635
4	64	50	5.E-07	0.E+00	0.73149
4	64	50	5.E-07	1.E-07	0.72359
4	64	50	5.E-07	1.E-05	0.71524
4	64	50	5.E-07	1.E-03	0.71944
4	64	50	5.E-06	0.E+00	0.71931
4	64	50	5.E-06	1.E-07	0.72226
4	64	50	5.E-06	1.E-05	0.75134
4	64	50	5.E-06	1.E-03	0.72675
4	64	100	5.E-07	0.E+00	0.72736
4	64	100	5.E-07	1.E-07	0.73682
4	64	100	5.E-07	1.E-05	0.71975
4	64	100	5.E-07	1.E-03	0.74369
4	64	100	5.E-06	0.E+00	0.72949
4	64	100	5.E-06	1.E-07	0.72526
4	64	100	5.E-06	1.E-05	0.73535
4	64	100	5.E-06	1.E-03	0.73051
4	128	50	5.E-07	0.E+00	0.72785
4	128	50	5.E-07	1.E-07	0.72667
4	128	50	5.E-07	1.E-05	0.72747
4	128	50	5.E-07	1.E-03	0.73758
4	128	50	5.E-06	0.E+00	0.72502
4	128	50	5.E-06	1.E-07	0.73003
4	128	50	5.E-06	1.E-05	0.74665
4	128	50	5.E-06	1.E-03	0.73274
4	128	100	5.E-07	0.E+00	0.72785
4	128	100	5.E-07	1.E-07	0.73505
4	128	100	5.E-07	1.E-05	0.71253
4	128	100	5.E-07	1.E-03	0.73953
4	128	100	5.E-06	0.E+00	0.73391
4	128	100	5.E-06	1.E-07	0.73565
4	128	100	5.E-06	1.E-05	0.72856
4	128	100	5.E-06	1.E-03	0.71842

Outer CV	# of Module	Early Stopping Patience	Learning Rate	Weight Decay	AUC
5	32	50	5.E-07	0.E+00	0.8271
5	32	50	5.E-07	1.E-07	0.8012
5	32	50	5.E-07	1.E-05	0.77812
5	32	50	5.E-07	1.E-03	0.78914
5	32	50	5.E-06	0.E+00	0.80723
5	32	50	5.E-06	1.E-07	0.80857
5	32	50	5.E-06	1.E-05	0.80276
5	32	50	5.E-06	1.E-03	0.79546
5	32	100	5.E-07	0.E+00	0.8004
5	32	100	5.E-07	1.E-07	0.78903
5	32	100	5.E-07	1.E-05	0.80052
5	32	100	5.E-07	1.E-03	0.80184
5	32	100	5.E-06	0.E+00	0.80571
5	32	100	5.E-06	1.E-07	0.78699
5	32	100	5.E-06	1.E-05	0.8054
5	32	100	5.E-06	1.E-03	0.78508
5	64	50	5.E-07	0.E+00	0.80072
5	64	50	5.E-07	1.E-07	0.81467
5	64	50	5.E-07	1.E-05	0.81205
5	64	50	5.E-07	1.E-03	0.80774
5	64	50	5.E-06	0.E+00	0.81462
5	64	50	5.E-06	1.E-07	0.8208
5	64	50	5.E-06	1.E-05	0.8123
5	64	50	5.E-06	1.E-03	0.81562
5	64	100	5.E-07	0.E+00	0.80989
5	64	100	5.E-07	1.E-07	0.7987
5	64	100	5.E-07	1.E-05	0.81249
5	64	100	5.E-07	1.E-03	0.81274
5	64	100	5.E-06	0.E+00	0.80361
5	64	100	5.E-06	1.E-07	0.81069
5	64	100	5.E-06	1.E-05	0.81217
5	64	100	5.E-06	1.E-03	0.81285
5	128	50	5.E-07	0.E+00	0.80679
5	128	50	5.E-07	1.E-07	0.81543
5	128	50	5.E-07	1.E-05	0.82375
5	128	50	5.E-07	1.E-03	0.81941
5	128	50	5.E-06	0.E+00	0.8208
5	128	50	5.E-06	1.E-07	0.82024
5	128	50	5.E-06	1.E-05	0.82279
5	128	50	5.E-06	1.E-03	0.81353
5	128	100	5.E-07	0.E+00	0.80815
5	128	100	5.E-07	1.E-07	0.8168
5	128	100	5.E-07	1.E-05	0.81261
5	128	100	5.E-07	1.E-03	0.81609
5	128	100	5.E-06	0.E+00	0.81005
5	128	100	5.E-06	1.E-07	0.82495
5	128	100	5.E-06	1.E-05	0.82536
5	128	100	5.E-06	1.E-03	0.82657

Table S3. Classification performance of XGBoost, DNN, and MOMA on ROSMAP NL and AD classifications, TCGA 34 classes classification, and TCGA early- and late-stage classification. XGBoost and DNN use single omics data gene expression (GE) and DNA methylation (DM) respectively. Gray and bold text indicate the best performance in each metric, ACC, F1-score, AUC, MCC, and AP.

Task & Datasets		XGBoost					XGBoost					DNN					DNN					MOMA									
		GE					DM					GE					DM					GE*					DM*				
		ACC	F1	AUC	MCC	AP	ACC	F1	AUC	MCC	AP	ACC	F1	AUC	MCC	AP	ACC	F1	AUC	MCC	AP	ACC	F1	AUC	MCC	AP	ACC	F1	AUC	MCC	AP
ROSMAP NL / AD		.686	.723	.757	.363	.823	.596	.664	.632	.168	.691	.653	.740	.764	.284	.806	.589	.581	.616	.180	.696	**.737**	.740	**.812**	**.488**	**.859**	.720	**.753**	.807	.432	.854
TCGA 34 Classes		.953	.951	**.998**	.951	.979	.950	.949	**.998**	.948	**.980**	.878	.848	.988	.874	.935	.875	.837	.985	.870	.924	.950	.947	.996	.948	.977	**.955**	**.952**	.996	**.953**	.978
TCGA early- and late-stage	ACC	.633	.455	.681	.191	.578	.672	.435	.673	.241	.621	.618	.386	.671	.157	.549	.578	.305	.710	.118	**.678**	**.723**	**.654**	**.725**	**.429**	.596	.710	.611	.720	.393	.609
	BLCA	**.756**	**.832**	**.772**	**.410**	**.873**	.685	.785	.688	.221	.805	.675	.806	.636	.000	.780	.699	.800	.712	.223	.816	.705	.771	.736	.352	.849	.663	.716	.747	.335	.856
	BRCA	**.718**	.179	**.630**	.101	.383	.716	.154	.584	.094	.374	.701	.188	.573	.079	.350	.705	.232	.602	.095	.367	.692	.288	.616	**.134**	.420	.651	**.331**	.618	.125	**.424**
	COAD	.634	.539	.664	.247	.637	.605	.499	.627	.185	.588	.579	.513	.626	.178	.578	.609	.367	.644	.179	.590	.630	**.624**	.703	.310	**.671**	**.667**	.605	**.726**	**.322**	**.671**
	ESCA	.547	.369	.508	.033	.477	.548	.319	.457	.013	.436	.479	.347	.472	.000	.470	**.566**	.218	**.542**	**.046**	**.505**	.509	**.467**	.526	.026	.500	.510	.417	.441	-.001	.440
	HNSC	**.780**	.874	.607	.131	.830	.764	.864	.598	.060	.839	.778	**.875**	.542	.000	.804	.778	**.875**	.501	.000	.789	.728	.832	.631	.079	.858	.648	.727	**.643**	**.196**	**.863**
	KICH	.631	.280	**.617**	.088	.522	.646	.253	.567	.062	.476	.692	.067	.533	.036	.512	.554	.201	.450	.018	.396	.631	**.357**	.556	.141	.518	**.708**	.286	**.617**	**.176**	**.554**
	KIRC	.719	.647	.781	.421	.743	.735	.662	.798	.453	.736	.719	.641	.772	.420	.716	.735	.584	.752	.429	.680	**.754**	**.717**	**.836**	.507	**.785**	**.754**	.714	.830	**.510**	.760
	KIRP	.845	.638	.845	.573	**.771**	.833	.592	.848	.523	.734	.817	.463	.819	.430	.703	.805	.447	.789	.400	.662	.813	.637	.842	.522	.749	**.849**	**.684**	**.858**	**.602**	.767
	LIHC	.725	.226	.593	.120	.387	.696	.127	.644	-.004	.353	**.733**	.088	.547	.034	.355	**.733**	.236	.665	.134	.407	.699	**.366**	.680	.188	.418	.730	.314	**.688**	**.190**	**.431**
	LUAD	.779	.056	.605	.020	.314	.792	.130	**.643**	**.132**	**.349**	**.794**	.000	.612	.000	.275	**.794**	.000	.509	.000	.237	.711	**.241**	.595	.092	.316	.761	.177	.612	.092	.316
	LUSC	.834	.029	.564	.020	**.248**	.831	.084	.460	**.085**	.209	.703	.057	**.567**	.000	.225	**.837**	.000	.533	.000	.197	.698	.190	.506	.023	.231	.695	**.224**	.500	.044	.226
	MESO	**.758**	**.837**	**.682**	**.401**	**.840**	.610	.740	.511	-.027	.757	.701	.822	.622	.033	.802	.690	.816	.421	-.032	.723	.643	.773	.572	-.064	.803	.609	.719	.607	.019	.796
	READ	**.571**	**.624**	.519	**.138**	.601	.536	.560	.582	.063	.655	.523	.620	**.605**	-.007	**.680**	.501	.415	.521	.000	.633	.410	.473	.461	-.203	.561	.465	.516	.479	-.121	.607
	SKCM	.622	.160	.457	-.059	.356	.623	.237	**.590**	.027	**.477**	.614	.092	.528	.000	.421	**.684**	.000	.382	-.030	.285	.654	**.292**	.490	**.101**	.465	.644	.153	.446	.001	.357
	STAD	.550	.582	.592	.096	.607	.569	.617	.595	.132	.625	.581	.579	.626	.173	.647	.500	.301	.541	.035	.612	**.594**	**.637**	.627	.183	**.651**	**.594**	.587	**.643**	**.198**	.649
	THCA	.719	.460	.718	.307	.616	.731	.487	.732	.343	.609	.705	.350	.694	.262	.588	**.735**	.434	.708	.322	.601	.687	.595	.761	.378	.685	.727	**.618**	**.769**	**.418**	**.694**
	UVM	.378	.358	.404	-.252	.548	.480	.454	.509	-.038	.592	.506	**.592**	**.608**	.013	**.677**	.493	.267	.469	.000	.545	.533	.501	.587	.060	.638	**.545**	.417	.580	**.115**	.625

* The dataset used for the performance measurement (a task data set) is specified.

Table S4. P-value for two-tailed t-tests compard with the MOMA (5CVs x 20 Tasks). Bold text indicate statistical significance with p < 0.05.

	MORONET	MOFA+SVM	SMSPL	P-NET
ACC	9.99.E-01	9.99.E-01	9.62.E-01	6.52.E-01
F1	**1.11.E-03**	**1.11.E-03**	**1.82.E-17**	5.80.E-02
AUC	**4.82.E-02**	**4.82.E-02**	**1.34.E-07**	**2.58.E-02**
MCC	**1.01.E-11**	**1.01.E-11**	**3.45.E-07**	**7.88.E-09**
AP	**3.36.E-07**	**3.36.E-07**	**7.32.E-03**	**2.04.E-04**

Table S5. Computational time, the number of parameters, memory usage, and GPU memory usage to train each model on the ROSMAP NL/AD classification task.

	XGBoost	DNN	MORONET	MOFA	SMSPL	P-NET	Ensem-MOMA
Computational time (s)	20	325	37	1038	177	115	697
Number of parameters	-	63947 K	576 K	-	-	132 K	3857 K
Memory usage (MB)	509	3443	2432	4834	1099	1883	3286
GPU memory usage (MB)	-	2833	607	1627	-	525	1449

B. Experiment

We compared our model with XGBoost, MOFA, MORENET, and SMSPL. Nested 5x3 CV was used for hyperparameter tuning and evaluation. For XGBoost, the parameters of 'the max depth' from the set {3, 5, 7, 9}, 'the regularization lambda' from the set {100, 10, 1, 0.1, 0.01}, and 'the learning rate' from the set {0.3, 0.2, 0.1, 0.01} were optimized. For MORONET, the parameters of 'the threshold of affinity values' from the set {2, 4, 6, 8, 10}, 'the learning rate for pretraining' from the set $\{5 \times 10^{-3}, 5 \times 10^{-4}\}$, 'the learning rate for graph convolutional network' from the set $\{5 \times 10^{-3}, 5 \times 10^{-4}\}$, 'the learning rate for classifier' from the set $\{5 \times 10^{-3}, 5 \times 10^{-4}\}$, and 'the number of significant features for each omics data type' from the set {200, 400} were used. MOFA was used for integrating multi-omics data, and the performance was measured with SVM based on the extracted factors. For SVM, the parameters of 'Regularization parameter' from set $\{0.001, 0.01, 0.1, 1, 10\}$, and the parameters of 'kernel' from the set {linear and radial basis function} were used. For SMSPL, the parameters of 'the parameter for adjusting influence from other modalities' from the set {0.66, 0.1, 0.01}, 'the age parameter' from the set $\{(0.66, 0.66, 0.66), (0.1, 0.1, 0.1), (0.01, 0.01, 0.01)\}$, 'the size of increasing the age parameter with each iteration' from the set $\{0.01, 0.02, 0.04, 0.08\}$, and 'The size to increase the selected sample for each iteration' from the set {2, 4} were optimized. For MOMA, the parameters of 'the number of modules' from the set {16, 32, 64, 128}, 'the learning rate' from the set $\{1 \times 10^{-4}, 1 \times 10^{-5}\}$, 'the weight decay' from the set $\{10^{-3}, 10^{-5}, 0\}$, and 'the early stopping patience' from the set {500, 1000} were optimized. Ensem-MOMA indicates the stacking ensemble approach using the results of each task of MOMA. Table S6 shows that our model outperformed other methods with the three data sets.

4. Our proposed multi-task attention learning algorithm for single-cell multi-omics data

We applied MOMA with single-cell multi-omics data. We used mouse embryonic stem cells (mESCs) datasets preprocessed from the previous study [6]. The mESCs datasets are composed of

[6] Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." Molecular systems biology 14.6 (2018): e8124.

Table S6. Classification performance of different methods with MRI, PET, and gene expression (GE) data of the ADNI cohort.

Model	Datasets	ACC	F1	AUC	MCC
XGBoost	MRI	0.849	0.675	0.897	0.587
XGBoost	PET	0.871	0.742	0.944	0.661
XGBoost	GE	0.727	0.104	0.594	0.078
MORONET	MRI,PET,GE	0.591	0.698	0.577	0.078
MOMA(32)+SVM	MRI,PET,GE	0.773	0.482	0.776	0.357
MOMA(64)+SVM	MRI,PET,GE	0.795	0.530	0.801	0.420
MOMA(128)+SVM	MRI,PET,GE	0.792	0.490	0.794	0.397
MOMA(256)+SVM	MRI,PET,GE	0.776	0.451	0.783	0.344
SMSPL	MRI,PET,GE	0.894	0.786	0.844[*]	0.722
MOMA	MRI[†]	0.894	0.782	0.958	0.726
	PET[†]	0.905	0.819	0.960	0.761
	GE[†]	0.902	0.809	0.958	0.750
Ensem-MOMA	MRI,PET,GE	**0.913**	**0.837**	**0.961**	**0.782**

( ) denotes the number of factors.

[*] In SMSPL, predictor is categorical. AUC is computed by categorical values.

[†] The data set used for the performance measurement (a task data set) was specified.

Bold text indicate the best performance.

Table S7. Classification performance of different methods on mouse embryonic stem cells multi-omics data.

	ACC	F1	AUC	MCC	AP
XGBoost (sc-RNA)	**0.948**	0.947	0.990	0.808	0.998
XGBoost (sc-DM)	0.844	0.781	0.866	0.137	0.959
MORONET	0.831	0.908	0.987	0.000	0.997
MOFA+SVM	0.871	0.929	0.992	0.291	0.999
SMSPL	0.875	0.928	0.683[*]	0.555	0.938
Ensem-MOMA	**0.948**	**0.969**	**1.000**	**0.820**	**1.000**

( ) denotes the dataset used for training.

[*] In SMSPL, predictor is categorical. AUC is computed by categorical values.

Bold text indicate the best performance.

64 serum-grown cells and 13 cells cultured in ’2i’ media with the single-cell RNA-seq (sc-RNA) and the single-cell DNA-methylation (sc-DM). We performed 2i/serum condition classification experiment in 3-fold CV. For all models, we used the hyperparameter set with the highest average validation performance in ROSMAP NL/AD classification task. Table S7 shows that our model is applicable to single-cell multi-omics datasets and outperformed other methods.

5. Further Analysis

We compared the performance of the module attention turning on and off according to various hyperparameter sets through inner CV results. Figures S3 - S7 shows the results of each inner CV on the TCGA cohort; each point shows a different set of hyperparameters.

Figure S8 shows that the ROSMAP samples were well clustered in the heatmap of similarity score. Figure S9 also shows the well-separated clustering results of the similarity score across the training and test dataset for the TCGA 34 class classification. Figures S10 - S11 show clustering results of the similarity score and enrichment analysis for early- and late-stage classification on KIRC and KIRP, which have the high-performance and large samples. For module analysis, only genes with a Z-score greater than the threshold Z-score (empirically set to gene expression threshold percentile = 95, DNA methylation threshold percentile = 99.99) were selected to be part of a module.

We identified cancer-specific module using the TCGA 34 classes classification model. The attention matrix was calculated based on the input values for each cancer type. For each module of one omics data, the similarity score with all modules of another omics data was averaged. And, we identified which module had the highest averaged similarity score for each cancer type. In this experiment, we used MOMA model trained in CV1. Figure S12 shows cancer-specific modules and NL-specific modules. All three examples have a high similarity score between the gene expression module and the DNA methylation module. The similarity score between BRCA-cancer-specific gene expression module 31 and DNA methylation module 19 was 0.944, the similarity score between UCEC-cancer-specific gene expression module 2 and DNA methylation module 14 was 0.976, and the similarity score between NL-specific gene expression module 10 and DNA methylation module 29 was 0.991. Figure S12 D shows the results of Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis and shows that different modules tend to be enriched in different pathways.

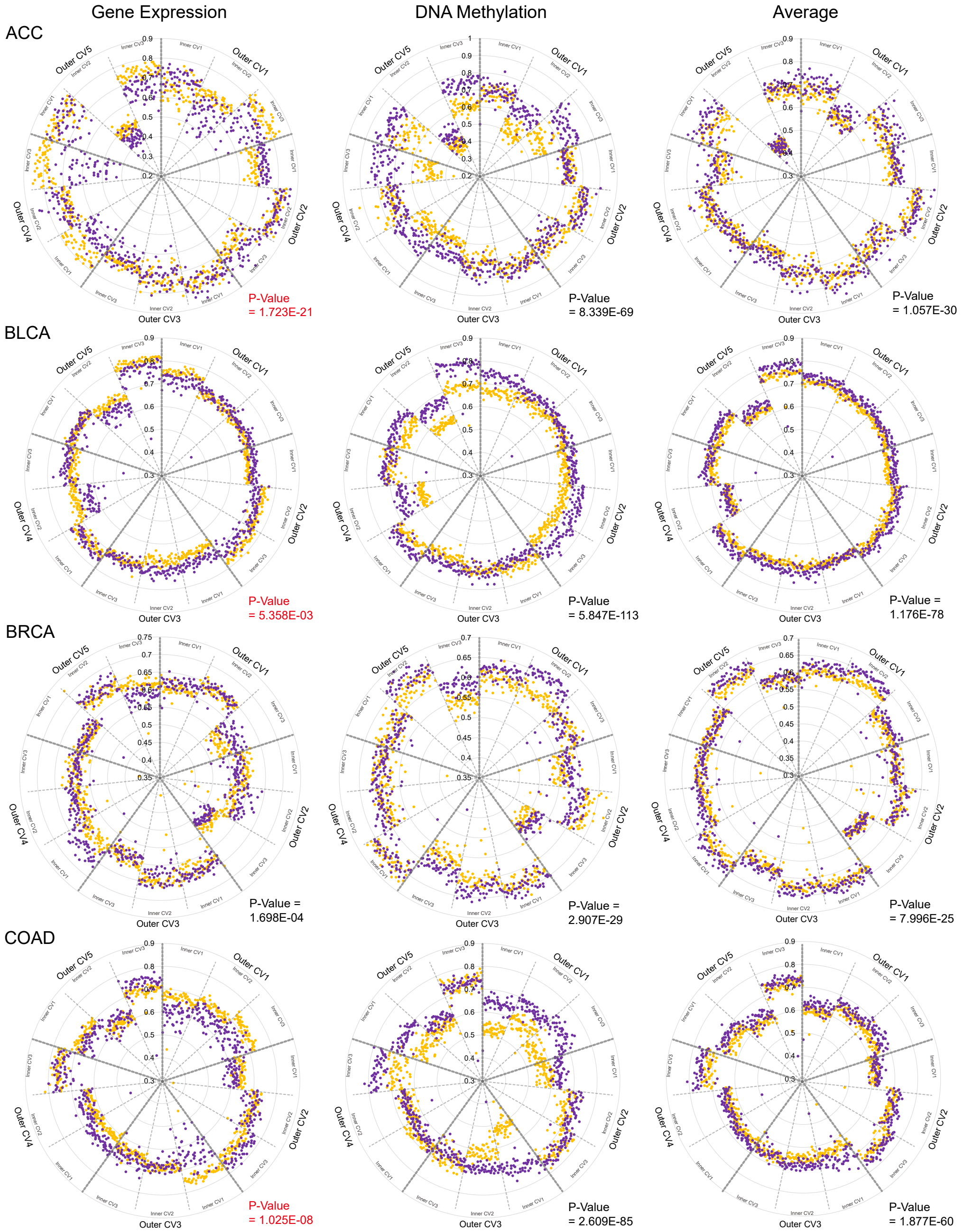


Fig. S3. Comparison of module attention turning On/Off on ACC, BLCA, BRCA, and COAD. Every inner cross validation result is reported along with the area under the receiver operating characteristic curve (AUC). Every point shows a different set of hyperparameters. Each columns shows the results of gene expression classification, the results of DNA methylation classification and the average results, respectively. P-values for a paired *t*-test for the AUC difference between turning on and off is shown, and *p*-values with a higher average AUC at turning off are marked in red.

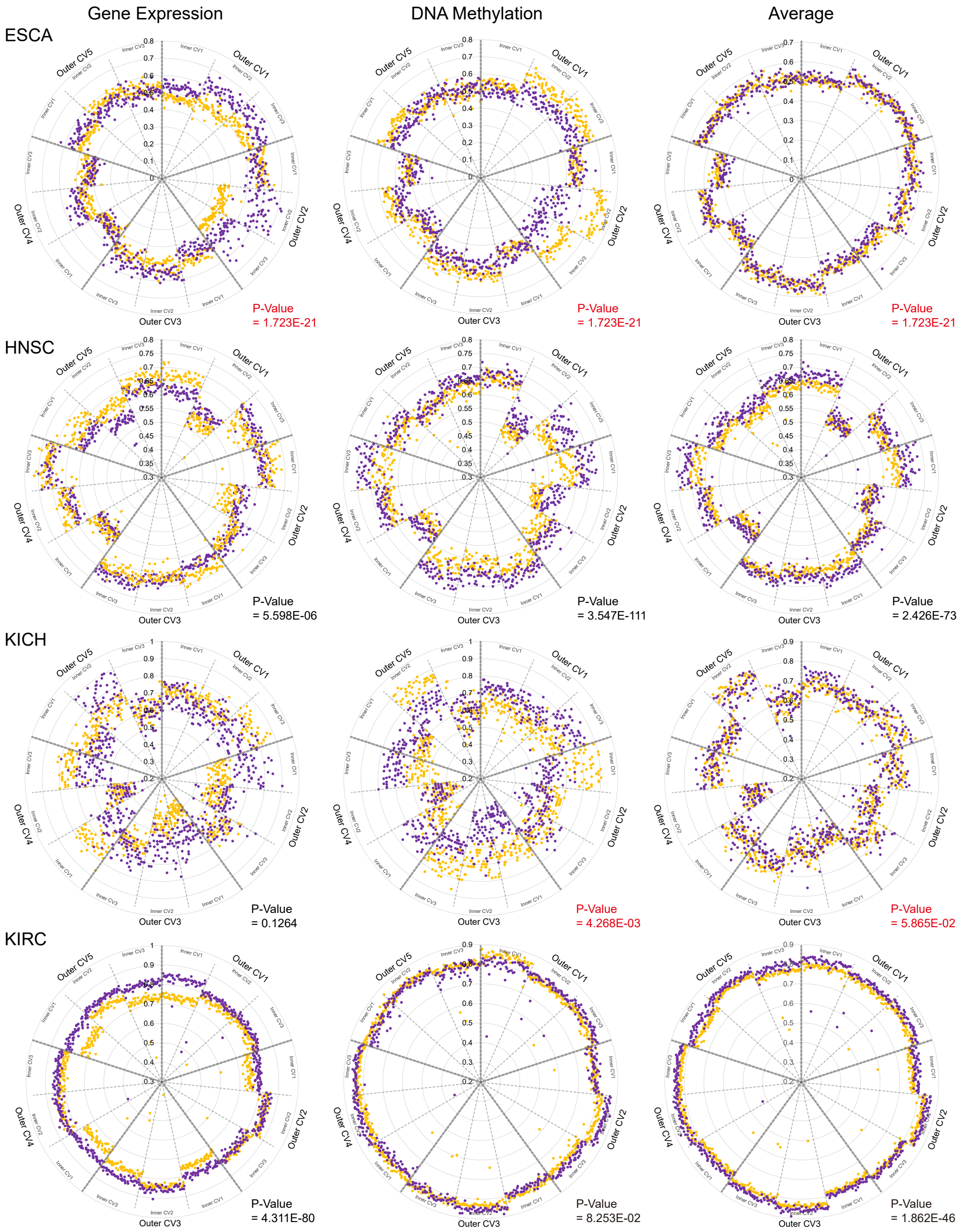


Fig. S4. Comparison of module attention turning on and off on ESCA, HNSC, KICH, and KIRC. Every inner cross validation result is reported along with the area under the receiver operating characteristic curve (AUC). Every point shows a different set of hyperparameters. Each column shows the results of gene expression classification, the results of DNA methylation classification and the average results. P-values for a paired *t*-test for the AUC difference between turning on and off is shown, and *p*-values with a higher average AUC at turning off are marked in red.

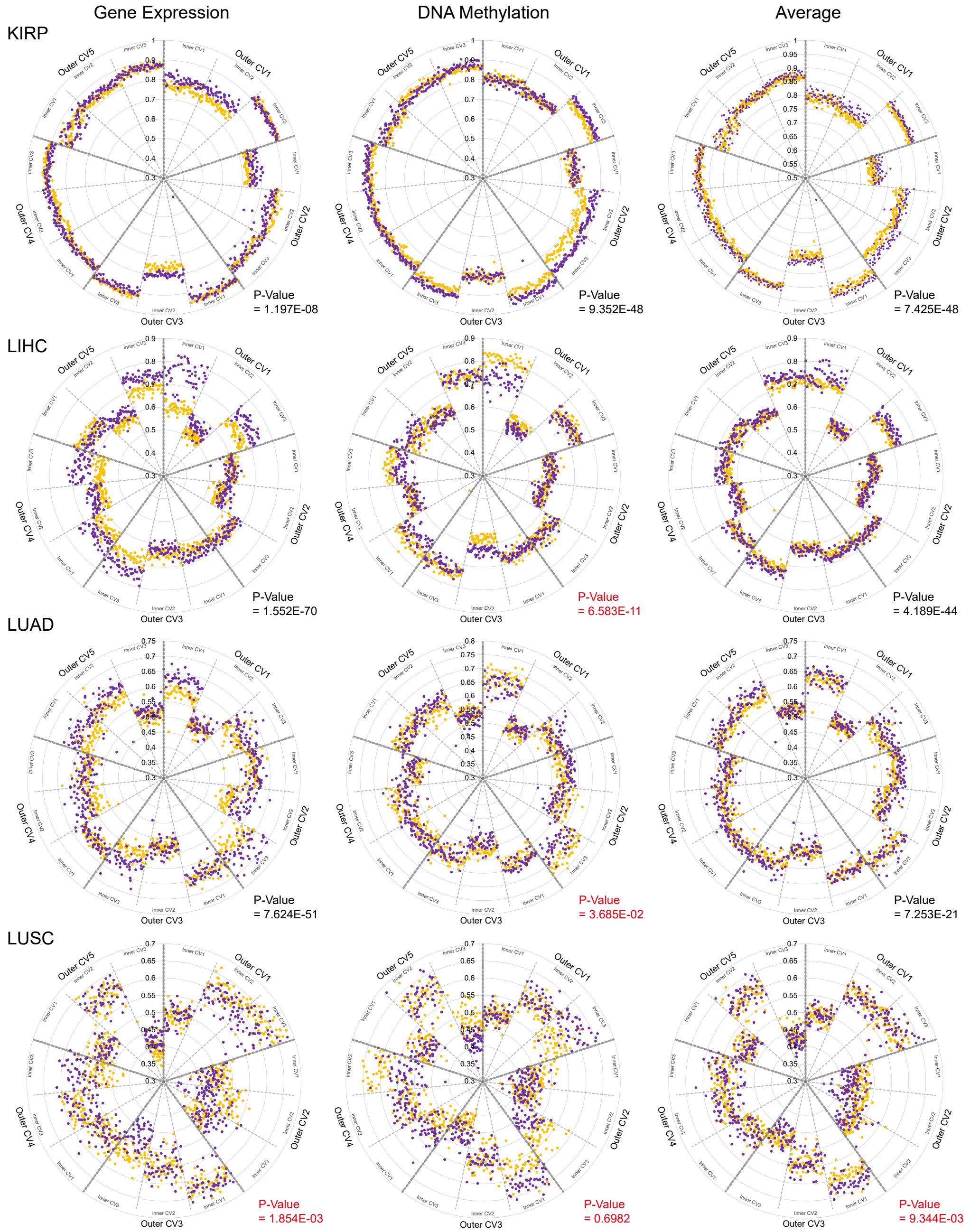


Fig. S5. Comparison of module attention turning on and off on KIRP, LIHC, LUAD, and LUSC. Every inner cross validation result is reported along with the area under the receiver operating characteristic curve (AUC). Every point shows a different set of hyperparameters. Each column shows the results of gene expression classification, the results of DNA methylation classification and the average results. P-values for a paired *t*-test for the AUC difference between turning on and off is shown, and *p*-values with a higher average AUC at turning off are marked in red.

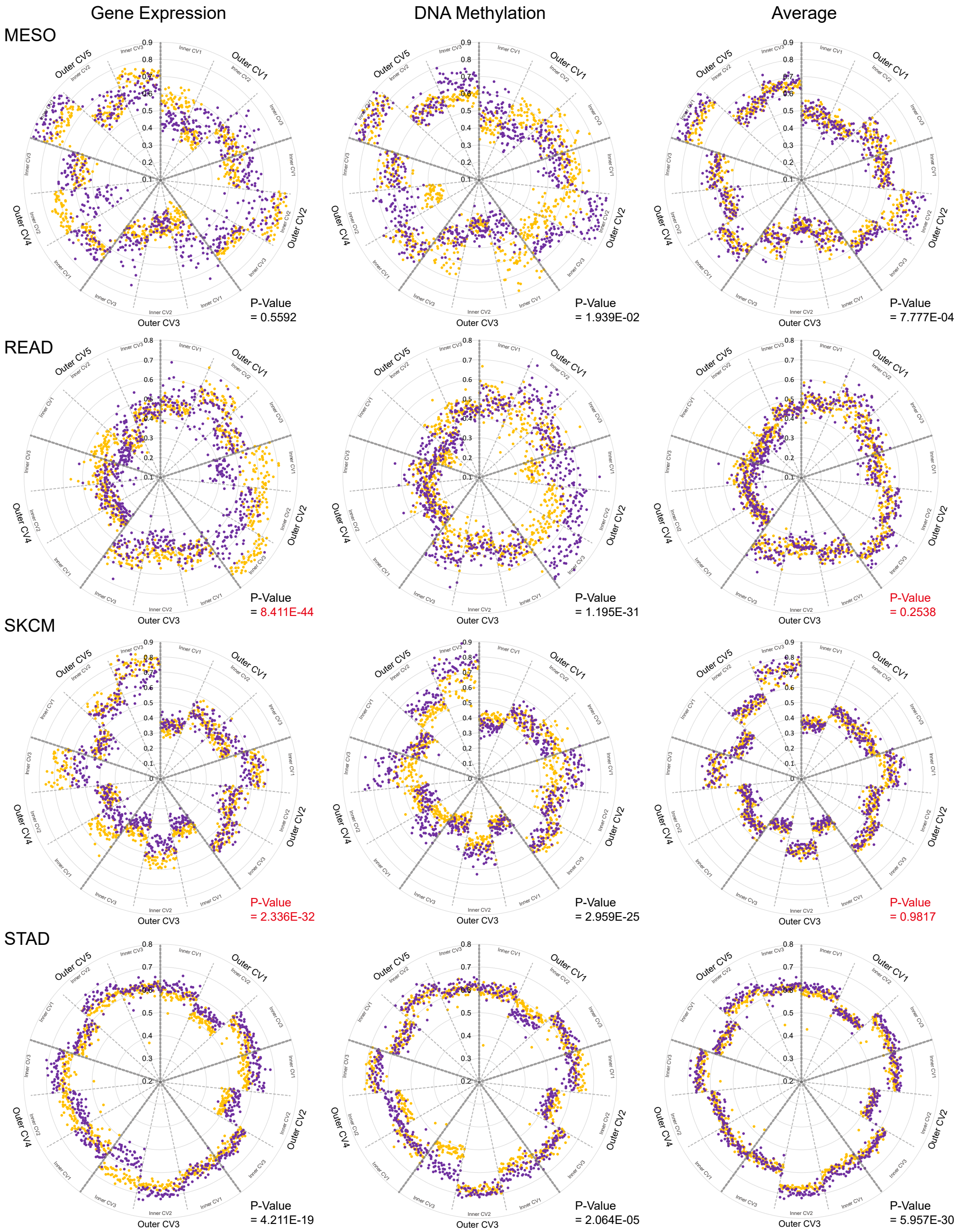


Fig. S6. Comparison of module attention turning on and off on MESO, READ, SKCM, and STAD. Every inner cross validation result is reported along with the area under the receiver operating characteristic curve (AUC). Every point shows a different set of hyperparameters. Each column shows the results of gene expression classification, the results of DNA methylation classification and the average results. P-values for a paired *t*-test for the AUC difference between turning on and off is shown, and *p*-values with a higher average AUC at turning off are marked in red.

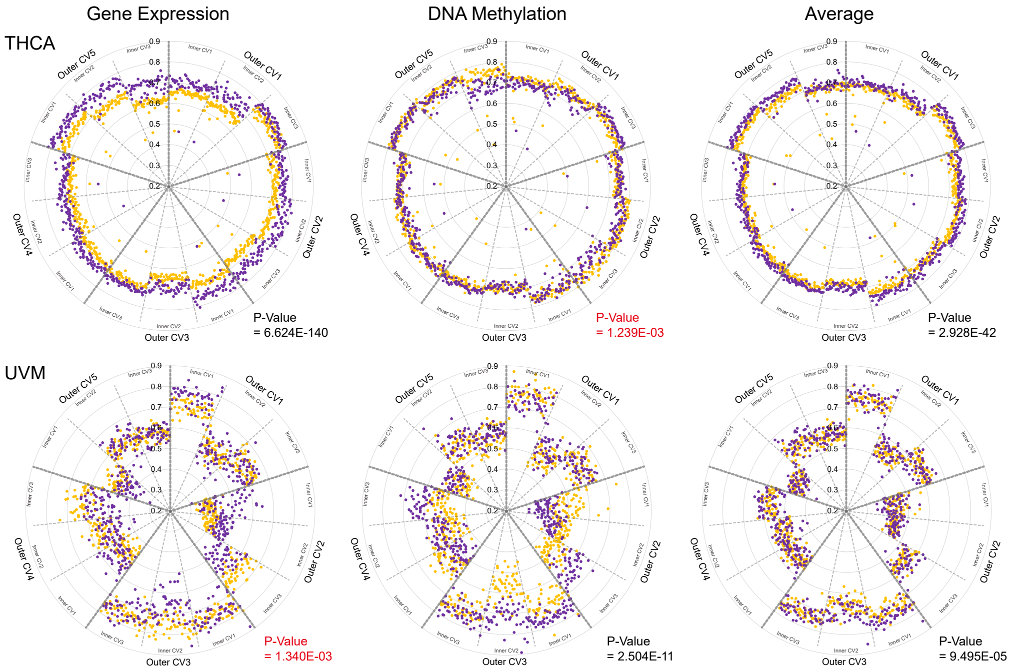


Fig. S7. Comparison of module attention turning on and off on THCA and UVM. Every inner cross validation result is reported along with the area under the receiver operating characteristic curve (AUC). Every point shows a different set of hyperparameters. Each column shows the results of gene expression classification, the results of DNA methylation classification and the average results. P-values for a paired *t*-test for the AUC difference between turning on and off is shown, and *p*-values with a higher average AUC at turning off are marked in red.

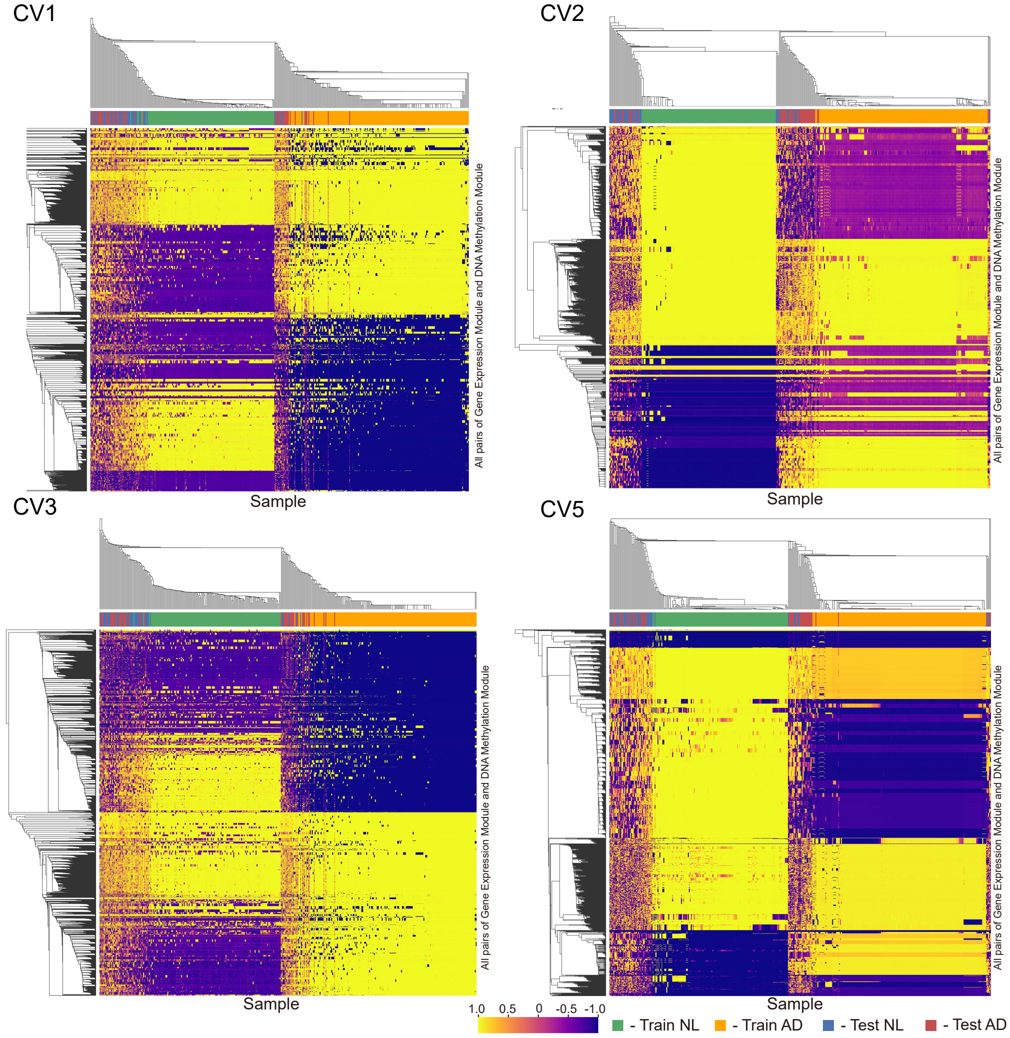


Fig. S8. Heatmap of hierarchical clustered similarity scores of the MOMA model on the ROSMAP cohort. Cosine similarity scores between the gene expression and DNA methylation modules of each sample were extracted from the each fold.

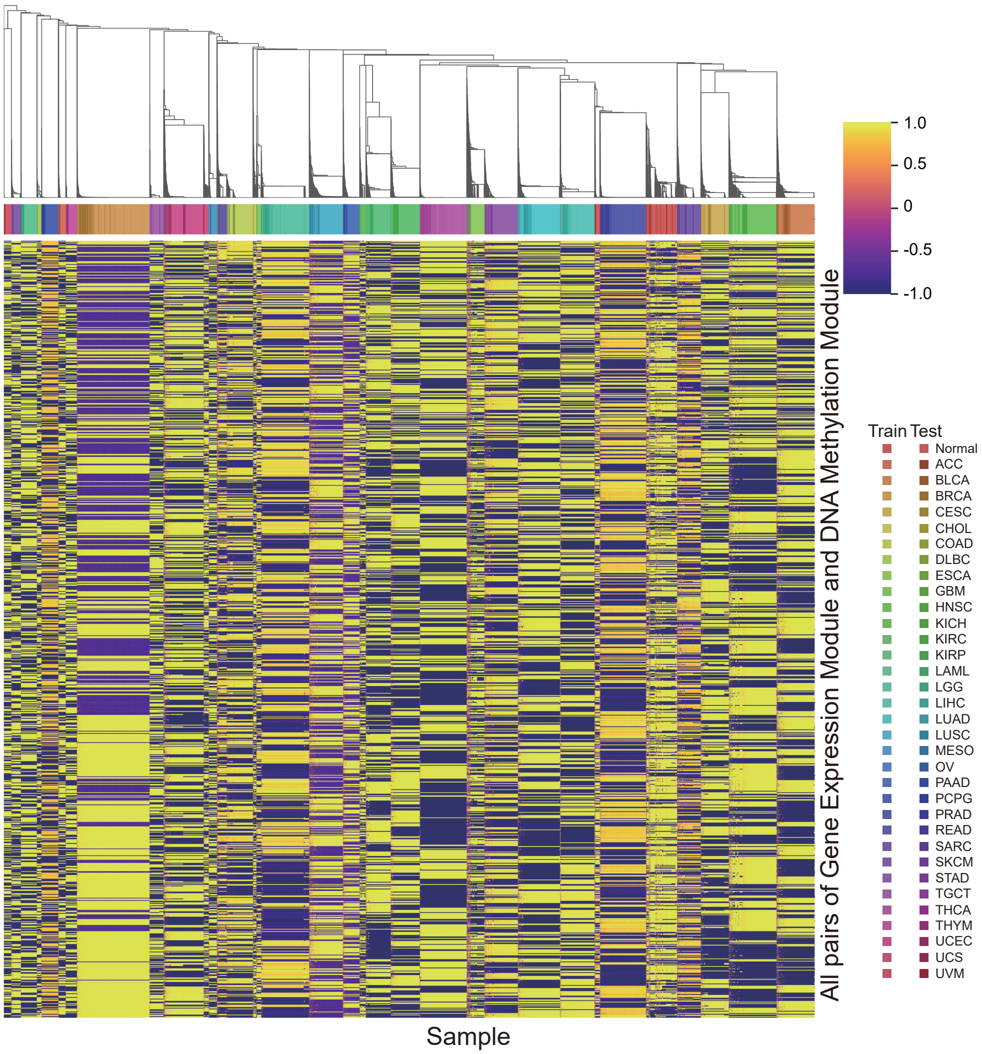


Fig. S9. Heatmap of hierarchical clustered similarity scores of MOMA model on TCGA 34 classes classification. Cosine similarity scores between the gene expression and DNA methylation modules of each sample were extracted from the each fold.

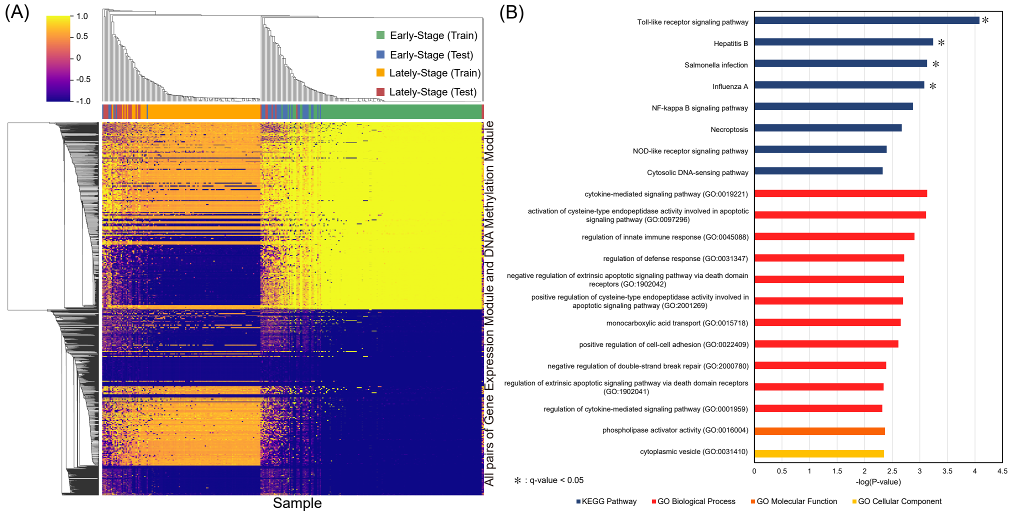


Fig. S10. (A) Heatmap of hierarchical clustered similarity scores of MOMA model for discriminating early- and late-stage of KIRC. (B) The Kyoto Encyclopedia of Genes and Genomes pathway and Gene Ontology term enrichment analysis results (P-value < 0.005).

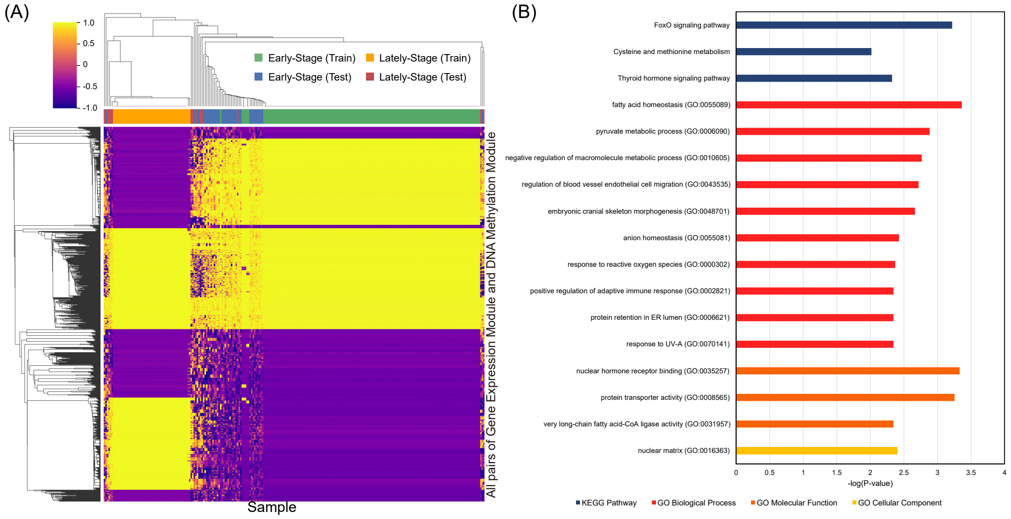


Fig. S11. (A) Heatmap of hierarchical clustered similarity scores of MOMA model for discriminating early- and late-stage of KIRP. (B) The Kyoto Encyclopedia of Genes and Genomes pathway and Gene Ontology term enrichment analysis results (P-value < 0.005).

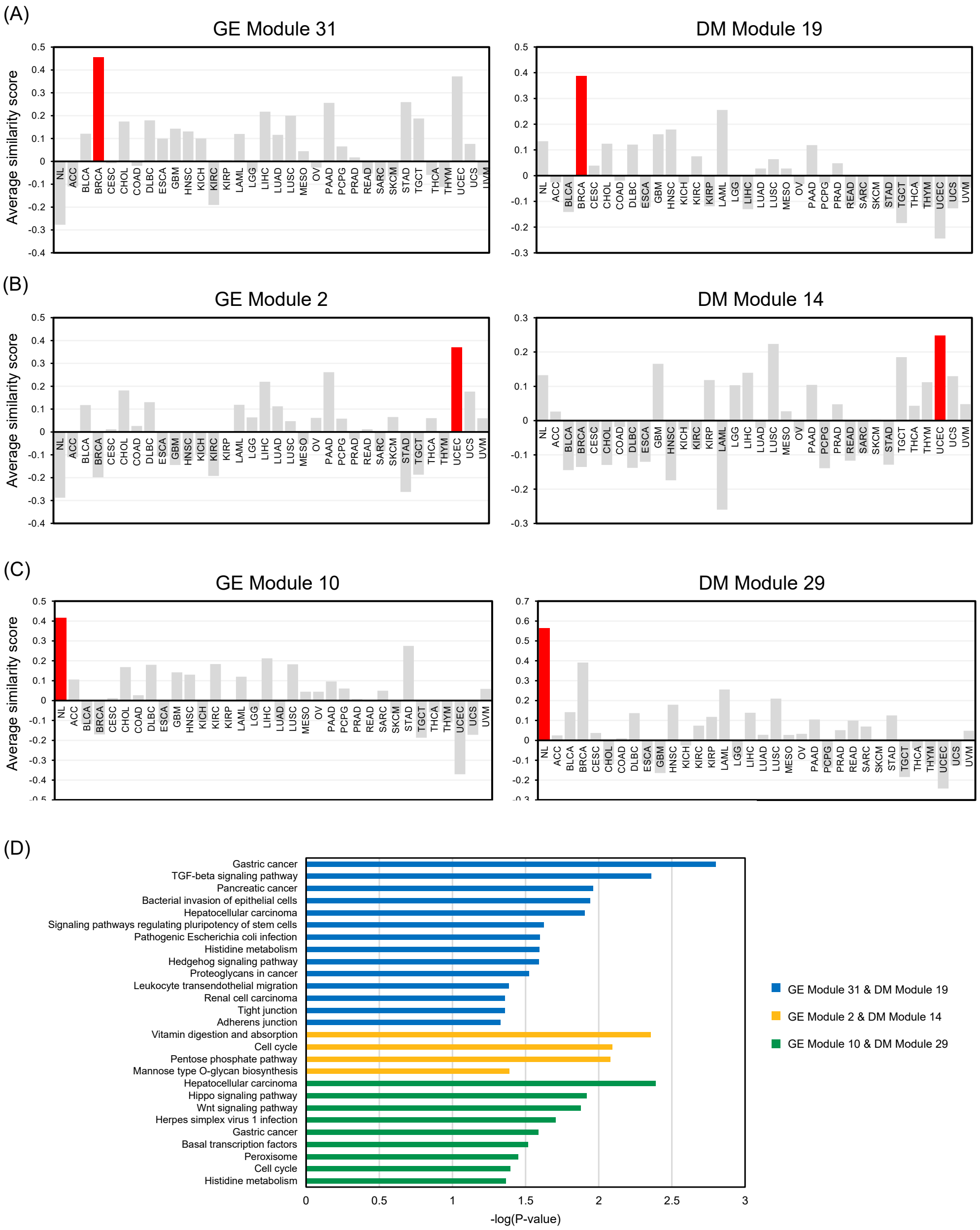


Fig. S12. (A) BRCA-cancer-specific gene expression (GE) module 31 and BRCA-cancer-specific DNA methylation (DM) module 19. (B) UCEC-cancer-specific GE module 2 and BRCA-cancer-specific DM module 14. (C) NL-specific GE module 10 and NL-specific DM module 29. (D) The Kyoto Encyclopedia of Genes enrichment analysis results of BRCA-cancer-specific GE and DM module, UCEC-cancer-specific GE and DM module,and NL-specific GE and DM module (P-value $<$ 0.05).