Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(12): 21098–21119.
DOI: 10.3934/mbe.2023933
Received: 22 September 2023
Revised: 14 November 2023
Accepted: 21 November 2023
Published: 27 November 2023

Research article

Autoencoder-assisted latent representation learning for survival prediction and multi-view clustering on multi-omics cancer subtyping

Shuwei Zhu[1], **Wenping Wang**[1], **Wei Fang**[1] **and Meiji Cui**[2,*]

[1] School of Artificial Intelligence and Computer Science, Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214122, China
[2] School of Intelligent Manufacturing, Nanjing University of Science and Technology, Nanjing 210094, China

* **Correspondence:** Email: cui_mj@163.com.

Abstract: Cancer subtyping (or cancer subtypes identification) based on multi-omics data has played an important role in advancing diagnosis, prognosis and treatment, which triggers the development of advanced multi-view clustering algorithms. However, the high-dimension and heterogeneity of multi-omics data make great effects on the performance of these methods. In this paper, we propose to learn the informative latent representation based on autoencoder (AE) to naturally capture nonlinear omic features in lower dimensions, which is helpful for identifying the similarity of patients. Moreover, to take advantage of survival information or clinical information, a multi-omic survival analysis approach is embedded when integrating the similarity graph of heterogeneous data at the multi-omics level. Then, the clustering method is performed on the integrated similarity to generate subtype groups. In the experimental part, the effectiveness of the proposed framework is confirmed by evaluating five different multi-omics datasets, taken from The Cancer Genome Atlas. The results show that AE-assisted multi-omics clustering method can identify clinically significant cancer subtypes.

Keywords: multi-omic data; cancer subtyping; multi-view clustering; autoencoder; latent space; data integration

1. Introduction

Currently, the research and application of big data technologies have penetrated from the internet fields to many other industries. Among which, the rapid development of high-throughput sequencing technologies accelerates the speed of biological big data accumulation, which has triggered a multi-faceted revolution in the research of advanced biology and medicine. Like other fields benefiting from big data, these biological big datas bring new opportunities and challenges to bioinformatics as well.

The key issue is how to discover some insightful knowledge from the biological big data efficiently, which has attracted a lot of attention from both academia and industry.

Thanks to advanced sequencing technologies, multi-omics data are generated in large quantities [1], which usually contain genomes, transcriptomes, proteomes, metabolomes, etc. It is worth noting that multi-omics data can be regarded as a specific type of multi-view data. As analyzed in the survey [2], in the past several years, more and more researchers have paid attention on analyzing multi-omics data via machine learning methods with the aim to obtain new knowledge. The general framework of Cox-regularized-model-based (or machine learning-based) multi-omics analysis is shown in Figure 1, which visually shows that each omics is represented by a data matrix from the perspective of a specific view. Although there are certain connections among each omic, the multi-source heterogeneity among multiple omics through data integration brings more potentials for discovering new knowledge, which is beneficial for disease identification and drug development. Also, some priori knowledge can be used to enhance the performance of machine learning methods. For example, a priori information about relationships between the different omics data can be considered, so as to diminish false-positive results and enhance the relevance of true molecular interactions as well.

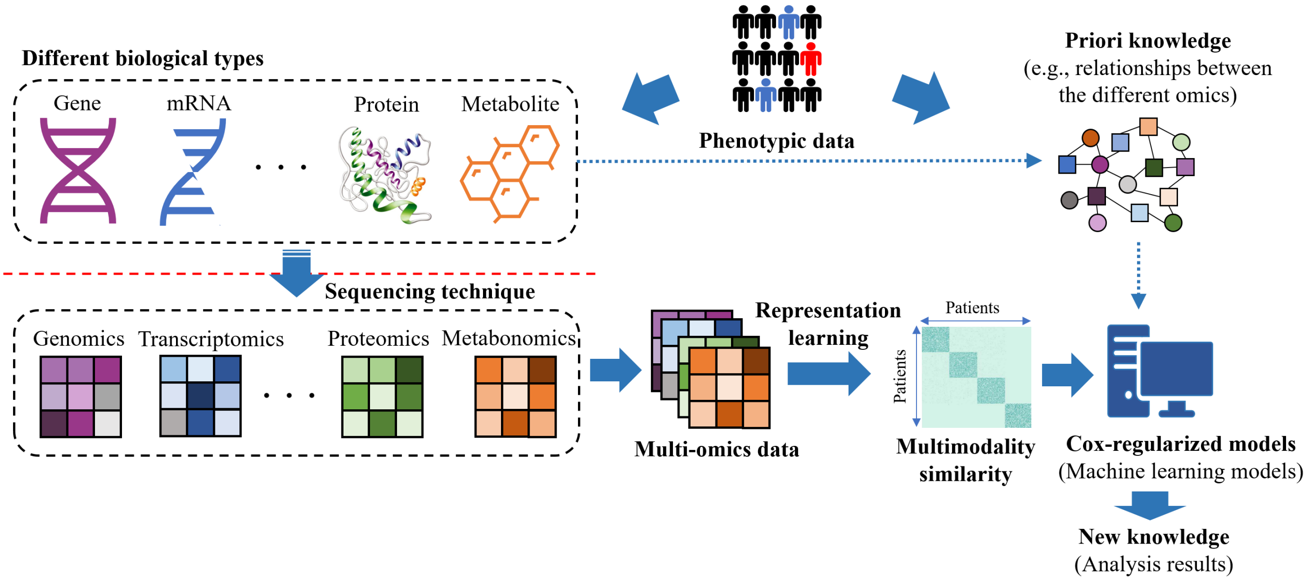


Figure 1. Multi-omics data analysis based on Cox-regularized models (or other machine learning models).

Over the last decade, considerable efforts have been devoted to the development of numerous computational methods for multi-omics data integration [3], which is the fundamental of knowledge discovery. These approaches can be roughly categorized into three classes in terms of the major strategies they used: Early, intermediate and late integration [4]. Early integration methods perform a simple concatenation of features from the omic data into a single feature combination, while late integration methods separately learn each omic layer and then merge the clustering results into a single solution. Both early and late integration methods fail to model the interactions among the features in different omics data levels. Instead, intermediate integration methods have gradually become mainstream, which consolidate data by constructing a holistic model for joint dimensionality reduction and clustering without simply concatenating features or merging results.

Nowadays, clustering of histology-oriented data has generated significant value for research in biology and medicine (e.g., disease typing, drug research, precision medicine, etc.) [5]. Among them, multi-omics data clustering which considers connections among different omics, belonging to intermediate data integration, can lead to more systematic discoveries [6]: 1) it can reduce the effects of experimental and biological noise in the data; 2) different groups can reveal different cellular levels; 3) even at the same molecular level, each group may contain data that are not available in other groups and 4) different groups can represent data from different levels of organisms. Although the existing multi-omics clustering algorithms have gained progress during the past years [7–12], they still have a large room of performance improvement by developing efficient algorithms based on advanced multi-view learning techniques, especially for large-scale multi-omics data.

In this paper, we propose the autoencoder-assisted latent space learning for survival analysis and multi-omics clustering (AELSMC) to identify meaningful cancer subtypes. First, the autoencoder (AE) aims to obtain nonlinear high-dimensional omic features in the lower dimensional space, so as to determine more accurate similarity of patients. Next, the clinical information assisted by embedding a multi-omic survival analysis approach is incorporated to learn the similarity graph of heterogeneous data at the multi-omics level. Then, we perform spectral clustering on the similarity matrix of patients given a number of clusters, and hence, generate the result of subtype groups. The proposed method is compared with some other representative algorithms on five multi-omics datasets. Experimental results have validated the promising potential of AE in capturing multi-omics feature information in lower dimensions, and the superiority of the proposed method in generating more distinguished subtypes.

2. Background

2.1. Survival prediction

Survival analysis is related to the time going by when an event begins until a censoring point. It is usually used to estimate the survival time of the observed patient [13], namely, the time from diagnosis of a disease to death. Nevertheless, it can be also concerned to any time-dependent event, which is often termed as disease-free survival, such as time in hospital or time until a disease recurs. In the literature, various survival prediction techniques are developed for clinical analysis of diseases, among them some categories are widely-used, like multi-task learning based analysis [14], deep learning based analysis model [15], and reweighted regression model [16].

Note that most of the existing survival prediction methods are developed based on a single type of data. When there exists different types of data, multi-view learning can exploit the complementary information between them by the joint optimization model so as to improve the generalization performance. Research on various multi-view learning techniques has gained a lot of attention [17], however, the development of survival prediction methods based on multi-view learning on multi-omics data is still under-explored. In view of this, we attempt to take full advantage of multi-view survival prediction on multi-omics data, thereby facilitating to deal with the tasks (e.g., cancer subtyping) in clinical analysis.

2.2. Cancer subtyping based on multi-omics analysis

During the past decade, cancer subtyping (or cancer subtypes identification) has become one of the vital steps for advancing diagnosis, prognosis and treatment. The essence of cancer subtyping is to classify patient samples with similar features of omics data, which usually adopts unsupervised or semi-supervised clustering methods.

In early time, the research of cancer subtyping focuses on clustering single omic data, such as gene expression data, which is similar to that of survival prediction techniques. However, it is insufficient today since a large quantity of multi-omics data has been generated quickly in this field. Under further research, various multi-omics clustering methods were proposed and applied to cancer subtyping, e.g., [18–21], which can be briefly summarized into three main categories: Multi-view clustering (MvC) methods, model-based methods, and similarity-based methods. Among them, MvC techniques seem to be more prevalent, as witnessed in literature. For example, a novel multi-view clustering with low-rank and sparsity constraints (MVCLRS) was proposed to capture both the global and the local structures by integrating the multi-omics data [18]. A multi-view spectral clustering with latent representation learning method was proposed in [20], which can deal with the incomplete multi-omics data with missing values. Most existing cancer subtyping methods are developed in an unsupervised manner, however, some knowledge like multi-view (omic) survival analysis (survival prediction) is very helpful in the MvC [13, 16].

Generally, omics (or multi-omics) data have the characteristics of sample scarcity and high dimensionality, hence dimension reduction or subspace learning techniques are very useful, as can be found in some recent multi-omics clustering algorithms. For example, a learning vector quantized representation based on vector-quantized variational autoEncoder was developed in [19]. The noise and redundant information in high-dimensional omics data has been addressed by the latent representation learning in [20]. The principal component analysis (PCA)-based feature extraction and singular value decomposition (SVD) were utilized for latent subspace learning in [22]. To simultaneously deal with the issues of high-level noise and high heterogeneity existing in multi-omics data, the deep latent space fusion (DLSF) model [23] was proposed based on a cycle AE with a shared self-expressive layer, which can learn consistent manifold in the sample latent space.

2.3. Multi-view clustering

In this section, we provide a more detailed discussion about MvC, owing to its predominance for multi-omics analysis. The research of MvC algorithms is a hot topic in the field of unsupervised learning, as clustering is performed by utilizing the heterogeneous perspectives of features in multi-view data to achieve accurate and meaningful solutions. In recent years, various MvC algorithms [24–27] have been proposed. Roughly, the existing MvC methods are mainly classified into three categories: 1) matrix decomposition methods [24]; 2) subspace-based methods that identify consensus low-dimensional subspaces [28] and 3) graph-based methods that utilize consensus nearest-neighbor matrices in graphs [25, 27]. Meanwhile, based on the data information employed, they can be further simply divided into two types [25]: Feature-driven methods (containing matrix decomposition and subspace-based categories) and relation-driven methods (graph-based category). Among them, the former one establishes explicit models via data features to estimate the distributions of data. The latter types aim at analyzing the point-to-point relationships of data as commonly-presented in graphs, which seeks to apply various optimization methods [25, 29, 30] on the graphs to get high-quality data partitioning.

As claimed in [25, 27], graph-based MvC methods have the advantages of simplicity and efficiency (e.g., efficiently handling nonlinear data), which have gained more attention recently. In this study, we employ the graph-based MvC technique for multi-omics clustering, for which a patient-to-patient similarity graph is learned via the AE-assisted informative latent space. At the same time, the graph learning and survival prediction are simultaneously optimized in the joint built on the latent embedding space, as discussed in the next section.

3. The proposed methods

3.1. The workflow of the proposed framework and basic definitions

An overall workflow of the proposed framework on multi-omics cancer subtyping is illustrated in Figure 2, and it has three major components. First, the more informative latent space representation (i.e., $\boldsymbol{Z}$) is captured by training the AE neural network on the multi-view (multi-omics) dataset $\boldsymbol{X}$. Therefore, two tasks–i.e., the survival analysis and similarity graph learning, are integrated into a unified optimization procedure based on the latent space representation. Finally, the cancer subtypes can be obtained by performing the spectral clustering algorithm on the affinity matrix of the patient-to-patient similarity graph, which can further provide insights for survival analysis and biological enrichment.

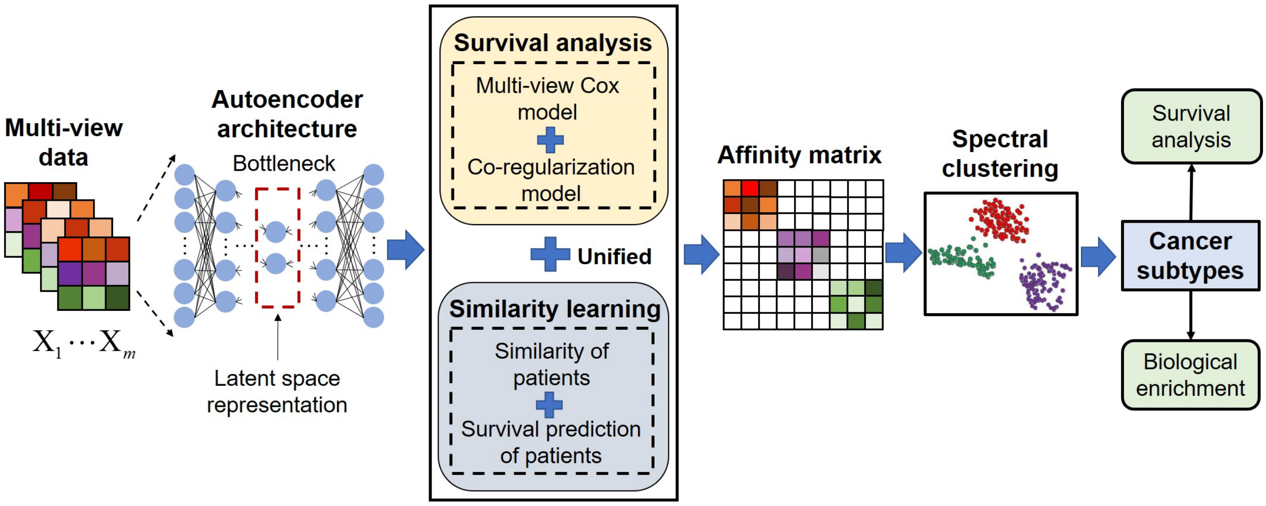


Figure 2. Workflow of the proposed framework on multi-omics cancer subtyping.

Given a dataset $\boldsymbol{X} = \{X_1^v, X_2^v, \ldots, X_n^v\}$ of n points, and there are m views, namely $\{v_1, v_2, \ldots, v_m\}$. Then for the k-th view, the feature matrix of data is $X^k \in \mathbb{R}^{p^k \times n}$, where p^k is the number of features in this view. Here, the multi-omics dataset of the k-th omic is defined as $D^k = \{d_1^k, d_2^k, \ldots, d_n^k\}$, where $d_i^k = \{X_i^k, T_i, \delta_i\}$ denotes the i-th patient. X_i^k is the aforementioned feature matrix, T_i is the observation time, and δ_i is the censoring indicator which indicates whether the patient is censored ($\delta = 0$) or observed ($\delta = 1$). Thus, T_i is defined as follows:

$$T_i = \begin{cases} O_i & \text{if } \delta_i = 1 \\ C_i & \text{if } \delta_i = 0, \end{cases} \tag{3.1}$$

where O_i denotes a survival time while C_i is a censored time.

3.2. Latent representation learning based on autoencoders

Generally, the number of samples is much less than the number of features in current biological datasets. The AE, as an efficient dimension reduction tool, can map the high-dimensional data into the low-dimensional hidden representation Z_i^v, where "v" represents a particular view of the input data.

AEs, unsupervised neural networks, attempt to restore inputs from their outputs through the process of encoding and decoding [31,32]. As shown in Figure 3, the general structure of an AE is made up of three parts, i.e., encoder, code (bottleneck), and decoder. To be specific, an encoder is a function that compresses the input into various latent representations, and thereby only useful information/features can be left by squeezing the input through a bottleneck. Accordingly, a decoder attempts to reconstruct the learned representation from the encoder back to the original format.

Here we adopt the minimization of reconstruction error as the training goal of an AE, i.e., calculating the difference between output $\boldsymbol{X'}$ and input $\boldsymbol{X}$. More specifically, each node Z_j^v in the hidden layer can be obtained as:

$$Z_j^v = \sigma(a_j^v + \sum\nolimits_i X_i^v w_{ij}^v), \tag{3.2}$$

where $\sigma(\cdot)$ is the encode function, a and w are parameters of the encoder. Hence, data in the latent space is represented as $\boldsymbol{Z}$.

The decoder takes the hidden representation as input and tries to reconstruct the original input, hence the value of each node $X_i^{v'}$ in the output layer can be calculated as:

$$X_i^{v'} = \sigma(a_i^{v'} + \sum\nolimits_j Z_j^v w_{ji}^{v'}), \tag{3.3}$$

where $\sigma(\cdot)$ is the decode function, $a^{'}$ and $w^{'}$ are parameters of the decoder.

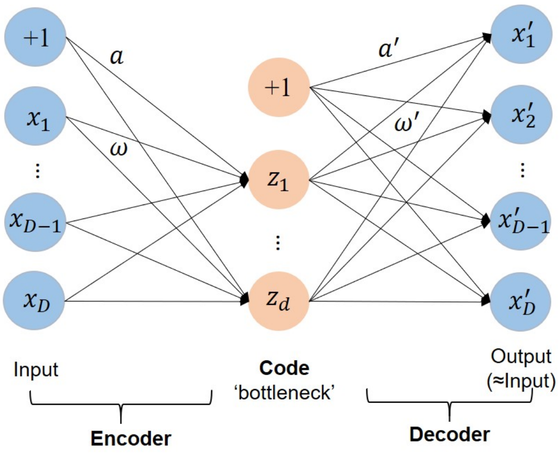


Figure 3. General structure of an AE.

The goal of training an AE is to optimize a predefined loss function. Through minimizing the loss function to reconstruct the input, the weight parameters can be updated. Regarding a particular view "v", the reconstruction loss (L_r^v) can be formulated as

$$L_r^v = \frac{1}{n} \sum_{i=1}^{n} \left\| X_i^v - X_i^{v'} \right\|^2 . \tag{3.4}$$

Therefore, the above loss function is used to update the weight parameters to reconstruct the original data $\boldsymbol{X}$, and the more informative latent space representation (i.e., $\boldsymbol{Z}$) can be captured by training the AE neural network. Then, either the survival analysis model or the similarity graph is generated from the optimized latent embedding space of $\boldsymbol{Z}$.

3.3. Survival analysis and similarity graph learning via the latent embedding space

As mentioned in Section 2.1, the survival information or clinical information is helpful for identifying meaningful cancer subtypes from the biological perspectives, as also claimed in [13,16,20]. To be more specific, the quality of the patient-to-patient similarity graph can be improved by simultaneously optimizing the learning process of the survival analysis model and the similarity matrix. Therefore, the two tasks, namely the survival analysis and similarity graph learning, are integrated into a unified optimization procedure.

Here, we employ the multi-view (multi-omics) Cox model [33] to be the survival analysis function. In addition, an omics-consistency co-regularization term [34] is introduced to explore consistent and complementary information within different omics. It aims at shrinking the agreement of the prediction between each pair of views among the multi-omic data, and hence, improves the learning performance. Therefore, the regularized survival analysis model of multi-omic data is formulated as:

$$L_{\text{survival}} = \min_{w} \underbrace{\sum_{k=1}^{m} \left(- \sum_{i=1}^{n} \delta_i \left(Z_i^k w^k - \log \sum_{j \in R_i} \exp(Z_j^k w^k) \right) \right)}_{\text{Multi-view Cox model loss function}} + \underbrace{\lambda \sum_{k \neq j} \left\| Z^k w^k - Z^j w^j \right\|_2^2}_{\text{Co-regularization (data interation)}} , \tag{3.5}$$

where $w = [w^1, w^2, \ldots, w^m]$ is the survival prediction coefficient across all views (omics) and R_i denotes the risk set of T_i, containing instances with observed time not less than T_i. Actually, function (3.5) is optimized via the latent representation of data $\boldsymbol{Z}$, rather than the original data $\boldsymbol{X}$.

Let S be the similarity matrix (or affinity matrix), which reflects the global structure of data. In S, s_{ij} is the similarity weight between the i-th and the j-th samples. A pair of similar data points can get a large weight, and vice versa.

Note that, in clinical analysis, it is usually observed that patients grouped in the same disease subtypes have similar distributions of features as well as similar survival times. Hence, a joint learning model of patient-to-patient similarity graph and survival prediction tends to discover disease subtypes more precisely. To this end, an adaptive affinity learning to measure the edge weights based on both the similarity of patient samples [35] and survival analysis [16] is defined as follows:

$$
\begin{aligned}
L_{\text{similarity}} =& \underbrace{\min_S \gamma \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left(\left\| Z_i - Z_j \right\|^2 S_{i,j} + \mu S_{i,j}^2 \right)}_{\text{Similarity of patient samples}} + \underbrace{\min_S \gamma \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left\| Z_i w - Z_j w \right\|^2 S_{i,j}}_{\text{Survival prediction of patients}} \\
=& \min_S \gamma \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left(\left\| Z_i - Z_j \right\|^2 + \left\| Z_i w - Z_j w \right\|^2 \right) S_{i,j} + \mu S_{i,j}^2 \\
& \text{s.t.} \sum_{j=1, j\neq i}^{n} S_{i,j} = 1, S_i \geq 0;\ i = 1, 2, \ldots, n,
\end{aligned}
\tag{3.6}
$$

where γ and μ are the tradeoff parameters. For simplicity, we set the value of γ and μ to 1. In the affinity matrix, a larger weight value is assigned for two patients if they get a smaller distance and similar prediction value.

It should be pointed out that the coefficient w is obtained by optimizing the survival prediction model (i.e., Eq (3.5)), such that it cannot be estimated only based on Eq (3.6). To address this issue, a collaborative learning model of survival prediction and graph affinity learning is developed by combining Eqs (3.5) and (3.6) into a joint optimization model. Hence, the unified loss function is formulated as:

$$
L_{\text{unified}} = L_{\text{survival}} + L_{\text{similarity}}. \tag{3.7}
$$

Note that it is interesting to include the AE training step into the multi-view Cox model of Eq (3.5). Specifically, a more unified optimization model can be built by adding Eq (3.4) to Eq (3.7). That is to say, step 1 (AE training) and step 2 (unified optimization of survival analysis and similarity graph learning) in Figure 2 are combined into a whole function. It needs additional theoretical analysis, which is beyond the scope of this work, and we will attempt to explore it in the future.

Therefore, the two tasks of graph learning and survival analysis are simultaneously optimized. The quality of similarity matrix S could be improved based on the distance of patient instances and prediction time, in order to identify more reasonable results of disease subtypes. Meanwhile, the similarity matrix S can in turn positively reinforce the survival analysis model. Consequently, the above two tasks can improve each other. Specifically, through the joint alternating optimization strategy [13, 16], the coefficient w of survival analysis model and the similarity matrix S iteratively update one while keeping the other one fixed. In addition, the proximal gradient algorithm [36] and the Lagrange multipliers [35, 37] need to be adopted to obtain or approximately reach the closed-form solution of w and S, respectively. For simplicity, the detailed formula derivation is not presented here, which can be referred to [13, 16]. Instead, we directly provide the iterative solution of w and S.

The updating formulation of w is achieved by keeping S fixed in L_{unified}, and let $h(w)$ be a part of L_{unified} only regarding w, i.e., $h(w) = L_{\text{survival}} + \gamma \sum_{i=1}^{n} \sum_{j=1, j\neq n}^{n} \left\| Z_i w - Z_j w \right\|^2 S_{i,j}$. According to the proximal gradient algorithm utilized in [36], w is calculated iteratively by:

$$
\begin{aligned}
w(t+1) =& \min_w h(w(t)) + < w(t+1) - w(t), \nabla h(w(t)) > + \frac{1}{2} \left\| w(t+1) - w(t) \right\|_2^2 \\
=& \min_w \frac{1}{2} \left\| w - v \right\|_2^2,
\end{aligned}
\tag{3.8}
$$

where $v = w(t) - (\tau/2) \nabla h(w(t))$ with learning step τ that can be estimated by linear search. Thus, the coefficient vector w is updated based on the following closed form solution:

$$w(t+1) = sign(v)(v - \eta)_+, \tag{3.9}$$

where $(a)_+$ is the positive function.

Similarly, the updating formulation of S is obtained by keeping w fixed in L_{unified} (L_{unified} actually equals to $L_{\text{similarity}}$ in this case). In addition, let $d_{ij} = \frac{\gamma}{\mu}\left(\left\|Z_i - Z_j\right\|^2 + \left\|Z_i w - Z_j w\right\|^2\right)$, then function (3.7) can be redefined as $\min\limits_{S_i} \sum_{j=1,j\neq i}^{n} S_{i,j} d_{i,j} + S_{i,j}^2$ with $S_{i,j}$ and S_i subject to the constraint of Eq (3.6). According to the Lagrange multipliers utilized in [35, 37], the closed-form solution of S is approximately calculated by:

$$S_{ij} = \left(\frac{1 + \sum_{j=1}^{n} \widehat{d_{ij}}}{K} - d_{ij}\right)_+, \tag{3.10}$$

where $K \in (1, n)$ is a pre-specified constant value to control the neighbour size (K=20 is set here), $\widehat{d_i}$ is obtained by sorting d_i in ascending order, and $(A)_+$ is the positive function.

After obtaining the similarity matrix (S) of the patient-to-patient graph via the latent embedding space, we can perform spectral clustering on S with a given number of clusters, and hence, generate the clustering result, namely the subtype groups.

The complete optimization procedure of the AELSMC algorithm is presented as Algorithm 1. The source code of AELSMC is freely available at https://github.com/ShuweiZhu/AELSMC.git.

Algorithm 1: The complete optimization procedure of the AELSMC algorithm

```
Input: Multi-omics dataset: $X$; Parameters: $\lambda, \beta$, and the number of subtypes $k$
Output: The identified subtypes (clustering result).
Train autoencoder on multi-omics dataset $X$ to generate the latent space representation $Z$.
Set $t = 0$, and initialize model coefficient $w$ and similarity matrix $S$.
while the unified model not converge do
    Step 1. Suvival analysis.
    Fix $S$ and estimate $\tilde{w}$ by proximal gradient algorithm, i.e., repeat:
        Compute the gradient $\nabla h(w(t))$ and let $v = w(t) - (\tau/2) \nabla h(w(t))$.
        Update $\tilde{w} = sign(v)(v - \eta)_+$ and increase the learning rate of $\tau$.
    Update $w(t+1) = \tilde{w}$.
    Step 2. Similarity learning.
    Estimate $S$ by fixing $w$: Compute $d_{ij} = \frac{\gamma}{\mu}\left(\left\|Z_i - Z_j\right\|^2 + \left\|Z_i w - Z_j w\right\|^2\right)$ and update each
      element of $S$ by Eq(3.10).
    $t = t + 1$.
end
Perform spectral clustering algorithm based on the similarity matrix $S$ to generate $k$ subtypes.
```

3.4. Clustering ensemble for a consensus solution

As mentioned in Section 1, late integration strategy can take advantage of a set of clustering solutions (based solutions) from different perspectives, such as using different methods, parameters and subspaces. As a typical late integration strategy, clustering ensemble is performed to fuse these base solutions, aiming at generating a consensus clustering [38, 39]. Usually, such a consensus clustering is beneficial from the complementary information of multiple views of omics data while making up their shortcomings. For example, a graph-based multi-method and multi-source consensus clustering strategy [12], named as ClustOmics, has been proposed based on evidence accumulation clustering (EAC) [40] to improve the robustness of predictions. A clustering ensemble method is usually composed of two main components: 1) clustering generation, and 2) consensus function. For the former, the quality and diversity of the generated clusterings are two key factors. During the last two decades, a majority number of clustering ensemble methods have been proposed, and among them, three consensus methods proposed in [41], the hyper-graph partitioning algorithm (HGPA), cluster-based similarity partitioning algorithm (CSPA) and meta-clustering algorithm (MCLA), are the most classical but still prevalent, due to their efficiency and efficacy.

In this study, we propose to take advantage of the cluster ensemble strategy by integrating useful information of multiple solutions from different perspectives, for example, to set different parameters. To do this, the powerful locally weighted clustering ensemble algorithm named LWGP [42] is used here. First, a set of clustering candidates is generated as the ensemble pool. Note that the dimension of latent space for each omic is set to be $\beta \times d$, where β is the control parameter of the size for latent space, and it takes value from interval $B = [0.01, 0.02, \ldots, 0.1]$. Thus, we can get the ensemble pool by setting different values of β in the proposed framework. Thereafter, the LWGP algorithm is executed on the ensemble pool to obtain the final solution.

4. Experimental results and analysis

In this section, we present the experimental settings, the comparison with the competitive methods of five multi-omics datasets, and the corresponding analysis (effectiveness verification and parameter analysis of the proposed framework).

4.1. Experimental settings

The proposed AELSMC algorithm is evaluated on five multi-omics datasets from the cancer genome atlas (TCGA) repository: Breast invasive carcinoma (BIC), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KRCCC), lung squamous cell carcinoma (LSCC), glioblastoma multiforme (GBM), ovarian serous cystadenocarcinoma (OV), skim cutaneous melanoma (SKCM) and sarcoma (SARC), which are commonly used in the literature [13, 22, 43]. There are three views in these datasets—i.e., miRNA expression, mRNA expression and DNA methylation. In this study, the preprocessed datasets provided by [6] are adopted and the original dataset can be found at https://gdac.broadinstitute.org/. (All the processed raw data are available at http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html.) In addition, it is worth pointing out that more different omics data can be further explored via [44–46].

Table 1 shows the properties of these multi-omics datasets, where the number of samples, censored times and the dimension of three omics (miRNA, mRNA, methylation) is presented, respectively. Note that, the Z-score standard normalization is adopted to map the original data $\boldsymbol{X}$ to the normal distribution with a mean of 0 and a standard deviation of 1, as defined below:

$$\boldsymbol{X}_{\text{stand}} = \frac{\boldsymbol{X} - \text{mean}(\boldsymbol{X})}{\text{standard deviation}(\boldsymbol{X})}. \tag{4.1}$$

Table 1. Properties of the multi-omics datasets.

	Samples	Censored	miRNA	mRNA	methylation
BIC	105	18	354	17,814	23,094
COAD	92	9	312	17,814	23,088
GBM	215	199	534	12,042	1305
KRCCC	122	33	329	17,899	24,960
LSCC	106	66	352	12,042	23,074
OV	290	174	705	5000	25,031
SARC	260	98	1046	5000	25,031
SKCM	439	213	1046	5000	25,031

The proposed AELSMC method is compared with several competing approaches: affinity network fusion (ANF) [7], similarity regression fusion (SRF) [8], similarity network fusion (SNF) [47], subspace Merging (SM) [9], DLSF [23], AutoCox [48], survival supervised graph clustering (S2GC) [13] and SparseAE (note that the author has not defined the algorithm name, here we define it as SparseAE for convenience) [43]. We conduct experiments on a PC with an Intel Core i7-1065G7 CPU and 16 GB RAM. Moreover, the Cox log-rank test (or -ln of log-rank's test p-value) [6] and the widely-used internal validity index–Silhouette [49] are adopted to evaluate the clustering performance.

Firstly, the -ln of log-rank's test p-value can measure whether the survival time is significantly different among subgroups. The log-rank test is defined as:

$$\chi^2_{\text{subgroups}} = \sum_{k=1}^{c} \frac{(O_k - E_k)^2}{E_k}, \tag{4.2}$$

where c denotes the number of clusters. O_k is the number of identified instances in the k-th subgroup while E_k is the number of expected instances. A smaller value of the p-value indicates a better result.

The Silhouette index [49] can quantify the goodness of clustering solutions by measuring the compactness within clusters and the separation between clusters [38, 39, 50], since no standard clustering result is available for the multi-omics datasets. It takes value from [-1, 1], and a larger value of the Silhouette indicates a better result.

4.2. Comparison with the state-of-the-art methods

The experimental results are obtained by using the source code provided by the authors or the results reported in the papers in the case that the source code is not available. For example, the codes of DLSF and S2GC are provided by the authors, which is very helpful for us. For convenience, the control parameter β of the size for latent space is set as 0.05 here, but the sensitivity analysis of β is presented in later subsection. Note that, it is necessary to make comparison experiments under the same conditions, so in this subsection the number of clusters k keeps consistent for all the methods on each dataset, i.e., $k = 5$ for BIC, $k = 3$ for COAD, $k = 3$ for GBM, $k = 4$ for KRCCC, $k = 4$ for LSCC, $k = 3$ for OV, $k = 2$ for SARC and $k = 3$ for SKCM, as suggested or used in literature [13, 43].

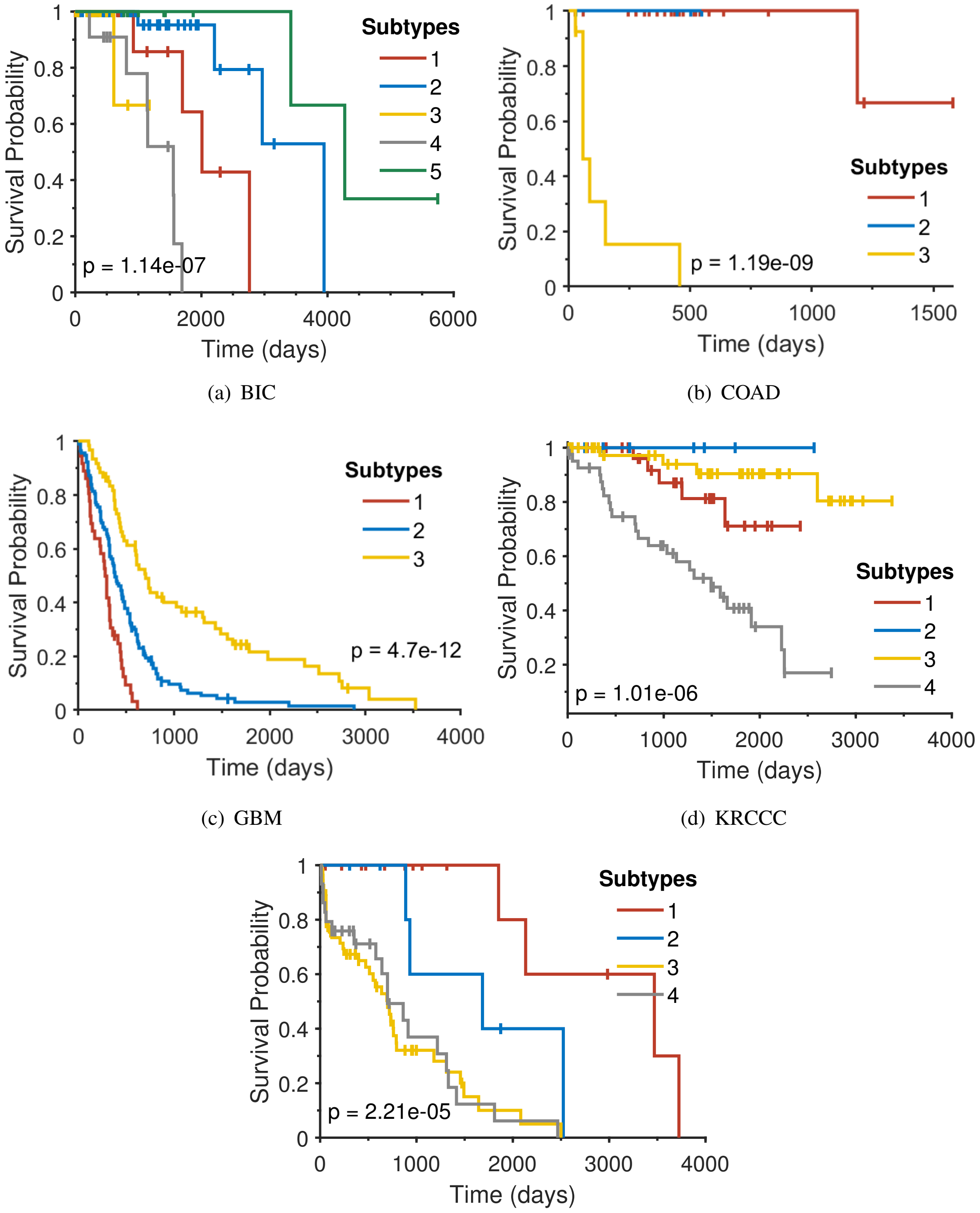
The clustering performance of different cancer subtype identification methods in terms of Cox log-rank test p-value is shown in Table 2, where the best and the second best for each dataset is shown with a dark (in bold type) and light gray background, respectively. For the competitive algorithms, the S2GC algorithm obtains the best result on dataset LSCC and the two second best results; SparseAE obtains the best result on datasets KRCCC and OV. Thus, it is observed that among the nine state-of-the-art cancer subtyping algorithms, no one can significantly outperform the others in all cases. For our AELSMC method, there are significant differences between the cancer subtypes identified, as the Cox log-rank test p-values for all datasets are quite low. Overall, the proposed AELSMC method shows very obvious superiority over other cancer subtyping algorithms, since it obtains the best result on 5 out of 8 datasets and the second best for the remaining 3 dataset LSCC. All these observations demonstrate the effectiveness of the proposed AELSMC model. It deserves noticing that AELSMC can obtain a much lower result (p-value=4.7E-12) than the others on the GBM dataset, which may be owing to the high-quality similarity graph learned via the latent embedding space.

Table 2. Clustering performance of different cancer subtype identification methods in terms of Cox log-rank test p-value.

Dataset	ANF	SRF	SNF	SM	DLSF	AutoCox	S2GC	SparseAE	AELSMC
BIC	7.6E-04	1.2E-06	1.1E-03	2.0E-04	4.0E-04	2.0E-02	3.1E-03	2.7E-05	**1.1E-07**
COAD	3.9E-02	1.9E-03	8.8E-04	7.3E-03	3.0E-01	8.1E-02	6.4E-02	4.2E-02	**1.2E-09**
GBM	1.7E-02	2.7E-04	2.0E-04	4.3E-03	1.0E-03	2.5E-02	4.0E-05	2.1E-04	**4.7E-12**
KRCCC	2.8E-02	4.6E-02	2.9E-02	2.8E-02	1.0E+00	3.0E-01	3.7E-03	**5.8E-10**	1.0E-06
LSCC	1.8E-02	2.0E-03	1.2E-02	8.4E-03	2.0E-02	6.1E-02	**1.4E-05**	4.4E-04	2.2E-05
OV	3.9E-02	5.1E-03	3.2E-03	6.7E-03	4.2E-03	1.9E-02	2.9E-04	**5.2E-06**	1.6E-05
SARC	7.8E-03	2.2E-03	1.8E-03	2.4E-03	2.1E-03	1.4E-02	2.9E-06	8.3E-08	**2.1E-08**
SKCM	3.3E-03	4.2E-02	7.1E-03	1.3E-03	1.5E-02	3.4E-01	4.4E-08	2.9E-04	**3.5E-10**

By using the Kaplan-Meier survival analysis, the survival curves of AELSMC on five multi-omics datasets are generated as shown in Figure 4. Note that Kaplan-Meier probability is a commonly used approach to discriminate the survival time of different groups. It is observed from Figure 4 that the subtypes of all the datasets can mostly be well distinguished, except for two subtypes of the COAD dataset. That is to say, in most cases there is a significant difference between the survival probability curves of various clusters identified by our AELSMC method. To take the BIC dataset as an example, among all the five subtypes, subtype 5 has the longest average survival time, followed by subtypes 1 and 2. The prognosis of subtypes 3 and 4 is relatively poor. For the other four datasets, the prognosis of different subtypes can be observed in the similar way.

Moreover, the average values of Silhouette obtained by six multi-omics clustering methods, SRF, SNF, DLSF, S2GC, SparseAE and AELSMC are presented in Table 3, where the best score is highlighted in bold. We can observe that the proposed AELSMC method achieves the best result on 5 out of all 8 multi-omics datasets, and for the remaining three multi-omics datasets, it can still achieve relatively good results.



(a) BIC

(b) COAD

(c) GBM

(d) KRCCC

(e) LSCC

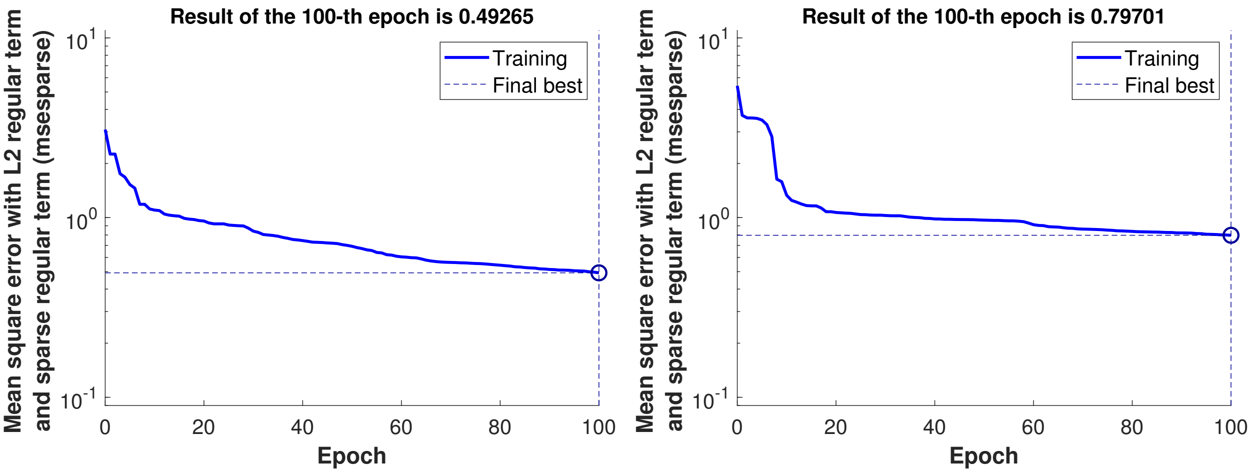
Figure 4. Kaplan-Meier survival curves of AELSMC on five multi-omics datasets.

Table 3. Comparative cluster evaluation of average Silhouette.

Dataset	SRF	SNF	DLSF	S2GC	SparseAE	AELSMC
BIC	0.593	0.237	0.529	0.558	0.614	**0.679**
COAD	0.335	0.444	0.348	0.376	0.603	**0.656**
GBM	0.429	0.491	0.372	0.763	0.747	**0.826**
KRCCC	0.399	0.425	0.724	0.755	**0.911**	0.892
LSCC	0.437	0.326	0.576	**0.704**	0.653	0.686
OV	0.636	0.611	0.647	0.686	0.708	**0.733**
SARC	0.603	0.627	0.596	0.610	**0.720**	0.704
SKCM	0.561	0.523	0.549	0.614	0.592	**0.643**

4.3. Effectiveness verification

First, we present the effectiveness verification of the AE in our proposed framework. To take dataset GBM as an example, the training process of an AE on the view of mRNA and methylation, respectively, is shown in Figure 5. The loss function of an AE is the mean square error with L2 regular term and sparse regular term (msesparse for short), and the total number of epochs is set as 100 here. It is obvious that the AE can quickly converge toward a low value of msesparse, which shows the effectiveness and efficiency of the latent space learned by the AE model. Also, the decreasing trend of both msesparses becomes slower at the late stage, such that it is not necessary to set a large value for the total number of epochs. Generally, we suggest to set the number of total epochs as less than 100, and sometimes 50 or smaller is enough. From Figure 5(a) the msesparse of epoch = 50 approximately reaches 0.75, while that of epoch = 100 is 0.49; and from Figure 5(b) the msesparse of epoch = 50 approximately reaches 1.0, while that of epoch = 100 is 0.79. Hence, there is actually no significant difference between the learned latent spaces in the cases of epoch = 50 and epoch = 100 here.

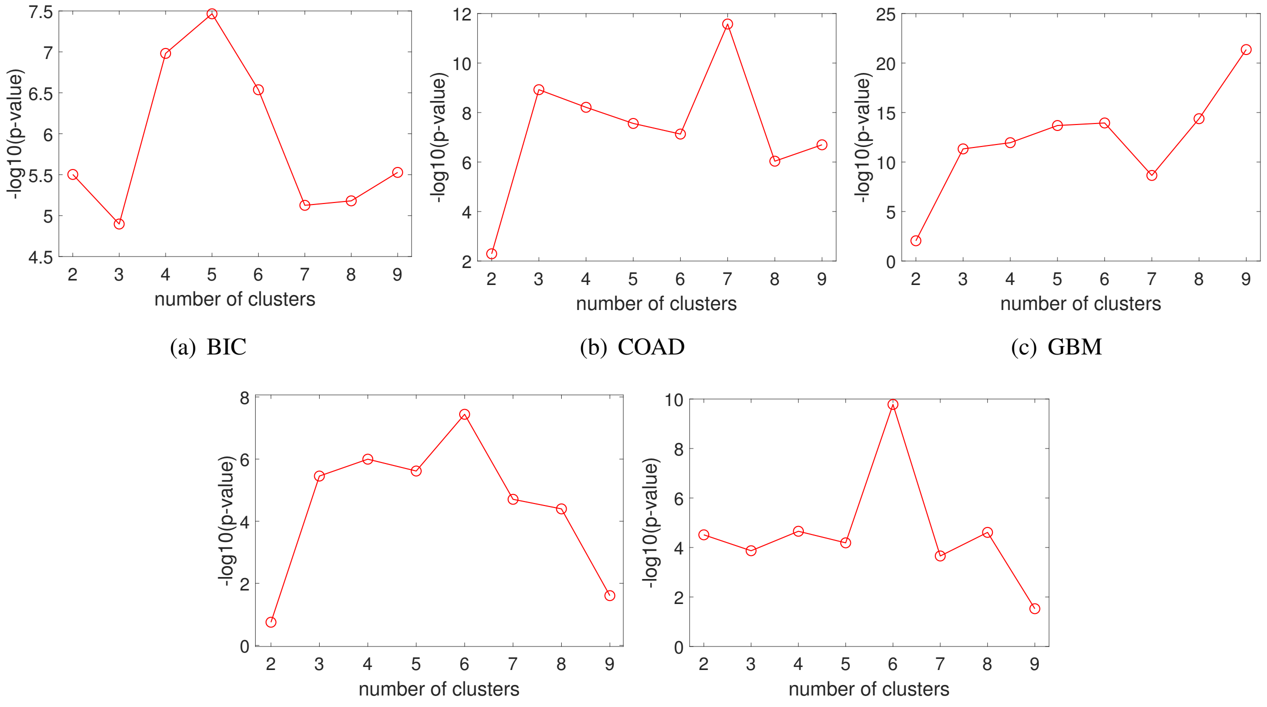


(a) mRNA

(b) methylation

Figure 5. Training process of AE on BIC dataset.

The number of clusters k was set in the interval from 2 to 9 and we conduct our AELSMC method on the five datasets, respectively, by setting the control parameter of the size for latent space as 0.05. Figure 6 shows the -log10(p-value) of the AELSMC method with varying k. We can see that an appropriate k is helpful for obtaining more significant cancer subtypes. If the number of subgroups of AELSMC is based on the value of Cox log-rank test p-value, we can generate better results than that of Table 2, which shows the promising potential of AELSMC.



(a) BIC (b) COAD (c) GBM

(d) KRCCC (e) LSCC

Figure 6. The -log10(p-value) of the AELSMC method on five multi-omics datasets with varying number of subtypes.

Table 4. Performances comparison of the AELSMC method (Integration) on multi-omic datasets with that of the framework on each single omic type and simply concatenating the encoder features (Concatenation) in terms of Cox log-rank test p-value.

Dataset	mRNA	miRNA	Methylation	Concatenation	Integration
BIC	1.7E-02	6.9E-05	5.7E-04	6.6E-03	**1.1E-07**
COAD	4.1E-02	7.1E-02	1.7E-03	4.7E-02	**1.2E-09**
GBM	7.2E-04	8.6E-07	1.9E-11	3.2E-05	**4.7E-12**
KRCCC	7.4E-03	1.8E-06	1.1E-02	3.6E-06	**1.0E-06**
LSCC	1.0E-03	6.2E-02	1.3E-04	1.8E-02	**2.2E-05**
OV	1.2E-04	2.3E-07	4.1E-06	3.3E-06	**2.9E-09**
SARC	9.5E-07	2.0E-07	3.2E-01	1.8E-06	**2.9E-08**
SKCM	3.9E-04	8.4E-08	8.5E-01	2.5E-05	**4.4E-08**

To better investigate the effectiveness of the AELSMC method, we compare the result of integrating all three omic types with that of our framework using each single omic type as well as simply concatenating the encoder features. The comparison result is shown in Table 4, where the best result in each case is shown in bold type. It is obvious that our AELSMC method of integrating multi-views obtains much better results than that of using only a single view. Moreover, the results of simply concatenating the encoder features shown in the concatenation column can obviously demonstrate the advantages of the proposed method in practice, since the p-value on each dataset is much worse than that in the integration column. Hence, we can conclude that the proposed data integration technique can significantly improve clustering performance and generate different survival profiles.

4.4. *Parameter analysis*

In this section, the analysis of parameter sensitivity is presented. To be specific, the parameter sensitivity of an AE and cancer subtyping process are investigated, respectively.

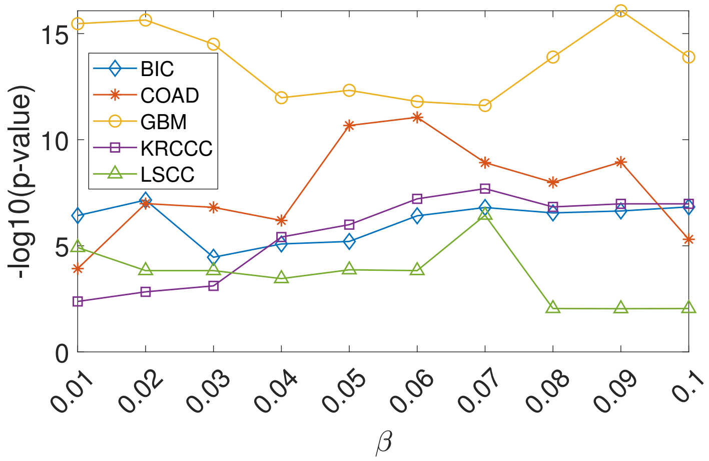


Figure 7. Performances comparison of setting different values of parameter β for AE-assisted latent representation learning on five multi-omics datasets in terms of -log10(p-value).

For the latent representation learning based on an AE, the dimension of reduced space is set to be $\beta \times d$, where β takes its value from interval $B = [0.01, 0.02, \ldots, 0.1]$. Figure 7 shows the performance of the AELSMC method on five multi-omics datasets in terms of -log10(p-value) under the given number of clusters. As we can see, neither a high value nor a low value of β is capable of obtaining desirable results in most cases. For example, the result on dataset LSCC is very poor by setting β with a higher value, while the result on datasets COAD and KRCCC is relatively poor by setting β with a lower value. It may be owing to the fact that the over-reduced latent space loses too much information of the original dataset, while a relatively high-dimensional latent space is not very beneficial for learning the similarity matrix of datasets. On the contrary, a moderate value of β (i.e., in the middle of interval B) can usually generate a desirable result, even though not the best one among the 10 cases ($|B| = 10$) for each dataset. Note that a worse result (which takes a value around 12) is obtained by setting the moderate value of β on dataset GBM. However, the performance in this case is still much better than other comparative algorithms since the second best -log10(p-value) is just 3.01 generated by S2GC from Table 2. Hence, we suggest to set β with a moderate value in interval B, e.g., $\beta = 0.05$ as the default value. However, it still needs more investigation of exploring how to set the best value of β for each multi-omics dataset.

For the survival analysis, the parameter λ needs to be set in Eq (3.5), which may play a significant impact on the final performance. In view of this, we provide the performances comparison of the AELSMC on five multi-omic datasets with different values of λ in terms of -log10(p-value), as illustrated in Figure 8, where λ takes the value from set $\{0.1, 0.5, 1, 5, 10\}$. We can observe that the results of setting $\lambda = 1$ are desirable in general, and it especially shows obvious superiority on datasets COAD and GBM by $\lambda = 1$ over that of setting other λ values. Hence, $\lambda = 1$ is suggested in our proposed method for analyzing the multi-omics datasets.

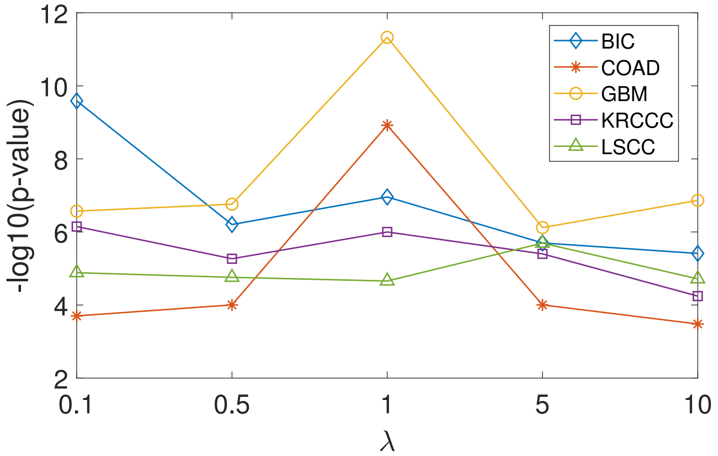


Figure 8. Performances comparison of the AELSMC method on five multi-omics datasets with different values of λ in terms of -log10(p-value).

Additionally, it is interesting to develop multi-objective clustering methods [38,39] for multi-omics cancer subtyping, since they can take advantage of multi-objective evolutionary algorithms [29, 30] to simultaneously balance multiple clustering objectives in terms of clustering quality from different perspectives (or views/omics). Meanwhile, some parameters may not be needed in the optimization model, like Eq (3.7).

5. Conclusions

To deal with multi-omics data characterised by high-dimension and heterogeneity, we propose an AE-assisted latent representation learning method for survival prediction and multi-omics clustering. The role of AEs benefits to capture nonlinear omic features in a lower-dimensional space, so as to identify the similarity of patients. Moreover, we make full use of survival information or clinical information by the means of embedding a multi-omic survival analysis approach when integrating the similarity graph of heterogeneous data at the multi-omics level. To validate its superiority, the performance of our method has been compared with other state-of-the-art algorithms on five multi-omics datasets. The experimental results reveal that the proposed AELSMC algorithm is a highly competitive method to deal with cancer subtype identification under the condition of biological big data. The promising performance of AEs encourages that more deep learning methods [51, 52] can be further investigated to advance the development of biology and medicine.

In addition, it is worth pointing out that more different omics data can be further explored via [44–46], and these datasets help discover some insightful knowledge that facilitates to deal with the tasks in clinical analysis. For example, the use of blood bioenergetics and metabolomics can be predictive biomarkers of patient response to immune checkpoint inhibitor therapy [45], which plays an important role in guiding treatment decisions and developing approaches to the treatment of therapeutic resistance. Moreover, OmicsDI [44] is an open-source platform that can be used to access, discover and disseminate omics. It can integrate proteomics, genomics, metabolomics and transcriptomics datasets, which has great potential for our further study.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Natural Science Foundation of Jiangsu Province under Grant BK20221067 and BK20230923, the Natural Science Foundation of China under Grant 62206113, and the Youth Project of Fundamental Research Funds for the Central Universities of China under Grant JUSRP123034.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. A. Conesa, S. Beck, Making multi-omics data accessible to researchers, *Sci. Data*, **6** (2019), 251. https://doi.org/10.1038/s41597-019-0258-4
2. P. S. Reel, S. Reel, E. Pearson, E. Trucco, E. Jefferson, Using machine learning approaches for multi-omics data analysis: A review, *Biotechnol. Adv.*, **49** (2021), 107739. https://doi.org/10.1016/j.biotechadv.2021.107739
3. M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, M. M. Hoffman, Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities, *Inf. Fusion*, **50** (2019), 71–91. https://doi.org/10.1016/j.inffus.2018.09.012
4. J. Lipkova, R. J. Chen, B. Chen, M. Y. Lu, M. Barbieri, D. Shao, et al., Artificial intelligence for multimodal data integration in oncology, *Cancer Cell*, **40** (2022), 1095–1110. https://doi.org/10.1016/j.ccell.2022.09.012
5. G. Cammarota, G. Ianiro, A. Ahern, C. Carbone, A. Temko, M. J. Claesson, et al., Gut microbiome, big data and machine learning to promote precision medicine for cancer, *Nat. Rev. Gastroenterol. Hepatol.*, **17** (2020), 635–648. https://doi.org/10.1038/s41575-020-0327-3
6. N. Rappoport, R. Shamir, Multi-omic and multi-view clustering algorithms: review and cancer benchmark, *Nucleic Acids Res.*, **46** (2018), 10546–10562. https://doi.org/10.1093/nar/gky889

7. T. Ma, A. Zhang, Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering, in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (2017), 398–403. https://doi.org/10.1109/BIBM.2017.8217682

8. Y. Guo, J. Zheng, X. Shang, Z. Li, A similarity regression fusion model for integrating multi-omics data to identify cancer subtypes, *Genes*, **9** (2018), 314. https://doi.org/10.3390/genes9070314

9. H. Ding, M. Sharpnack, C. Wang, K. Huang, R. Machiraju, Integrative cancer patient stratification via subspace merging, *Bioinformatics*, **35** (2019), 1653–1659. https://doi.org/10.1093/bioinformatics/bty866

10. C. Chauvel, A. Novoloaca, P. Veyre, F. Reynier, J. Becker, Evaluation of integrative clustering methods for the analysis of multi-omics data, *Briefings Bioinf.*, **21** (2020), 541–552. https://doi.org/10.1093/bib/bbz015

11. B. Pfeifer, M. G. Schimek, A hierarchical clustering and data fusion approach for disease subtype discovery, *J. Biomed. Inf.*, **113** (2021), 103636. https://doi.org/10.1016/j.jbi.2020.103636

12. G. Brière, É. Darbo, P. Thébault, R. Uricaru, Consensus clustering applied to multi-omics disease subtyping, *BMC Bioinf.*, **22** (2021), 1–29. https://doi.org/10.1186/s12859-021-04279-1

13. C. Liu, W. Cao, S. Wu, W. Shen, D. Jiang, Z. Yu, et al., Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **19** (2022), 1193–1202. https://doi.org/10.1109/TCBB.2020.3010509

14. Y. Li, J. Wang, J. Ye, C. K. Reddy, A multi-task learning formulation for survival analysis, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 1715–1724. https://doi.org/10.1145/2939672.2939857

15. H. Chai, X. Zhou, Z. Zhang, J. Rao, H. Zhao, Y. Yang, Integrating multi-omics data through deep learning for accurate cancer prognosis prediction, *Comput. Biol. Med.*, **134** (2021), 104481. https://doi.org/10.1016/j.compbiomed.2021.104481

16. C. Liu, S. Wu, D. Jiang, Z. Yu, H. S. Wong, View-aware collaborative learning for survival prediction and subgroup identification, *IEEE Trans. Biomed. Eng.*, **70** (2022), 307–317. https://doi.org/10.1109/TBME.2022.3190050

17. J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, *Inf. Fusion*, **38** (2017), 43–54. https://doi.org/10.1016/j.inffus.2017.02.007

18. Z. Huang, J. Wu, A multiview clustering method with low-rank and sparsity constraints for cancer subtyping, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **19** (2022), 3213–3223. https://doi.org/10.1109/TCBB.2021.3122917

19. Z. Chen, Z. Yang, L. Zhu, P. Gao, T. Matsubara, S. Kanaya, M. Altaf-Ul-Amin, Learning vector quantized representation for cancer subtypes identification, *Comput. Methods Programs Biomed.*, **236** (2023), 107543. https://doi.org/10.1016/j.cmpb.2023.107543

20. S. Ge, J. Liu, Y. Cheng, X. Meng, X. Wang, Multi-view spectral clustering with latent representation learning for applications on multi-omics cancer subtyping, *Briefings Bioinf.*, **24** (2023), bbac500. https://doi.org/10.1093/bib/bbac500

21. J. Zhao, B. Zhao, X. Song, C. Lyu, W. Chen, Y. Xiong, et al., Subtype-DCC: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data, *Brief. Bioinf.*, **24** (2023), bbad025. https://doi.org/10.1093/bib/bbad025

22. X. Ye, Y. Shang, T. Shi, W. Zhang, T. Sakurai, Multi-omics clustering for cancer subtyping based on latent subspace learning, *Comput. Biol. Med.*, **164** (2023), 107223. https://doi.org/10.1016/j.compbiomed.2023.107223

23. C. Zhang, Y. Chen, T. Zeng, C. Zhang, L. Chen, Deep latent space fusion for adaptive representation of heterogeneous multi-omics data, *Briefings Bioinf.*, **23** (2022), bbab600. https://doi.org/10.1093/bib/bbab600

24. L. Zong, X. Zhang, L. Zhao, H. Yu, Q. Zhao, Multi-view clustering via multi-manifold regularized non-negative matrix factorization, *Neural Networks*, **88** (2017), 74–89. https://doi.org/10.1016/j.neunet.2017.02.003

25. X. Li, H. Zhang, R. Wang, F. Nie, Multiview clustering: A scalable and parameter-free bipartite graph fusion method, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 330–344. https://doi.org/10.1109/TPAMI.2020.3011148

26. Y. Pan, C. Q. Huang, D. Wang, Multiview spectral clustering via robust subspace segmentation, *IEEE Trans. Cybern.*, **52** (2022), 2467–2476. https://doi.org/10.1109/TCYB.2020.3004220

27. H. Wang, Y. Yang, B. Liu, GMC: Graph-based multi-view clustering, *IEEE Trans. Knowl. Data Eng.*, **32** (2019), 1116–1129. https://doi.org/10.1109/TKDE.2019.2903810

28. B. B. Avants, N. J. Tustison, J. R. Stone, Similarity-driven multi-view embeddings from highdimensional biomedical data, *Nat. Comput. Sci.*, **1** (2021), 143–152. https://doi.org/10.1038/s43588-021-00029-8

29. Z. Zhao, M. Zhou, S. Liu, Iterated greedy algorithms for flow-shop scheduling problems: A tutorial, *IEEE Trans. Autom. Sci. Eng.*, **19** (2021), 1941–1959. https://doi.org/10.1109/TASE.2021.3062994

30. S. Zhu, L. Xu, E. D. Goodman, Z. Lu, A new many-objective evolutionary algorithm based on generalized pareto dominance, *IEEE Trans. Cybern.*, **52** (2022), 7776–7790. https://doi.org/10.1109/TCYB.2021.3051078

31. M. Cui, L. Li, M. Zhou, A. Abusorrah, Surrogate-assisted autoencoder-embedded evolutionary optimization algorithm to solve high-dimensional expensive problems, *IEEE Trans. Evol. Comput.*, **26** (2022), 676–689. https://doi.org/10.1109/TEVC.2021.3113923

32. M. Cui, L. Li, M. Zhou, J. Li, A. Abusorrah, K. Sedraoui, A bi-population cooperative optimization algorithm assisted by an autoencoder for medium-scale expensive problems, *IEEE/CAA J. Autom. Sin.*, **9** (2022), 1952–1966. https://doi.org/10.1109/JAS.2022.105425

33. R. Tibshirani, The lasso method for variable selection in the Cox model, *Stat. Med.*, **16** (1997), 385–395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4%3C385::AID-SIM380%3E3.0.CO;2-3

34. J. Zhang, J. Huan, Inductive multi-task learning with multiple view data, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2012), 543–551. https://doi.org/10.1145/2339530.2339617

35. F. Nie, X. Wang, M. Jordan, H. Huang, The constrained laplacian rank algorithm for graph-based clustering, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **30** (2016), 1969–1976. https://doi.org/10.1609/aaai.v30i1.10302

36. A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.*, **2** (2009), 183–202. https://doi.org/10.1137/08071654

37. X. Guo, Robust subspace segmentation by simultaneously learning data representations and their affinity matrix, in *Twenty-fourth International Joint Conference on Artificial Intelligence*, (2015), 3547–3553. https://dl.acm.org/doi/abs/10.5555/2832581.2832743

38. S. Zhu, L. Xu, E. D. Goodman, Hierarchical topology-based cluster representation for scalable evolutionary multiobjective clustering, *IEEE Trans. Cybern.*, **52** (2022), 9846–9860. https://doi.org/10.1109/TCYB.2021.3081988

39. S. Zhu, L. Xu, E. D. Goodman, Evolutionary multi-objective automatic clustering enhanced with quality metrics and ensemble strategy, *Knowledge-Based Syst.*, **188** (2020), 105018. https://doi.org/10.1016/j.knosys.2019.105018

40. A. L. Fred, A. K. Jain, Combining multiple clusterings using evidence accumulation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **27** (2005), 835–850. https://doi.org/10.1109/TPAMI.2005.113

41. A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, *J. Mach. Learn. Res.*, **3** (2002), 583–617. https://doi.org/10.1162/153244303321897735

42. D. Huang, C. D. Wang, J. H. Lai, Locally weighted ensemble clustering, *IEEE Trans. Cybern.*, **5** (2018), 1460–1473. https://doi.org/10.1109/TCYB.2017.2702343

43. S. Paul, Capturing the latent space of an autoencoder for multi-omics integration and cancer subtyping, *Comput. Biol. Med.*, **148** (2022), 105832. https://doi.org/10.1016/j.compbiomed.2022.105832

44. Y. Perez-Riverol, M. Bai, F. da Veiga Leprevost, S. Squizzato, Y. M. Park, K. Haug, et al., Discovering and linking public omics data sets using the omics discovery index, *Nat. Biotechnol.*, **35** (2017), 406–409. https://doi.org/10.1038/nbt.3790

45. P. L. Triozzi, E. R. Stirling, Q. Song, B. Westwood, M. Kooshki, M. E. Forbes, et al., Circulating immune bioenergetic, metabolic, and genetic signatures predict melanoma patients' response to anti–pd-1 immune checkpoint blockade, *Clin. Cancer Res.*, **28** (2022), 1192–1202. https://doi.org/10.1158/1078-0432.CCR-21-3114

46. A. K. Pullikuth, E. D. Routh, K. D. Zimmerman, J. Chifman, J. W. Chou, M. H. Soike, et al., Bulk and single-cell profiling of breast tumors identifies trem-1 as a dominant immune suppressive marker associated with poor outcomes, *Front. Oncol.*, **11** (2021), 734959. https://doi.org/10.3389/fonc.2021.734959

47. B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, et al., Similarity network fusion for aggregating data types on a genomic scale, *Nat. Methods*, **11** (2014), 333–337. https://doi.org/10.1038/nmeth.2810

48. H. Torkey, M. Atlam, N. El-Fishawy, H. Salem, A novel deep autoencoder based survival analysis approach for microarray dataset, *PeerJ Comput. Sci.*, **7** (2021), e492. https://doi.org/10.7717/peerj-cs.492

49. P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, **20** (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

50. S. Zhu, L. Xu, Many-objective fuzzy centroids clustering algorithm for categorical data, *Expert Syst. Appl.*, **96** (2018), 230–248. https://doi.org/10.1016/j.eswa.2017.12.013

51. Z. Lu, I. Whalen, Y. Dhebar, K. Deb, E. Goodman, W. Banzhaf, et al., Multi-objective evolutionary design of deep convolutional neural networks for image classification, *IEEE Trans. Evol. Comput.*, **25** (2020), 277–291. https://doi.org/10.1109/TEVC.2020.3024708

52. Z. Lu, G. Sreekumar, E. Goodman, W. Banzhaf, K. Deb, V. N. Boddeti, Neural architecture transfer, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2021), 2971–2989. https://doi.org/10.1109/TPAMI.2021.3052758