MCluster-VAEs: an end-to-end variational deep learning-based clustering method for subtype discovery using multi-omics data

Zhiwei Rong
Peking University
Jiali Song
Peking University
Lei Cao
Harbin Medical University
Zhilin Liu
Peking University
Yipei Yu
Peking University
Mantang Qiu
Peking University People's Hospital
Yan Hou (houyan@bjmu.edu.cn)
Peking University

Research Article

Keywords: cancer subtype discovery, multi-omics data integration, cluster, deep learning, variational bayes

Posted Date: May 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1668552/v1

MCluster-VAEs: an end-to-end variational deep learning-based clustering method for subtype discovery using multi-omics data

Zhiwei Rong[a], Jiali Song[a], Lei Cao[b], Zhilin Liu[a], Yipei Yu[a], Mantang Qiu[c,*] and Yan Hou[a,d,*]

[a] Department of Biostatistics, School of Public Health, Peking University, Beijing 100191, China

[b] Department of Epidemiology and Biostatistics, School of Public Health, Harbin Medical University, Harbin 150086, China

[c] Department of Thoracic Surgery, Peking University People's Hospital, No.11 Xizhimen South Street, Xicheng District, Beijing 100044, China

[d] Peking University Clinical Research Center, Peking University, 38 Xueyuan Rd., Haidian District, Beijing, China

[*] To whom correspondence should be addressed.

Yan Hou, Tel: +86 (10)82805952; E-mail: houyan@bjmu.edu.cn.

Mantang Qiu, Tel: +86 (10)88325983; E-mail: qiumantang@163.com.

Abstract

Background: The discovery of cancer subtype based on unsupervised clustering helps provide precise diagnoses, guide treatment and improve patients' prognoses. Instead of single-omics data, multi-omics data can improve performance of the clustering because it obtains a comprehensive landscape for understanding biological systems and mechanisms. However, heterogeneous data from multiple sources raises high complexity and different kinds of noise, which will be detrimental to the extraction of clustering information.

Methods: We propose an end-to-end deep learning-based method, Multi-omics Clustering Variational Autoencoders (MCluster-VAEs), that can extract cluster-friendly representations on multi-omics data. First, unified network architecture with an attention mechanism is developed for modeling multi-omics data precisely. Then, using a novel objective function built from the Variational Bayes technique, the model is trained to effectively obtain the posterior estimation of clustering assignments.

Results: Compared with twelve other state-of-the-art multi-omics clustering methods, MCluster-VAEs achieved outstanding performance on benchmark datasets from the TCGA database. On the Pan Cancer dataset, MCluster-VAEs achieved adjusted Rand index of around 0.78 for cancer category recognition, an increase of more than 18% compared with other methods. Furthermore, the survival analysis and clinical parameters enrichment tests on ten cancer datasets demonstrate that MCluster-VAEs delivered comparable or even better results than many typical integrative methods.

Conclusions: These results demonstrate that MCluster-VAEs is a new powerful tool

for dissecting complex multi-omics relationships and providing new insights for cancer subtype discovery.

Key words: cancer subtype discovery; multi-omics data integration; cluster; deep learning; variational bayes

1. Background

Cancer being a highly complex and heterogeneous genomics disease, showcases variability in tumor responsiveness to the therapy (1). This problem forms the basis of one of the critical areas of cancer research, i.e. the development of excellent subtype discovery methods. These methods are designed to decouple the heterogeneity of cancer, and accordingly, divide cancer patients into different groups to better understand the pathogenesis of cancer and to boost clinical treatment (2). Since the occurrence and development of cancer is related to several different biological layers and molecular systems, it is more pertinent to identify cancer subtypes based on multi-omics data rather than single-omics data (3).

The most commonly used methods to recognize cancer subtypes are clustering, such as hierarchical clustering (4), spectral clustering (5) and K-means (6). These methods, designed for single source data, cannot address the dimensional redundancy and heterogeneous information integration brought by multi-omics data. Lately, the advent of high-throughput technologies has proliferated multi-omics data, leading to the development of multi-omics clustering methods (7). These methods are based on either additional regularization controlling the problem of dimensionality (3,8,9), the fusion of similarity networks (10–14), the factorization of multiple matrices (15,16), or the simple linear probabilistic model (17,18). However, all these methods can only deal with simple hypotheses to describe molecular systems, and cannot accurately depict the complex regulation of multi-omics data. They still face significant challenges of data

complexity.

More recently, deep learning is emerging rapidly in many fields, exhibiting innovative performance in processing images (19), texts (20) and graphs data (21). Several deep learning-based methods have been developed to try to solve multi-omics clustering tasks, such as autoencoder (AE) (22,23), variational autoencoder (VAE) (24,25) and Subtype-GAN (26). They use non-linear neural networks to learn an integrated representation of multi-omics data by the unsupervised framework and then apply a traditional clustering algorithm to this representation. We call them "two-steps" methods (Figure 1A (a)), because their entire process includes two separate steps: representation learning and clustering. Due to the substantial capacity of neural networks, the integrated representations contain rich high-level information. Thus, the clustering step is expected to obtain improved performance. However, representation learning is independent of the following clustering step, and the representation from the first step is not guaranteed to be suitable for the second. For example, most representation learning approaches try to squeeze all reconstruct-friendly information into lower-dimensional representations, and these representations contain a lot of information unrelated to clustering (27). The cluster-friendly information will be overwhelmed and interfered by cluster-unfriendly information in the following clustering step. Thus, if we can integrate these two steps as one step (Figure 1A (b)), the representation learning will be guided by the clustering target, and cluster-friendly representation will be extracted to improve clustering performance. This "one-step" idea has been studied in the field of single-modal deep learning (28–30), but not in the

field of multi-omics.

In this research, we propose Multi-omics Clustering Variational Autoencoders (MCluster-VAEs), an end-to-end deep learning-based method for clustering multi-omics data. It describes multi-omics data using a new probabilistic model with a global discrete latent variable considered as the clusters. Using Variational Bayes approach (31), we derive a unified end-to-end architecture and a novel objective function, to fit the multi-omics data and infer the posterior probability of the clustering assignments. They form a "one-step" framework, so the clustering target can guide representation learning and improve clustering performance. In addition, an attention module is used to effectively incorporate multiple omics and reveal the contribution of each omics to clustering results. To evaluate the performance of MCluster-VAEs, we collected the Pan Cancer dataset with 8,314 samples and the ten cancer datasets with a total of 4,154 samples from TCGA database. The results generated by MCluster-VAEs were fully validated, which showed great potential for novel cancer subtype discovery from deep learning model.

Figure 1. Overview of MCluster-VAEs. (A) Two deep learning-based multi-omics clustering pipelines. (B) The data generative model. (C) The posterior variational inference path. (D) The architecture of MCluster-VAEs model for the experiments using mRNA expression, miRNA expression, copy number alterations (CNA) and DNA methylation.

2. Materials and methods

2.1 Probabilistic model with discrete latent variables

Let us consider some multi-omics dataset $\boldsymbol{X} = \left\{ \boldsymbol{X}^1, \ldots, \boldsymbol{X}^M \right\}$ consisting of N i.i.d.

samples, and each sample contains M kinds of omics data which can be expressed as $\mathbf{x}_i = \left\{\mathbf{x}_i^1, \ldots, \mathbf{x}_i^M\right\}$, $i = 1, ..., N$. The vector $\mathbf{x}_i^m$ represents the features of omics m of sample i. We assume that the data are generated by some random process, involving an unobserved discrete random variable $\mathbf{y}$ with C categories and M unobserved continuous random variables $\mathbf{z} = \left\{\mathbf{z}^1, \ldots, \mathbf{z}^M\right\}$. The process consists of three steps: (1) a categorical value y_i is generated from a prior distribution $p(y)$; (2) a value $\mathbf{z}_i^m$ is assumed to follow a mixture of C Gaussian distributions and generate from distribution $p_\theta(\mathbf{z}^m \mid y_i)$ corresponding to omics m; (3) a value $\mathbf{x}_i^m$ is generated from a conditional distribution $p_\theta(\mathbf{x}_i^m \mid \mathbf{z}_i^m, y_i)$. This process can be expressed by the Figure 1B and the following formula:

$$p_\theta(\mathbf{x}_i) = \int_{\mathbf{z}_i^1, \ldots, \mathbf{z}_i^M, y_i} \prod_{m=1}^{M} p_\theta\left(\mathbf{x}_i^m | \mathbf{z}_i^m, y_i\right) p_\theta\left(\mathbf{z}_i^m | y_i\right) p(y_i) \, d\mathbf{z}_i^m \, dy_i \tag{1}$$

We assume that $p_\theta\left(\mathbf{x}_i^m | \mathbf{z}_i^m, y_i\right)$ and $p_\theta\left(\mathbf{z}_i^m | y_i\right)$ come from parametric families of distributions $p_{\theta'}(\mathbf{x}|\mathbf{z}, y)$ and $p_{\theta'}(\mathbf{x}|\mathbf{z})$, and that their probability density functions (PDFs) are differentiable almost everywhere w.r.t both θ', $\mathbf{z}$ and y. The above process describes a realistic data generative model and has enough flexibility to fit the complex multi-omics data. The unobserved y_i can be considered as the cluster assignments of sample i with C categories, and the clustering task can be equivalent to inference of the unknown parameters θ and posterior distribution $p_\theta(y_i \mid \mathbf{x}_i^1, \ldots, \mathbf{x}_i^M)$.

2.2 Variational Bayes

In order to solve the above inference task, we introduce a trainable parametric

approximation $q_\phi(y_i, \mathbf{z}_i|\mathbf{x}_i)$ of posterior distribution $p_\theta(y_i, \mathbf{z}_i \mid \mathbf{x}_i)$. The marginal likelihood is composed of a sum over the marginal likelihoods of individual samples $\log p_\theta(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i)$, which can each be rewritten as:

$$\begin{aligned} \log p_\theta(\mathbf{x}_i) &= D_{KL}\left(q_\phi(y_i, \mathbf{z}_i|\mathbf{x}_i) \,\|\, p_\theta(y_i, \mathbf{z}_i|\mathbf{x}_i)\right) + L(\theta, \phi; \mathbf{x}_i) \\ &\geq L(\theta, \phi; \mathbf{x}_i) \\ &= \mathrm{E}_{q_\phi(y_i, \mathbf{z}_i|\mathbf{x}_i)}\left(\log p_\theta(\mathbf{x}_i|y_i, \mathbf{z}_i)\right) - D_{KL}\left(q_\phi(y_i, \mathbf{z}_i|\mathbf{x}_i) \,\|\, p_\theta(y_i, \mathbf{z}_i)\right) \end{aligned} \tag{2}$$

The term $L(\theta, \phi; \mathbf{x}_i)$ is called (variational) evidence lower bound (ELBO), which strictly less than the marginal likelihood which can be maximized indirectly by updating generative parameters θ and variational parameters ϕ. This approach was proposed first by Kingma and Welling (31), and has been applied in several generative models, including the famous VAE (31).

We can assume that $q_\phi(y_i, \mathbf{z}_i|\mathbf{x}_i)$ has the mean field form and can be decomposed into product form:

$$q_\phi(y_i, \mathbf{z}_i|\mathbf{x}_i) = q_\phi(y_i|\mathbf{x}_i) \prod_{m=1}^{M} q_\phi\left(\mathbf{z}_i^m|\mathbf{x}_i^m, y_i\right) \tag{3}$$

Figure 1C illustrates this inference path. Therefore, the ELBO can be derived in the following form:

$$\begin{aligned} L(\theta, \phi; \mathbf{x}_i) = & \underbrace{\sum_{m=1}^{M}\left(\mathrm{E}_{q_\phi(y_i|\mathbf{x}_i)}\left(\mathrm{E}_{q_\phi\left(\mathbf{z}_i^m|\mathbf{x}_i^m, y_i\right)}\left(\log\left(p_\theta(\mathbf{x}_i^m \mid y_i, \mathbf{z}_i^m)\right)\right)\right)\right)}_{\text{reconstruction term}} \\ & \underbrace{-\sum_{m=1}^{M}\left(\mathrm{E}_{q_\phi(y_i|\mathbf{x}_i)}\left(D_{KL}\left(q_\phi\left(\mathbf{z}_i^m|\mathbf{x}_i^m, y_i\right) \,\|\, p_\theta\left(\mathbf{z}_i^m|y_i\right)\right)\right)\right)}_{\text{conditional prior term}} \\ & \underbrace{-D_{KL}\left(q_\phi(y_i|\mathbf{x}_i) \,\|\, p(y_i)\right)}_{\text{conditional entropy term}} \end{aligned} \tag{4}$$

The detailed derivation is given in Supplementary Note S1. The first Right-Hand-Side

(RHS) term is actually a measurement of reconstruction by samples of posterior distributions. The last two terms are the Kullback-Leibler (KL) divergences of approximated posterior and prior distributions. We refer to all RHS terms as the reconstruction term, conditional prior term and conditional entropy term, respectively. Now, with some reasonable and loose assumptions for $\mathbf{z}_i^m$ and y_i, equation (4) can be used as loss function and to optimize parameters by stochastic gradient descent (SGD) with Monte Carlo gradient estimator. After the training completed, $q_\phi(y_i|\mathbf{x}_i)$, as the approximation of posterior $p_\theta(y_i|\mathbf{x}_i)$, will be used to obtain the clustering assignment of each sample.

2.3 Multi-omics Clustering Variational Autoencoders (MCluster-VAEs)

Just like VAE, we use some neural networks for $p_\theta(\mathbf{x}_i^m \mid y_i, \mathbf{z}_i^m)$, $p_\theta(\mathbf{z}_i^m|y_i)$, $q_\phi(y_i|\mathbf{x}_i)$ and $q_\phi(\mathbf{z}_i^m|\mathbf{x}_i^m, y_i)$ in equation (4). The prior over the latent variables y_i is assumed as uniform categorical distribution $p_\theta(y_i) = 1/M$. We also assume that $p_\theta(\mathbf{z}_i^m|y_i)$ and $q_\phi(\mathbf{z}_i^m|\mathbf{x}_i^m, y_i)$ are multivariate Gaussian distributions with diagonal covariance, whose distribution parameters (means and covariances) are computed from $\mathbf{x}_i^m$ and y_i with two multilayer perceptrons (MLP, fully-connected neural networks with some hidden layers), respectively. $q_\phi(y_i|\mathbf{x}_i)$ is also a MLP with M output nodes with softmax activation function to get categorical probabilities of each clustering label. $p_\theta(\mathbf{x}_i^m \mid y_i, \mathbf{z}_i^m)$ is considered as multivariate Gaussian distribution for real-valued data, whose means are calculated from $\mathbf{z}_i^m$ and y_i through an additional MLP. The whole model architecture is shown in figure 1D.

There are two expectations in the above ELBO term. The first is w.r.t discrete

distribution $q_\phi(y_i|\mathbf{x}_i)$, which can be calculated by iterating over all possible values of y_i. The second is w.r.t continuous distribution $q_\phi(\mathbf{z}_i^m|\mathbf{x}_i^m, y_i)$, which should be estimated by reparameterization sampling trick (31). To summarize all of the above, the loss function can be rewritten as the following form (the derivation is given in Supplementary Note S2):

$$
\begin{aligned}
&\text{Loss} = -\sum_{i=1}^{N} L'(\theta, \phi; \mathbf{x}_i) = L_{rec} + L_{cprior} + L_{centropy} \\
&L_{rec} = \frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{c=1}^{C} \sum_{j=1}^{d_m^x} \left(x_{ijc}^m - x_{ijc}'^m \right)^2 \pi_{ic} \\
&L_{cprior} = -\frac{1}{2N} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{c=1}^{C} \sum_{j=1}^{d_m^z} \left(\log \left(\frac{\left(\sigma_{ijc}^m\right)^2}{\left(\sigma_{ijc}'^m\right)^2} \right) - \frac{\left(\sigma_{ijc}^m\right)^2}{\left(\sigma_{ijc}'^m\right)^2} - \frac{\left(\mu_{ijc}^m - \mu_{ijc}'^m\right)^2}{\left(\sigma_{ijc}'^m\right)^2} \right) \pi_{ic} \\
&L_{centropy} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \pi_{ic} \log \pi_{ic}
\end{aligned}
\tag{5}
$$

where $L'(\theta, \phi; \mathbf{x}_i)$ represents an estimator of $L(\theta, \phi; \mathbf{x}_i)$ using Gaussian reparameterization trick. d_m^x and d_m^z are the dimensions of original features and latent continuous embeddings of the omics m, respectively. $x_{ijc}'^m$ is the reconstruction of x_{ijc}^m, which is the value of feature j of omics m of sample i if its cluster assignment is c. μ_{ijc}^m and $\mu_{ijc}'^m$ are means of $q_\phi(\mathbf{z}_i^m|\mathbf{x}_i^m, y_i = c)$ and $p_\theta(\mathbf{z}_i^m|y_i = c)$, respectively. $\left(\sigma_{ijc}^m\right)^2$ and $\left(\sigma_{ijc}'^m\right)^2$ are variances of $q_\phi(\mathbf{z}_i^m|\mathbf{x}_i^m, y_i = c)$ and $p_\theta(\mathbf{z}_i^m|y_i = c)$, respectively. π_{ic} is the value of $q_\phi(y_i = c|\mathbf{x}_i)$.

2.4 Gated attention mechanism

$q_\phi(y_i|\mathbf{x}_i)$ is the main part of MCluster-VAEs, which merges all omics data together to infer the posterior probability of clustering assignments. The most common deep learning-based approach for multi-omics information integration is the *intermediate integration* (32), which first processes each omics data using dedicated layers,

concatenates the outputs of dedicated layers and then uses further layers to integrate the features extracted from each omics data. This approach enables the most suitable dedicated layers to be used for each omics data and can hence extract more predictive features.

However, the intermediate integration approach cannot consider the contribution of multiple omics data. In order to make more effective use of multi-omics data, we add an attention mechanism into the encoder $q_\phi(y_i|\mathbf{x}_i)$. The attention module is shown in figure 1D, which is implemented by a simple gated attention mechanism. Each omics data first maps into the hidden representation with the same dimension, then the representations are weighted sum according to attention scores calculated from the representations:

$$\begin{aligned} \mathbf{h}_i &= \sum_m^M s_i^m \mathbf{h}_i^m \\ s_i^m &= \frac{1}{1+\exp\left(-W^m \mathbf{h}_i^m\right)} \end{aligned} \tag{6}$$

where $\mathbf{h}_i^m$ and s_i^m are the hidden representation and attention scores of omics m of sample i, respectively. $\mathbf{h}_i$ represents the integrated information of all types of omics and then is used to predict clustering assignments.

2.5 Conditional entropy annealing trick

The most unusual term in the ELBO is the conditional entropy term. It tries to minimize the KL divergence of y-posterior and uniform prior. It seems to be in the opposite direction of clustering. However, it is the existence of this term which makes reasonable clustering possible. Actually, this term is a Bayesian regularization term against bias of

maximum likelihood, which enables every category of y_i to have enough chance to join the training process and the model does not quickly fall into a very bad local optimum.

In practice, model training often falls into collapse mode when the equation (5) is directly considered as loss function. In this case, all clustering assignments tend to become a same value. This problem may come from the enormous contribution of large dimensional multi-omics data to the training of reconstruction term L_{rec}. To solve it, we need to increase the weight of the conditional entropy term. Therefore, we propose a conditional entropy annealing trick for experiments on real data, which multiplies the conditional entropy term $L_{centropy}$ by a large value of γ at the beginning of training, and then gradually reduces the γ during training to take into account all categories of y_i:

$$\text{Loss} = L_{rec} + L_{cprior} + \gamma L_{centropy} \tag{7}$$

The value of γ is described in Supplementary Note 3. In the following experiments, we will show that this trick is very helpful in training of MCluster-VAEs.

2.6 Gumbel softmax reparameterization

The training speed of MCluster-VAEs is relatively slow, because all categories of y_i need to be passed in model when calculating the loss function. One way to solve this problem is to use reparameterization trick on the discrete variable y_i as well. According to the techniques introduced by Jang and Maddison (33,34), we use Gumbel softmax distribution to approximate the categorical distribution to achieve reparameterization. The Gumbel softmax distribution can be sampled by the following

process:

$$y_{ic} = \frac{\exp\left(\frac{\left(\log \pi_{ic} + g_c\right)}{\tau}\right)}{\sum_{k=1}^{C} \exp\left(\frac{\left(\log \pi_{ik} + g_k\right)}{\tau}\right)} \qquad (8)$$
$$g_k = -\log\left(-\log\left(u_k\right)\right), u_k \sim U(0,1), k = 1,...,C$$

where π_{ic} is the posterior probability of category c of sample i (the value of $q_\phi(y_i = c \mid \mathbf{x}_i)$). τ is an annealing factor that decreases gradually during whole training. The pseudo-code of MCluster-VAEs with and without the gumbel softmax trick is shown in Supplementary Note S3.

2.7 Datasets and preprocessing

The mRNA expression, miRNA expression, DNA methylation (450K) and copy number alterations (short for mRNA, miRNA, methylation, CNA) from TCGA database were used in this study. MCluster-VAEs and comparison methods were tested in the following two settings.

The Pan Cancer dataset consisted of a total of 8,314 samples across 32 types of cancers. These cancer types originated from different types of tissues and dissection positions. Therefore, the Pan Cancer dataset had natural categorical structure (32 cancer types) and could be used to test the performance of clustering algorithms. We then preprocessed this dataset using a 6-steps approach: 1. to filter samples to ensure that they exist in all four omics; 2. to perform log transformation of mRNA and miRNA; 3. to remove duplicated regions data of CNA and convert it into gene-level form; 4. to select the features using Yang's approaches (26), retaining 3105 CNA features, 3217

mRNA features, 383 miRNA features and 3139 methylation features; 5. to perform missing data imputation by the samples' mean value; 6. To standardize features in each omics by removing the mean and scaling the feature to unit variance. The sample sizes and abbreviations of 32 cancer types were included in the Supplementary Table S1.

Secondly, ten cancer types in the TCGA dataset were applied in this study (including 4154 tumors): 1031 BRCA tumors, 399 BLCA tumors, 488 KIRC tumors, 127 GBM tumors, 490 LUAD tumors, 176 PAAD tumors, 446 SKCM tumors, 407 STAD tumors, 510 UCEC tumors and 80 UVM tumors. These datasets were used to test survival and clinical differences of clustering assignments, which have been used in Yang's study (26). These datasets with the above-mentioned four omics were preprocessed by the above 6-steps approach.

2.8 Evaluation criteria and comparison algorithms

Since the cancer types were known, the Pan Cancer dataset was tested to determine whether the clustering algorithms could correctly identify the labels. It was measured by unsupervised clustering accuracy (ACC), adjusted Rand index (ARI), normalized mutual information (NMI), and clustering F measure (F1). They were defined in Supplementary Note S4.

On the ten single cancer datasets, the significance of survival analysis and the number of enriched clinical parameters were used to measure performance of MCluster-VAEs and the comparison algorithms. It assumed that if clusters of samples exhibit significant difference in clinical parameters and survival, they are different in a biologically meaningful way. The clinical parameters included gender, age at diagnosis, pathologic

T, pathologic M, pathologic N and pathologic stage. The four latter parameters were discrete pathological parameters, measuring the progression of the tumor (T), metastases (M) and cancer in lymph nodes (N), and the total progression (stage). We measured significance of survival analysis among the obtained clustering assignments using the log-rank test (35), enrichment for discrete parameters using the chi-square test, and enrichment for numeric parameters using Kruskal-Wallis test. According to Rappoport and Shamir (36), the accuracy of pure log-rank test, chi-square test and Kruskal-Wallis test might be influenced by small sample size and unbalanced cluster assignments. Therefore, we followed their advice and used the corresponding permutation tests to estimate empirical *P* values. More details on the permutation tests can be found in their study (36). After that, the *P* values were corrected for multiple hypotheses using Bonferroni correction for each cancer and corresponding method. Note that cancer subtypes that are biologically different may have similar survival, and this is also true for enrichment of clinical parameters. However, these measures are widely used for clustering assessment, including in the most multi-omics clustering studies (12,26,36–38).

To evaluate the clustering performance of MCluster-VAEs, we compared its performance with the performances of twelve state-of-the-art clustering methods (i.e. *k-means* (6), *spectral clustering* (5), *MCCA* (15), *SNF* (10,11), *COCA* (39), *ANF* (40), *iClusterBayes* (3), *CIMLR* (41), NEMO (12), *MAUI*(*VAE*) (24,25), *DCAP*(*AE*) (22,23), *SubtypeGAN* (26)). MUAI and DCAP used VAE and AE with *intermediate integration*, respectively. Hence, MUAI(VAE) and DCAP(AE) were used to represent them. These

methods were chosen to represent diverse approaches to multi-omics clustering. The detailed information and the parameter setting of all methods were shown in Supplementary Note S5, S6, Table S2 and S3.

3. Results

3.1 Comparison of MCluster-VAEs with state-of-the-art clustering methods on the Pan Cancer dataset

Table 1. Clustering performance on the Pan Cancer dataset using MCluster-VAEs and twelve other methods.

Method	ACC	ARI	F1	NMI
iCluster	0.0569±0.0022	0.0014±0.0004	0.0386±0.0004	0.0231±0.0012
MCCA	0.1275±0.0026	0.0557±0.0032	0.1067±0.0053	0.1320±0.0031
SNF	0.4809	0.3223	0.3659	0.6050
Spectral	0.5635	0.4423	0.4690	0.6857
COCA	0.5459±0.0099	0.5051±0.0030	0.5264±0.0029	0.6732±0.0042
DCAP(AE)	0.6656±0.0256	0.5267±0.0347	0.5482±0.0324	0.7725±0.0112
SubtypeGAN	0.6334±0.0374	0.5335±0.0505	0.5526±0.0484	0.7311±0.0265
CIMLR	0.6342±0.0100	0.5355±0.0209	0.5561±0.0197	0.7337±0.0081
MAUI(VAE)	0.6750±0.0241	0.5380±0.0528	0.5587±0.0494	0.7669±0.0079
K-means	0.6507±0.0308	0.5738±0.0275	0.5922±0.0264	0.7494±0.0124
ANF	0.6924±0.0672	0.6134±0.0991	0.6314±0.0965	0.7944±0.0843
NEMO	0.7475	0.6642	0.6782	0.7966
MCluster-VAEs	**0.8200±0.0152**	**0.7826±0.0291**	**0.7924±0.0279**	**0.8789±0.0088**

The proposed MCluster-VAEs and the state of art methods were applied to the Pan Cancer dataset labeled with known cancer types, with clusters set equal to the number of cancer types (i.e., 32). The clustering performance (ACC, ARI, F1, and NMI) was shown in Table 1. Except for SNF, Spectral, and NEMO, which produced deterministic results, all methods were repeated five times to avoid the influence of randomness. The results obtained by MCluster-VAEs were 0.8200 (ACC), 0.7826 (ARI), 0.7924 (F1),

and 0.8789 (NMI), which showed the best performance compared with the other methods. Even when compared to the second-placed NEMO, there was about 10-18% improvement. Moreover, MCluster-VAEs showed more minor variation than other deep learning-based methods (DCAP(AE), MAUI(VAE), SubtypeGAN) and were at the same level as k-means. The correspondence between clustering assignments and cancer types labels was shown in Supplementary Figures S1 and S2, demonstrating that the performance improvement of MCluster-VAEs is reflected in almost all cancer types, not just in individual cancer types. These findings indicated that the performance of MCluster-VAEs was better than those of the twelve state-of-the-art clustering methods on the Pan Cancer dataset.

Figure 2. The t-SNE visualization on the latent variables generated by four deep learning-based methods (MCluster-VAEs, AE, VAE, SubtypeGAN) used the Pan Cancer dataset. The color of each point indicated the cancers. MCluster-VAEs had a greater degree of separation.

As mentioned earlier, the excellent performance of MCluster-VAEs may come from its "one-step" framework, which guides the model to learn cluster-friendly representation. In order to prove it, we visualized the representations learned of MCluster-VAEs, DCAP(AE), MAUI(VAE), and SubtypeGAN through t-SNE (42). For MCluster-VAEs, the representations were the output of the second last layer of encoder $q_\phi(y_i|\mathbf{x}_i)$. For DCAP(AE), MAUI(VAE), and SubtypeGAN, the representations were the outputs of the encoder. As shown in Figure 2, the representations of MCluster-VAEs had more remarkable dissociation among different cancer types. For example, although data points representing HNSC, LUSC, CESC, and BLCA (red circle) tended to mix in subfigures of DCAP(AE), MAUI(VAE), and SubtypeGAN, MCluster-VAEs had a

greater degree of separation. In addition, the points representing STAD tended to mix with the points representing COAD for these three “two-steps” methods (light blue circle). The confusion could be caused by the fact that these two kinds of cancers are all digestive tract cancer. However, the representations learned by MCluster-VAEs separated STAD from COAD. MCluster-VAEs also made some improvements in LIHC and CHOL (green circle). The above results suggested that MCluster-VAEs could learn cluster-friendly representations and better clusters than “two-step” methods.

Figure 3. Distribution of metrics and attention scores of MCluster-VAEs using single omics data of the Pan Cancer datasets. **A.** ACC, ARI, NMI and F1 of MCluster-VAEs based on all four omics (MOmics) or single-omics data. **B.** The distribution of attention scores for each omics data. Here, mRNA represents mRNA expression, methy denotes DNA methylation (450K), miRNA represents miRNA expression and CNA represents copy number alterations.

To determine the necessity of usage of multi-omics data, we compared the clustering results of MCluster-VAEs based on four single-omics data and multi-omics data.MCluster-VAEs removed the attention module when processing single-omics data because the fusion of multiple omics features was no longer required, while other parameter settings remained unchanged. As shown in Figure 3A, CNA alone was insufficient for accurate cancer classification. In comparison to methylation, mRNA, and miRNA data alone, the combined use of multi-omics data increased ARI of approximately 18%, 19%, and 33%, respectively. In addition, the ARI of the multi-omics data had a smaller distribution range than the ARI values of single-omics data. ACC, F1, and NMI's conclusions are consistent with ARI's. These results indicate that MCluster-VAEs can give more accurate and stable clustering results using multi-omics information compared with single-omics datasets.

3.2 The role and influence of attentional mechanism

How to integrate multiple source data is a major challenge in multi-omics studies. Deep learning-based methods have an inherent advantage, because the neural networks can contain multiple independent layers to extract appropriate information for each omics, and the shared layer then fuse these information. This architecture, called *intermediate integration*, was easy to implement and has been used in many studies (22,25,26). However, different omics data contained varying amounts of clustering information, making it difficult for the *intermediate integration* architecture to identify these differences and lack interpretation of omics contribution. The solution provided by this study was the attention mechanism, which adaptively learned the weight of each omics data for each sample.

The distribution of the attention scores of each omics was shown in Figure 3B. It indicated that when clustering Pan Cancer dataset, MCluster-VAEs place a greater emphasis on mRNA, miRNA, and methylation while giving CNA less weight. It is worth noting that the distribution of attention scores matched that of clustering metrics on single omics data (Figure 3A) and that the metrics can be used to measure the amount of clustering information in each omics data to some extent. This similarity showed that the attention module allowed the model to focus more on the omics with more clustering information, allowing it to extract information more efficiently. We also implemented the non-attention version of MCluster-VAEs on the Pan Cancer dataset, compared with the standard version (with attention mechanism) of MCluster-VAEs. The results were shown in Supplementary Figure S3 and indicated that the performance

of the attention version was slightly better than the non-attention version (ARI: 0.7647±0.0257 vs. 0.7523±0.0188, ACC: 0.7978±0.0154 vs. 0.7704±0.0167). Thus, using an attention module does not affect clustering performance. Simultaneously, the attention module improved the interpretability of the clustering results, which helped investigate the biological significance of cancer subtypes. For example, methylation has a more negligible effect on PRAD and BRCA than on other cancers, whereas CNA has a more significant effect on OV (Supplementary Figure S4). In conclusion, the above results showed that the attention mechanism could help with more explainable and accurate multi-omics data integration.

3.3 Gumbel Softmax and conditional entropy weighted annealing

The exact form of MCluster-VAEs was time-consuming, especially when the number of clustering was huge (such as 32 of the Pan Cancer dataset). The reason was that MCluster-VAEs must traverse all possible clustering assignments and run $q_\phi\left(\mathbf{z}_i^m|\mathbf{x}_i^m, y_i\right)$ and $p_\theta\left(\mathbf{z}_i^m|y_i\right)$ repeatedly. The application of the Gumbel Softmax reparameterization trick would make the model just run one or few times to calculate the loss function. The running time of the Gumbel version and exact version was reported in Supplementary Figure S5, and we could see that the Gumbel version was more than 5 times faster than the exact version in the Pan Cancer dataset. The Gumbel version also had a better performance than the exact version (Figure 4). This improvement of ACC might be because this reparameterization trick improved gradient variance and gave more space for exploration.

Another challenge of multi-omics data analysis was the curse of dimensionality, which

also appeared in single-omics analysis but would be more severe in multi-omics analysis because of the integration of multiple data sources. One of its influences for MCluster-VAEs was that the effect of the reconstruction term overwhelmed the KL divergence regularization. As a result, it quickly led $q_\phi(y_i \mid \mathbf{x}_i)$ to converge into a unique clustering assignment during the early stage of training. This is a "Lazy behavior" of model training, resulting in convergence to a bad local minimum.

Figure 4. The monitoring records of the training process of MCluster-VAEs on the Pan Cancer dataset. "WA" means using conditional entropy weight decay annealing trick. "Gumbel" means using Gumbel Softmax reparameterization trick. "none" means neither used.

To solve this problem, we proposed the conditional entropy weight annealing trick. The γ factor was used to provide obvious gradient signals for each clustering category during backpropagation learning. Figure 4 showed the entropy, conditional entropy and ACC during the whole training period in the Pan Cancer dataset. When the weight annealing was not applied, conditional entropy would decline sharply to zero in first 10-20 epochs. Correspondingly, the entropy was soon stop at a lower level until training end. When using the weight annealing trick, conditional entropy would still decline to zero but more slowly, the entropy would rise to a higher level, and the ACC would also rise up to a higher value. Unexpectedly, the use of Gumbel softmax reparameterization also eased the convergence problem, which may be due to the reparameterization that improved efficiency of gradient backpropagation. Overall, by using the conditional entropy weight annealing trick and Gumbel Softmax trick, the convergence problem of MCluster-VAEs could be solved effectively.

3.4 Integrative clustering across ten cancer types with thirteen methods

We applied MCluster-VAEs and the other twelve state-of-the-art clustering methods on the ten TCGA multi-omics datasets. The clustering assignments and clinical parameters were used to calculate corrected empirical *P* values of survival analysis and enrichment test (36). To ensure the fairness of comparison, we followed the setup of Yang's study (26) and used the reasonable number of clusters obtained from the previous studies (2,43–48) as the setting for all methods (BRCA: 5, BLCA: 5, KIRC: 4, SKCM: 4, UCEC: 4, UVM: 4, GBM: 3, LUAD: 3, STAD: 3, PAAD: 2).

Figure 5. Performance of the algorithms on the ten cancer datasets. The x-axis was the number of clinical parameters enriched in the clusters, and the y-axis measured the differential survival between clusters (-log10 of permutated logrank's test *P* values). Colors indicated clustering methods.

The -log10 *P*-values of differential survival and the number of enriched clinical parameters of all methods on the ten datasets were shown in Figure 5 (the values were further reported detailly in Supplementary Tables S4 and S5). On all nine datasets except LUSC, MCluster-VAEs had the smallest log-rank *P* value and the most clinical parameters that were enriched. Even on the LUSC dataset, MCluster-VAEs still had the best survival analysis results and the second-highest number of clinical variables after SubtypeGAN. The mean of -log10 log-rank *P* values was 5.850 ($-\log_{10} 0.05 = 1.301$) obtained by MCluster-VAEs, while the second-best method (SubtypeGAN) achieved 3.511 and the third-best method (SNF) achieved 3.485. Especially on the KIRC and SKCM datasets, the -log10 *P* values obtained by MCluster-VAEs were bigger than 10 ($P < 1\times10^{-10}$). Figure 6 showed all the Kaplan-Meier survival plots of MCluster-VAEs, indicating that MCluster-VAEs can reveal cancer datasets' survival differences.

Figure 6. Kaplan-Meier survival plots of MCluster-VAEs on the ten cancer datasets.

For the number of enriched clinical parameters, the MCluster-VAEs achieved the mean of 4.0, while the second-best method (SubtypeGAN) achieved 3.4 and the third-best method (SNF) achieved 2.9. For example, on the STAD datasets, four clinical parameters' enrichment tests of MCluster-VAEs were significant, followed by three significant results of COCA. On the PAAD dataset, MCluster-VAEs had four enriched clinical parameters, whereas the other methods had no more than one. MCluster-VAEs also obtained better results than the comparison methods on the other datasets. Therefore, we can conclude that our method outperformed most of the typical methods on many TCGA datasets.

The MCluster-VAEs was also applied to single-omics data of the ten single cancer datasets. The results are shown in Figure 7. On five datasets (PAAD, UVM, GBM, STAD, UCEC), MCluster-VAEs had the smallest log-rank *P* value, while the remaining five datasets ranked second. The means of -log10 *P* values of survival analysis obtained by MCluster-VAEs were 3.845 (methylation-alone), 2.601 (mRNA-alone), 2.367 (miRNA-alone) and 1.407 (CNA-alone). They were worse than that of multi-omics data (5.850). In addition, MCluster-VAEs had the most clinical parameters which were enriched on nine datasets (except UVM). The means were 4.00 (multi-omics), 2.80 (methylation-alone), 2.78 (miRNA-alone), 1.6 (CNA-alone) and 1.3 (mRNA-alone). These results again proved that MCluster-VAEs could effectively use multiple information from multi-omics data. Combining the results of the Pan Cancer dataset, we had reason to believe that MCluster-VAEs can achieve better integration of multi-

omics data for clustering tasks.

Figure 7. Performance of MCluster-VAEs based on four omics or single-omics data on the ten cancer datasets. The x-axis was the number of clinical parameters enriched in the clusters, and the y-axis measured the differential survival between clusters (-log10 of permutated logrank's test P values). Colors indicated the omics data applied. Here, MOmics represents four omics data, mRNA represents mRNA expression, methy denotes DNA methylation (450K), miRNA represents miRNA expression and CNA represents copy number alterations.

3.5 Identification of marker genes of BRCA dataset

It is crucial to demonstrate that the subtypes identified by MCluster-VAEs are biologically interpretable as an application in medicine. To identify essential genes among the five subtypes of breast carcinoma (BRCA) identified by MCluster-VAEs (i.e., C1, C2, C3, C4, and C5), we calculated the signal-to-noise ratio (SNR) of each gene, proposing a one-vs-rest differential procedure for one subtype as one group with the other four subtypes as another group. The gene with a high absolute value of SNR indicates significant differential expression in its subtype. We selected ten genes with the highest SNR and ten genes with the lowest SNR for each BRCA subtype, such that a total of one hundred marker genes were filtered and listed in Supplementary Table S6. To interpret the biological role and potential functions of the marker genes identified by MCluster-VAEs, functional enrichment analysis of the marker genes was conducted with a gene list of GO (Gene Ontology) Molecular Functions, GO Biological Processes and GO Cellular Components in Metascape (https://metascape.org). All genes in the genome have been used as the enrichment background. Terms with a P-value <0.01, a minimum overlap of 3, and an enrichment factor >1.5 are collected and grouped into clusters based on their membership similarities. The results of GO analyses are

displayed in Figure 8A and 8B. In addition, the top 18 biological process items are listed in Supplementary Table S7.

Biological processes include epithelial cell differentiation, RAGE receptor binding, cell maturation, female sex differentiation, lipid binding, multicellular organismal homeostasis, cellular process involved in reproduction in multicellular organism, negative regulation of cell population proliferation and growth factor activity were significantly regulated by these genes. A protein–protein interaction (PPI) network was created with STRING, BioGrid, OmniPath, InWeb_IM. Only physical interactions in STRING (physical score > 0.132) and BioGrid are used. Moreover, significant modules of PPI were identified in Figures 8C. In addition, we utilized the molecular complex detection (MCODE) algorithm to analyze clusters of the PPI networks. The MCODE component extracted was mainly associated with RAGE receptor binding, sequestering of metal ion and epithelial cell differentiation (Figure 8 D–E). We also found that the molecular subtyping results of MCluster-VAEs are partially consistent with the previous analysis results obtained by gene expression data (Supplementary Table S8), confirming the ability of MCluster-VAEs to acquire reasonable cancer subtypes from the original multi-omics data without cancer-related prior information.

Figure 8. Enrichment analysis of the marker genes identified by MCluster-VAEs (Metascape). **A.** Pathway and process enrichment analysis has been carried out with Gene Ontology (GO) Biological Processes. Bar graph of enriched terms across the marker genes, colored by P-values; **B.** Network of enriched terms across the marker genes: colored by P-value, where terms containing more genes tend to have a more significant P-value; **C.** Protein–protein interaction network and the MCODE component identified in the marker genes; **D–E.** Pathway and process enrichment analysis has been applied to the MCODE component (**D**), and the three best-scoring terms by P-value have been retained as the functional description of the MCODE components, shown in the table (**E**).

4. Discussion

We presented a brand-new deep learning-based multi-omics clustering algorithm, called MCluster-VAEs. As far as we know, MCluster-VAEs is the first “one-step” deep learning-based clustering algorithm for multi-omics data. The experiments on the Pan Cancer dataset proved that MCluster-VAEs could better identify the intrinsic categorical information than other comparison methods, while the results on the ten cancer datasets showed that it could find biologically significant clusters.

The main insight of MCluster-VAEs is that it is better to use only cluster-friendly information than all information in multi-omics data. We believe that MCluster-VAEs’ better performance than “two-steps” deep learning-based methods stems considerably from this insight. Therefore, it is important that the probabilistic model contains both global categorical latent variable y and omic-special latent variables $\mathbf{z}^m$. $\mathbf{z}^m$ will be used to bear cluster-unfriendly information, which makes y focus on the cluster-friendly information. The attention module further enhances this ability, enabling the model to make trade-offs and selections among multiple omics data. The main differences between MCluster-VAEs and the "one-step" single-modal clustering methods are the omics-special $\mathbf{z}^m$ and attention module, which makes MCluster-VAEs more suitable for multi-omics data.

In the result section, the number of clusters k of all methods is set to a fair value obtained from known labels or previous large-scale studies for each tumor type. However, determining the optimal k is also a challenging task. For ten cancer types,

we explored a simple strategy to determine the most appropriate k based on MCluster-VAEs: (1) choose a possible k as the MCluster-VAEs setting and fit model; (2) extract the fusion hidden representation from the attention module's output hidden layer; (3) calculate silhouette score using this representation, and cluster assignments; (4) perform (1)-(3) on the all possible k and choose the k with the largest silhouette score. This strategy was applied to 10 cancer datasets, and the results were shown in Supplementary Figure S7. We found that MCluster-VAEs determined a reasonable k for all datasets. On six of the ten datasets (STAD, PAAD, LUAD, SKCM, GBM, UVM), the k values obtained by MCluster-VAEs are precisely equal to the values used in the benchmark. For other datasets, MCluster-VAEs obtained relatively conservative (i.e., fewer) k. Thsus, we believe that even when the number of clusters is unknown, the subtype results obtained by this strategy based on MCluster-VAEs and silhouette score are reliable.

There are still some limitations of the MCluster-VAEs. The first limitation is that the attention module is a little crude and cannot do gene-level evaluation. The second limitation is the instability in small sample size data, which could be seen in PAAD, GBM, and UVM datasets. Even if using the weight annealing trick and Gumbel softmax reparameterization, there is still a slight chance to train into the collapse. For this problem, our experience is that the entropy and conditional entropy should be monitored. The training must guarantee the gradual rise of the entropy and gradual fall of the conditional entropy by controlling the hyperparameters (such as learning rate, the weight of conditional entropy items, etc.). Simultaneously, the entropy finally needs to

be as large as enough. When the above conditions are satisfied, MCluster-VAEs will perform well in most cases.

5. Conclusions

Clustering cancer patients into subgroups has the potential to be used for personalized diagnosis and therapy. The increasing diversity of omics data, as well as their reduced cost, creates an opportunity to use multi-omics data to discover such subgroups. MCluster-VAEs' ability to handle complex relationships on multi-omics data makes it a valuable method. Its importance will become apparent with the further exploration of omics data.

6. List of abbreviations

- MCluster-VAEs: Multi-omics Clustering Variational Autoencoders
- AE: autoencoder
- VAE: variational autoencoder
- PDFs: probability density functions
- ELBO: evidence lower bound
- KL: Kullback-Leibler
- RHS: Right-Hand-Side
- SGD: stochastic gradient descent
- MLP: multilayer perceptrons
- ACC: unsupervised clustering accuracy

- ARI: adjusted Rand index
- NMI: normalized mutual information
- F1: clustering F measure
- GO: Gene Ontology
- PPI: protein-protein interaction
- The abbreviations of each cancer types are shown in Supplementary Table S1

7. Declarations

7.1 Ethics approval and consent to participate

Not applicable.

7.2 Consent for publication

Not applicable.

7.3 Availability of data and materials

The datasets used during the current study are available in the UCSC Xena data portal, https://xenabrowser.net/datapages/.

Code has been made publicly available on GitHub https://github.com/luyiyun/MCluster-VAEs.

7.4 Competing interests

The authors declare that they have no competing interests.

7.5 Funding

This work was supported by the National Natural Science Foundation of China [81773550, 82173615].

7.6 Authors' contributions

ZR proposed the main idea and implemented the codes of MCluster-VAEs, and was a major contributor in writing the manuscript. JS and LC collected and analyzed the benchmark datasets used in current research. ZL tested the comparison methods in the benchmark datasets. YY corrected the language of the manuscript. YH were the funder of the current research. All authors read and approved the final manuscript.

7.7 Acknowledgements

Not applicable.

8. References

1. Pavía-Jiménez A, Tcheuyap VT, Brugarolas J. Establishing a human renal cell carcinoma tumorgraft platform for preclinical drug testing. Nature Protocols. 2014;9(8):1848–59.

2. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell. 2018 Apr 9;33(4):690-705.e9.

3. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics. 2018 Jan 1;19(1):71–86.

4. Bridges CC. Hierarchical Cluster Analysis. Psychol Rep. 1966;18(3):851–4.

5. Shi J, Malik J. Normalized cuts and image segmentation. 2000.

6. MacQueen J, others. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Oakland, CA, USA; 1967. p. 281–97.

7. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nature Communications. 2021;12(1):124.

8. Wu D, Wang D, Zhang MQ, Gu J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genomics. 2015;16(1):1022.

9. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12.

10. Wang B, Jiang J, Wang W, Zhou ZH, Tu Z. Unsupervised metric fusion by cross diffusion. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012. p. 2997–3004.

11. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods. 2014;11(3):333–7.

12. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics. 2019 Sep 15;35(18):3348–56.

13. Zhu J, Oh JH, Deasy J, Tannenbaum A. vWCluster: A Network Based Clustering of Multi-omics Breast Cancer Data Based on Vector-Valued Optimal Transport. 2021.

14. Lemsara A, Ouadfel S, Fröhlich H. PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. BMC Bioinformatics. 2020 Apr 16;21(1):146.

15. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol. 2009;8(1):Article28.

16. Lock EF, Hoadley KA, Marron JS, Nobel AB. JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. Ann Appl Stat. 2013 Mar 1;7(1):523–42.

17. Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013;29(20):2610–6.

18. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012;28(24):3290–7.

19. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In 2016 [cited 2022 May 7]. p. 770–8. Available from: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional

Transformers for Language Understanding. arXiv e-prints. 2018;arXiv:1810.04805.

21. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907. 2016;

22. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. 2018;24(6):1248.

23. Chai H, Zhou X, Zhang Z, Rao J, Zhao H, Yang Y. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Comput Biol Med. 2021 Jul;134:104481.

24. Ronen J, Hayat S, Akalin A. Evaluation of colorectal cancer subtypes and cell lines using deep learning. Life Sci Alliance. 2019 Dec 1;2(6):e201900517.

25. Hira MT, Razzaque MA, Angione C, Scrivens J, Sawan S, Sarker M. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. Scientific Reports. 2021;11(1):6265.

26. Yang H, Chen R, Li D, Wang Z. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. Bioinformatics. 2021 Feb 18;

27. Yang B, Fu X, Sidiropoulos ND, Hong M. Towards K-Means-Friendly Spaces: Simultaneous Deep Learning and Clustering. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. JMLR.org; 2017. p. 3861–70. (ICML'17).

28. Min E, Guo X, Liu Q, Zhang G, Cui J, Long J. A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture. IEEE Access. 2018;6:39501–14.

29. Diallo B, Hu J, Li T, Khan GA, Liang X, Zhao Y. Deep embedding clustering based on contractive autoencoder. Neurocomputing. 2021;433:96–107.

30. Guo X, Gao L, Liu X, Yin J. Improved Deep Embedded Clustering with Local Structure Preservation. 2017;1753–9.

31. Kingma DP, Welling M. Auto-encoding variational bayes [J]. 2013;

32. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics. 2019;1.

33. Jang E, Gu S, Poole B. Categorical Reparameterization with Gumbel-Softmax. 2016;arXiv:1611.01144.

34. Maddison C, Mnih A, Teh Y. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. 2016 02;

35. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966 Mar;50(3):163–70.

36. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Research. 2018;46(20):10546–62.

37. Zhang C, Chen Y, Zeng T, Zhang C, Chen L. Deep latent space fusion for adaptive

representation of heterogeneous multi-omics data. Briefings in Bioinformatics. 2022 Mar 1;23(2):bbab600.

38. Yang Y, Tian S, Qiu Y, Zhao P, Zou Q. MDICC: novel method for multi-omics data integration and cancer subtype identification. Briefings in Bioinformatics. 2022 Apr 18;bbac132.

39. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014 Aug 14;158(4):929–44.

40. Ma T, Zhang A. Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017. p. 398–403.

41. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nature Communications. 2018 Oct 26;9(1):4453.

42. Maaten L van der, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9(86):2579–605.

43. Genomic Classification of Cutaneous Melanoma. Cell. 2015 Jun 18;161(7):1681–96.

44. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. Nature.

2014;511(7511):543–50.

45. Creighton CJ, Morgan M, Gunaratne PH, Wheeler DA, Gibbs RA, Gordon Robertson A, et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499(7456):43–9.

46. Levine DA, Getz G, Gabriel SB, Cibulskis K, Lander E, Sivachenko A, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497(7447):67–73.

47. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell. 2017 Oct 19;171(3):540-556.e25.

48. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010 Jan 19;17(1):98–110.

Figures

Figure 1

Overview of MCluster-VAEs. (A) Two deep learning-based multi-omics clustering pipelines. (B) The data generative model. (C) The posterior variational inference path. (D) The architecture of MCluster-VAEs model for the experiments using mRNA expression, miRNA expression, copy number alterations (CNA) and DNA methylation.

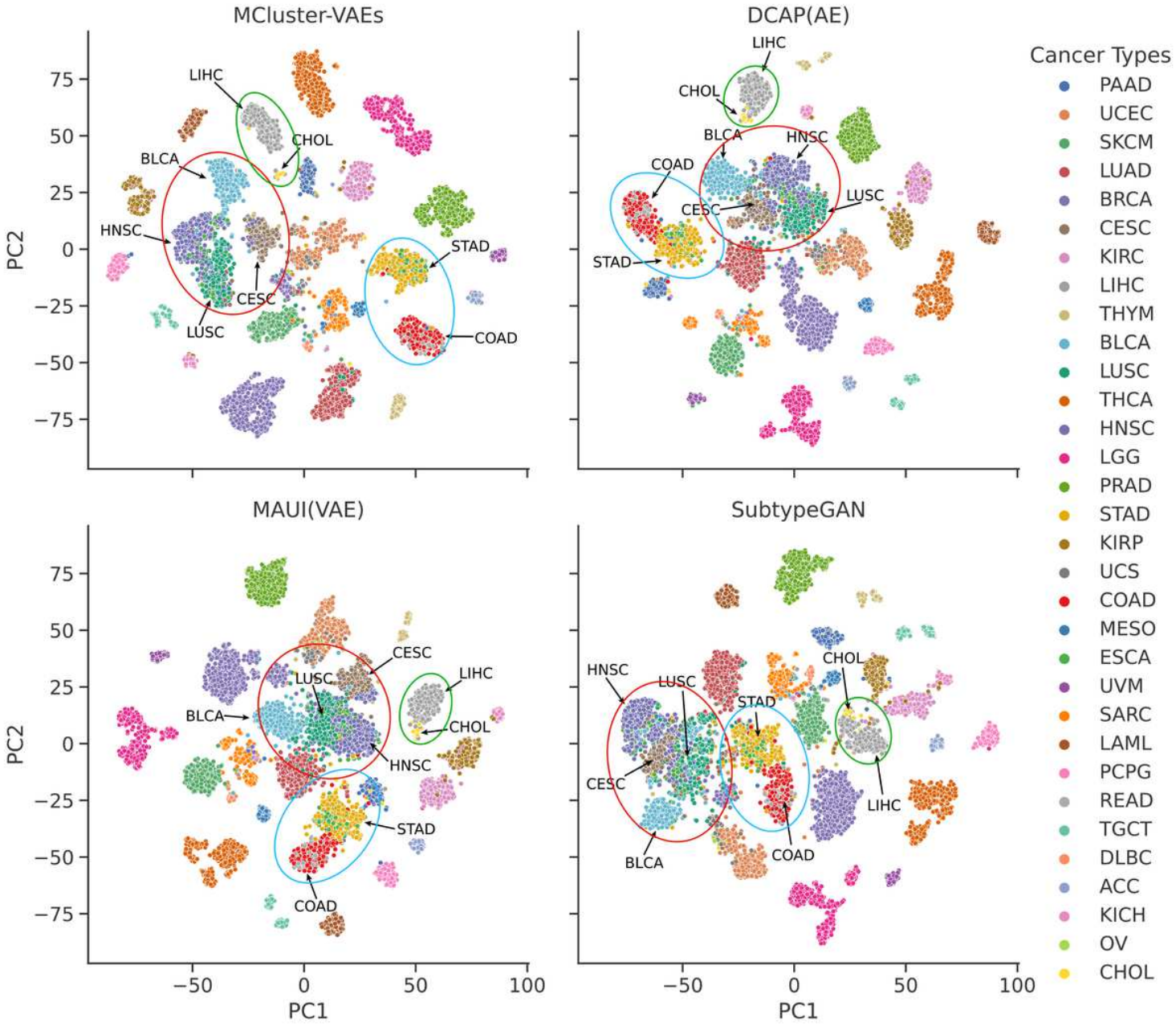

Figure 2

The t-SNE visualization on the latent variables generated by four deep learning-based methods (MCluster-VAEs, AE, VAE, SubtypeGAN) used the Pan Cancer dataset. The color of each point indicated the cancers. MCluster-VAEs had a greater degree of separation.

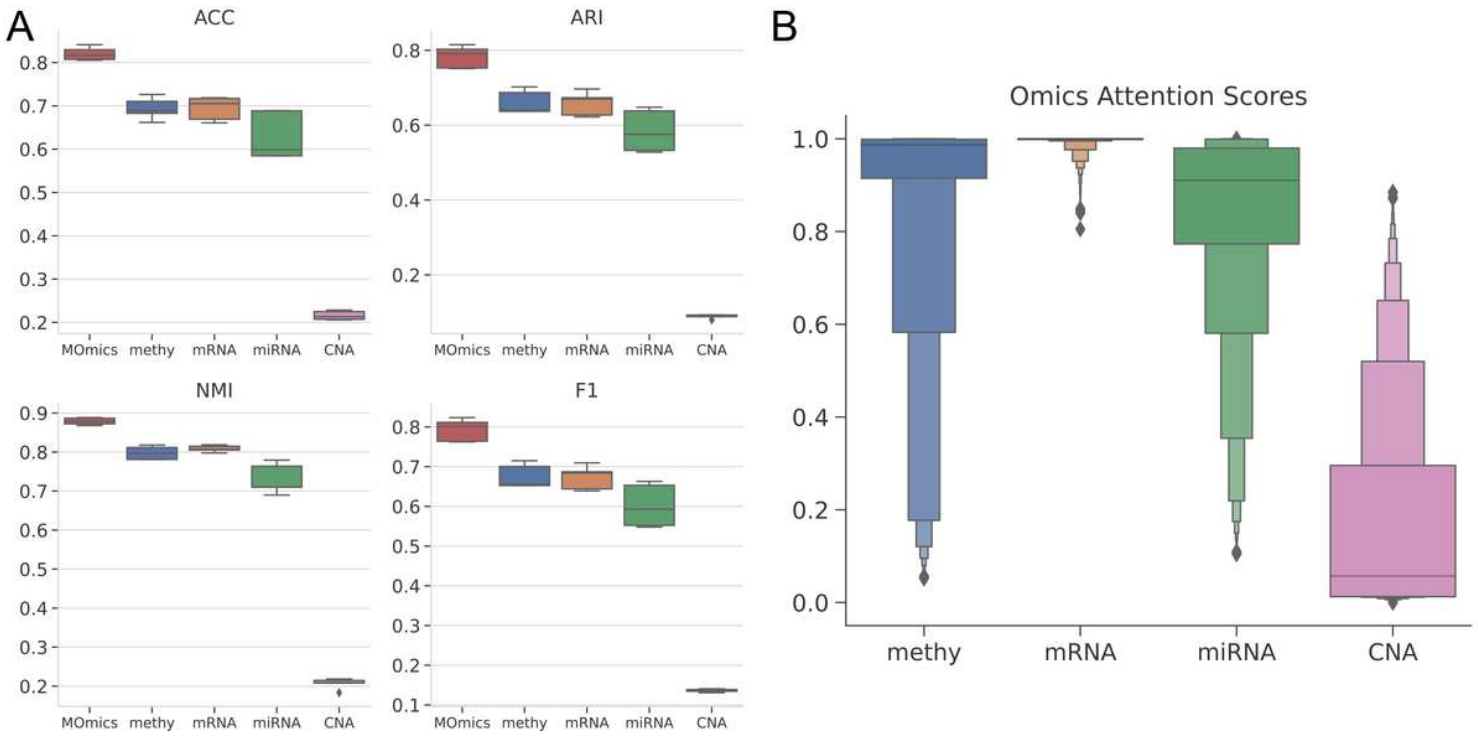

Figure 3

Distribution of metrics and attention scores of MCluster-VAEs using single omics data of the Pan Cancer datasets. **A.** ACC, ARI, NMI and F1 of MCluster-VAEs based on all four omics (MOmics) or single-omics data. **B.** The distribution of attention scores for each omics data. Here, mRNA represents mRNA expression, methy denotes DNA methylation (450K), miRNA represents miRNA expression and CNA represents copy number alterations.

Figure 4

The monitoring records of the training process of MCluster-VAEs on the Pan Cancer dataset. “WA” means using conditional entropy weight decay annealing trick. “Gumbel” means using Gumbel Softmax reparameterization trick. “none” means neither used.

Figure 5

Performance of the algorithms on the ten cancer datasets. The x-axis was the number of clinical parameters enriched in the clusters, and the y-axis measured the differential survival between clusters (-

log10 of permutated logrank's test *P* values). Colors indicated clustering methods.

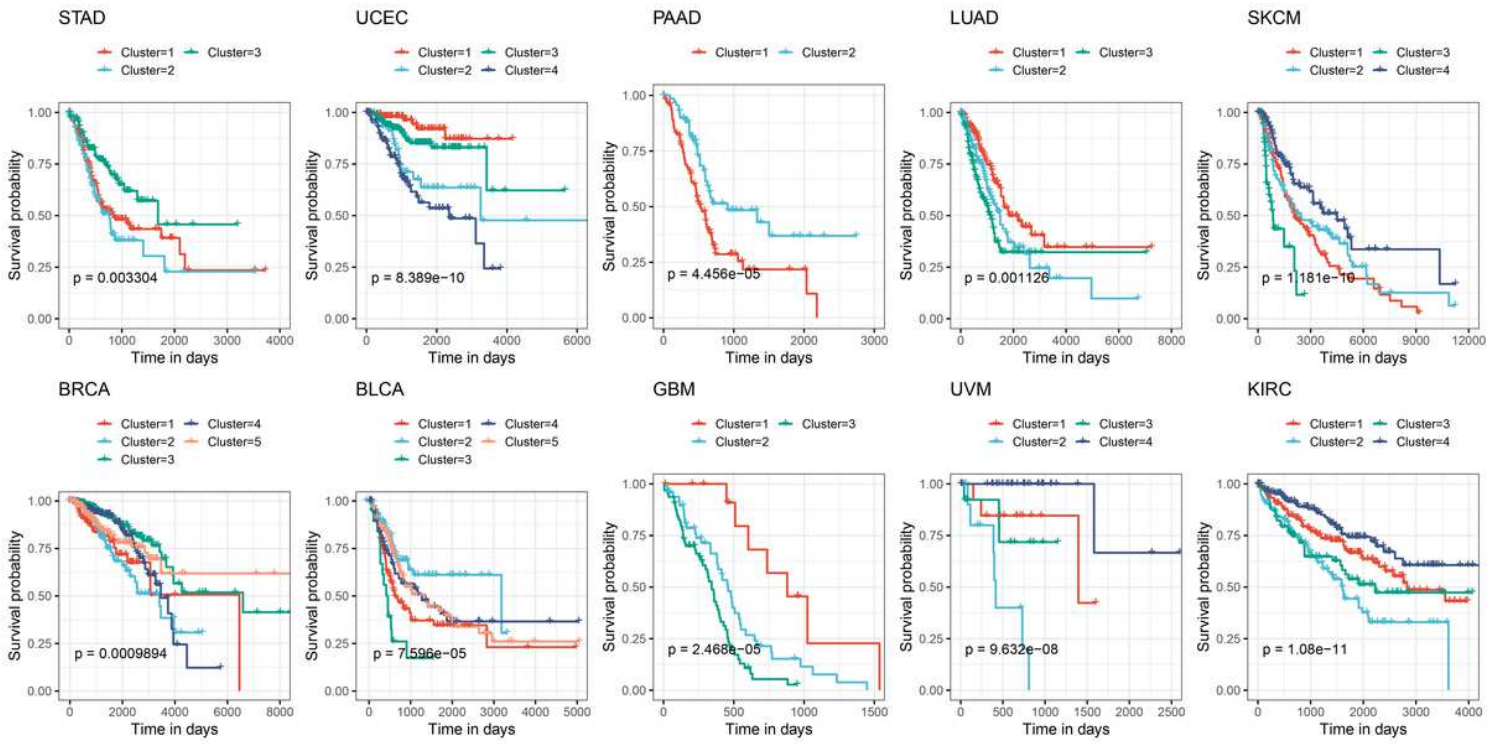

Figure 6

Kaplan-Meier survival plots of MCluster-VAEs on the ten cancer datasets.

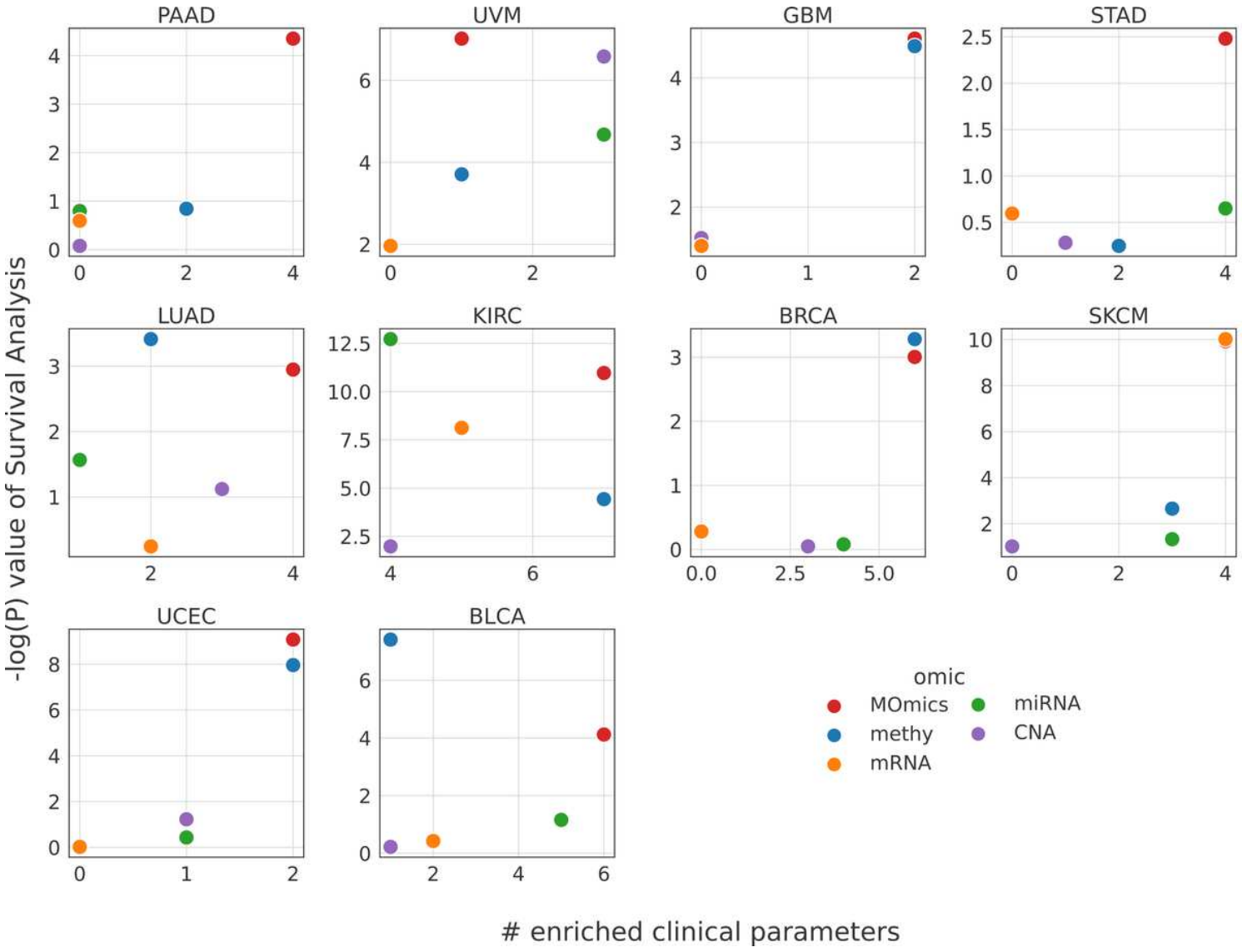

Figure 7

Performance of MCluster-VAEs based on four omics or single-omics data on the ten cancer datasets. The x-axis was the number of clinical parameters enriched in the clusters, and the y-axis measured the differential survival between clusters (-log10 of permutated logrank's test P values). Colors indicated the omics data applied. Here, MOmics represents four omics data, mRNA represents mRNA expression, methy denotes DNA methylation (450K), miRNA represents miRNA expression and CNA represents copy number alterations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.docx