A survey on data integration for multi-omics sample clustering

Neurocomputing 488 (2022) 494–508

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A survey on data integration for multi-omics sample clustering

Marta Lovino [a,*], Vincenzo Randazzo [d], Gabriele Ciravegna [c], Pietro Barbiero [b], Elisa Ficarra [e,1], Giansalvo Cirrincione [f,g,1]

[a] Politecnico di Torino, DAUIN, Turin, Italy
[b] University of Cambridge, Cambridge, UK
[c] University of Florence, DINFO, Florence, Italy
[d] Politecnico di Torino, DET, Turin, Italy
[e] Universitá degli Studi di Modena e Reggio Emilia, Modena e, Reggio Emilia, Italy
[f] University of Picardie Jules Verne, Amiens, France
[g] University of South Pacific, Suva, Fiji

ARTICLE INFO

Article history:
Received 10 April 2021
Revised 9 November 2021
Accepted 27 November 2021
Available online 2 December 2021

Keywords:
Cancer analysis
Competitive learning
Data fusion
Multi-omics clustering
Neural networks
NGL-F
Unsupervised learning
Unsupervised clustering

ABSTRACT

Due to the current high availability of omics, data-driven biology has greatly expanded, and several papers have reviewed state-of-the-art technologies. Nowadays, two main types of investigation are available for a multi-omics dataset: extraction of relevant features for a meaningful biological interpretation and clustering of the samples. In the latter case, a few reviews refer to some outdated or no longer available methods, whereas others lack the description of relevant clustering metrics to compare the main approaches. This work provides a general overview of the major techniques in this area, divided into four groups: graph, dimensionality reduction, statistical and neural-based. Besides, eight tools have been tested both on a synthetic and a real biological dataset. An extensive performance comparison has been provided using four clustering evaluation scores: Peak Signal-to-Noise Ratio (PSNR), Davies-Bouldin(DB) index, Silhouette value and the harmonic mean of cluster purity and efficiency. The best results were obtained by using the dimensionality reduction, either explicitly or implicitly, as in the neural architecture.

1. Introduction

Recently, the decrease in cost in next-generation sequencing (NGS) techniques has enabled the availability of a huge amount of biological data [1–7]. In particular, various types of omics can be obtained from the same sample [8]. The term *omics* refers to a particular type of molecular data providing a specific perspective of a biological phenomenon; indeed, it derives from the suffix of the type of investigation (e.g. genomics, proteomics, transcriptomics, epigenomics) [9–13]. Each of these omics carries partial information of the biological problem. Then, integrating several omics can provide a systemic approach for biological problem investigation. However, despite its informative potential, omic integration is still an open challenge [14].

Although a single-omic study can identify molecules and biomarkers of the main pathologies, it can provide only partial information; nowadays, multi-omics data is fundamental to gain a more accurate insight and more effective predictions [15–17]. The greater availability of data has allowed many multi-omics studies [18–24] and fostered the expansion and construction of public databases to ensemble the greatest amount of data in standardized file formats and user-friendly interfaces. Examples of such projects are the Ensemble Genome Project and the Human Proteome Project, which aim at collecting the major genes and proteins underlying the main biological processes in the cell [25,26]. Other important data repositories are the Genomic Data Commons (GDC), the Clinical Proteomics Tumor Analysis Consortium, and the International Cancer Genomics Consortium [27–31]. In such repositories, the main multi-omics data are RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, proteomics, whole genome sequencing, and the genomic variations data (somatic and germline mutation).

In the last decade, the availability of such an amount of data and information has led to various methodologies and algorithms for their analysis [32–38]. Concerning single-omic dataset processing, the two most common types of analysis are:

* Corresponding author.
E-mail addresses: marta.lovino@polito.it (M. Lovino), vincenzo.randazzo@polito.it (V. Randazzo).
[1] These Authors contributed equally.

https://doi.org/10.1016/j.neucom.2021.11.094
0925-2312/

1. Extraction of the most relevant features for the detection of new biological signatures or pathways.
2. Classification and clustering of samples (typically patients) to create predictive models for a pathology or discover new molecular subtypes.

In a multi-omics scenario, these two approaches are still valid, but the algorithms used to integrate and analyze the data need to be properly modified and optimized.

This work presents the state-of-the-art about multi-omics data integration, especially concerning the classification and clustering of samples.

Several papers reviewed the state-of-the-art for multi-omics integration [39–41]; however, some of these refer to outdated methods or, sometimes, no longer available [42]. More recent reviews [43,44] are complete about the sample clustering problem, but they lack the description of some relevant metrics to realize which method is more suitable in a specific context. Therefore, in this survey, specific clustering metrics (Peak Signal-to-Noise Ratio (PSNR), Davies-Bouldin (DB) index and cluster Silhouette value (S) are used to compare the various methods.

For simplicity and readability, the various multi-omics methods will be grouped into four major categories:

1. *Graph based*. Based on the description of samples such as graphs or similarity matrices (see Section 2).
2. *Dimensionality reduction based*. The integration is given by the joint reduction of the dimensionality among the various omics (see Section 3).
3. *Statistical based*. The prevailing approach for the integration is based on statistics, including Bayesian models (see Section 4).
4. *Neural Networks based*. Techniques based on the creation of artificial neural networks, and, in particular, deep learning methods to integrate multi-omics data (see Section 5).

An algorithm may belong to more than one of the above categories; in this sense, each method is placed in the most representative one. Section 6 provides a comparison among the most popular multi-omics data clustering algorithms, while Section 7 reports the final considerations in the multi-omics clustering domain.

Table 1 summarizes the methods discussed in the following sections for integrating multi-omic data.

2. Graph Based

The first group of techniques of the proposed taxonomy deals with those methods based on the construction of a graph from a similarity matrix: the nodes are the samples, while the edges represent their relationship intensity, measured as the distance (Euclidean or correlation-based) between the samples. Various approaches can be followed to generate a consensus from these similarity matrices. In the following, the major algorithms are presented.

2.1. SNF

The Similarity Network Fusion (SNF) [45] starts from the similarity matrices of the original data and creates a consensus through an iterative algorithm: at each step, the matrices from individual omics are updated, accounting for relevant contributions from the others. This approach has outperformed single-omic studies in some problems such as identification of cancer subtypes and prediction of survival rates when combining mRNA expression, DNA methylation and miRNA expression. The method is simple and fast but requires to have the same samples across all omics.

2.2. MultiSpC

Multi-view spectral clustering (MultiSpC) [46–50] is a generalization of the spectral clustering technique [51] to the multi-omics case. It is based on graphs in which the samples are the nodes and the distances between samples are the arcs. The generalization is based on the Minimizing Disagreement (M-D) algorithm, where samples in two (or more) omics should cluster the data in order to reduce the disagreement between the clustering. As per SNF, the algorithm needs to have the same samples across all omics, while the number of features may differ.

2.3. NEMO

Neighborhood-based multi-omics (NEMO) [52] clustering is a graph-based approach which computes, for each omic, a patient similarity matrix using the Euclidean distance. The similarity matrices are merged into a single matrix, which it is fed to the spectral clustering algorithm to determine, for each sample, the corresponding cluster. This approach is quite efficient because a high-dimensional problem is reduced to a lower one by computing the Euclidean distances among samples (the amount of data is some order of magnitude smaller than the number of features). Its main strength consists of the potential in dealing with partial datasets, where the data related to a patient can be available only for a subset of omics. Also, NEMO can perform data clustering without performing data imputation, and it proved to reach comparable results to state-of-the-art algorithms, which by contrast, work only on complete datasets. However, its major limitation resides in the use of Euclidean distance metric: in a high-dimensional space, samples are more spaced (large inter distances), thus disrupting the meaningful signal inside the dataset. A potential solution could be the use of other distance measures, such as L1-Minkowski and correlation-based distances.

2.4. PINSPlus

Perturbation clustering for data integration and disease subtyping (PINSPlus) [53,54] exploits a similarity-based algorithm to merge the connectivity among samples across all omics. Patient connectivity data are stored in a graph where samples are nodes and distance measures are edges. The novelty of this approach consists in the stability of the clusters, which are tested under three conditions: 1) perturbing the input data, 2) selecting different omics at a time, 3) changing the clustering algorithm. Samples are then grouped together according to the stability across these perturbations using a hierarchical structure search. Although this method is really flexible (the user can select the preferred clustering methods), some biological relationships can be clear only with some clustering methods and not with others according to the input data distribution. In this sense, the role of the user is important in obtaining significant results.

3. Dimensionality reduction based

Another approach, called *Joint Dimensionality Reduction* (jDR), consists in applying dimensionality reduction techniques on the input space, accounting for the features of the different omics. This is achieved through several algorithms aimed at extending to multiple input datasets the dimensionality reduction techniques applied to single matrix processing.

The goal is the projection of the high dimensional omics into a low dimensional space. This is achieved by decomposing the matrices representing each of the L different omic matrices M_i with $i = 1, \ldots, L$, each of size $n_i \times m$ (where m is the number of samples

Table 1
Summary of multi-omic data integration methods.

Method	Family	Core methodology	Optimization objective	Limitations
SNF	Graph	Iterative consensus algorithm	Similarity matrix	Same examples required
MultiSpC	Graph	Spectral clustering	Cluster quality scores	Same examples required
NEMO	Graph	Spectral clustering	Cluster quality scores	Euclidean distance metric
PIN-SPlus	Graph	Clustering	Connectivity matrix	User dependent
JIVE	Dimensionality reduction	Matrix factorization	Min residuals	Linearity assumption
RGCCA	Dimensionality reduction	Matrix factorization	Max correlation	Linearity assumption
tICA	Dimensionality reduction	Matrix factorization	Max correlation	Latent variables independence
MOFA	Dimensionality reduction	Matrix factorization	Max evidence lower bound	Linearity assumption
MSFA	Dimensionality reduction	Matrix factorization	Max evidence lower bound	Normality assumption
intNMF	Dimensionality reduction	Matrix factorization	Min distance	Linearity assumption
MCIA	Dimensionality reduction	Matrix factorization	Max covariance	Linearity assumption
iCluster	Statistical	K-means	Min variance	Linearity assumption
PARADIGM	Statistical	Hierarchical clustering	Max centroid distance	Known pathways only
LRAcluster	Statistical	K-means	Min variance	Linearity assumption
CCA	Neural	Multi Layer Perceptron	Max uncorrelation	Uncorrelation of embeddings
Split-AE	Neural	Auto Encoder	Min reconstruction error	Infinite equivalent latent spaces
DCCAE	Neural	Auto Encoder	Min reconstruction error	Infinite equivalent latent spaces
NGL-F	Neural	Multi Layer Perceptron	Min reconstruction error	Euclidean distance metric

and n_i the number of features) into the product of a $k_i \times m$ factor matrix (F) and $n_i \times k_i$ omics-specific weight/projection matrices (A_i).

There are many methods based on different mathematical formulations. Here are the most representative ones:

3.1. JIVE

Joint and Individual Variation Explained (JIVE) [55] is an extension of the Principal Component Analysis (PCA) [56–59] to multi-omics data. PCA seeks to describe the data with a reduced number of meta-features obtained by linear combination under the condition that the new meta-features are orthogonal and variance is maximized. JIVE decomposes each omic matrix into a joint factor matrix U, an omic-specific factor matrix A and residual noise E: $X_i = U_i S + A_i + E_i$ for $i = 1, \ldots, L$. S is the score matrix explaining variability across multiple types of data. E^i, A^i and U^i are (n x k) matrices. The objective function $||E||^2$ is minimized with E = $[E_1, \ldots, E_L]^T$.

3.2. RGCCA

Regularized Generalized Canonical Correlation Analysis (RGCCA) [60] is a generalization to multi-omics data of the Canonical Correlation Analysis (CCA) [61,62], a method looking for a linear combination of two matrices with the greatest correlation. RGCCA determines a factorization of the same form as JIVE but maximizes the correlation between omic specific factors by finding projection vectors u^i such that the correlation between projected data is maximized: $\underset{i,j}{\text{argmax}}(Corr(X_i u_i, X_j, u_j))$ for all $i, j = 1, \ldots, L$.

3.3. tICA

Tensorial Independent Component Analysis (tICA) [63] is an extension of tensor-based dimensionality reduction methods. In particular, it aims to overcome the limitation of such methods to share both samples and features. It starts from the correlation matrix, whose rows and columns are the samples common to all omics, while its elements (i,j) yield the correlation of sample i with sample j. Then, tICA solves the following equation:

$$X = S \overset{L}{\underset{i=1}{\otimes}} \Omega_i \quad (1)$$

where X represents the multi-omics data organized into a tensor ; S is a tensor with the same dimension of X, composed of $S_1, \ldots, S_L$ mutually statistically independent random variables with E $[S_1, \ldots, S_L]$ = 0 and Var$[S_1, \ldots, S_L]$ = 1 ; and $\otimes$ represents the tensor contraction operation. Since tICA searches for independent signals, the deconvolution of complex mixtures is improved; thus, it better identifies biological functions and pathways underlying the multi-omics data.

3.4. MOFA

Multi-Omics Factor Analysis (MOFA)[64] is an extension of factor analysis, which solves a joint latent variable model composed of a system of equations of the form $M_i = A_i F + E_i$, for $i = 1, \ldots, L$. Here, F represents the latent matrix variable, A_i is the omic-specific weight matrix, and E_i is an error term. A prior distribution is placed on all unobserved variables: a standard normal prior is used for the factors Z, while sparsity priors are used for the weight matrices; finally, various noise models are supported for the error term. The model is then solved by maximising the evidence lower bound (ELBO).

3.5. MSFA

Multi-Study Factor Analysis (MSFA) [65] is a generalization of factor analysis by means of modelling the omic matrices through the following sum: $X_i = \Phi F_i + \Lambda_i + E_i$ for $i = 1, \ldots, L$, where omic specific factors are multivariate normal.

3.6. intNMF

Integrative NMF (intNMF) [66] is an approach based on Non-negative Matrix Factorization, where a matrix A is factorized into two matrices under the assumption that all three matrices are non-negative. The matrix from each omic X_i is factorized into the product of a common factor matrix W and a non-negative, omic-specific matrix H_i, by minimizing the objective function Q = $min_{WH} \sum_{i=1}^{p} \Theta_i ||X_i - WH_i||$. Once the W and H_i matrices have been computed, samples are assigned to the cluster in which they have the highest weight according to W.

3.7. MCIA

Multiple Co-Inertia Analysis (MCIA) [67] is an extension of Co-Inertia analysis (CIA) to more than two omics. MCIA factorizes each matrix into omic-specific factors $X_i = A_i F_i + E_i$ for $i = 1, \ldots, L$, by separately applying the PCA to each omic matrix X_i and then max-

imizing the sum of the squared covariance between the scores of each factor, which corresponds to the global PCA projection:

$$argmax_{q_1^1 \ldots q_1^p} \sum_{k_1}^{L} cov^2(X_k^i q_k^i, X^i, q^i) \text{with} q^i \quad (2)$$

3.8. Scikit-fusion

Matrix tri-factorization (aka scikit-fusion) [68] computes a matrix R (say relation matrix), which encodes the relations inferred between features of different omics, and a matrix C (say constraint matrix), which links features of the same omic. Then, it factorizes all the R matrices by applying matrix tri-factorization under the constraints given by C. R and C matrices are block-matrices, with element R_i containing a relation between the elements of the *i-th* omic and those of the *j-th*; in this sense, the matrix tri-factorization is applied separately to each block.

4. Statistical based

Statistical methods are some of the most common and widely used clustering algorithms for multi-omics data integration. The adoption of probability distributions to model variable factors or the underlying data generation process is the distinguishing factor of statistical approaches [69].

The success of statistical-based methodologies is mainly due to their intrinsic interpretability and straightforward implementation. The possibility of incorporating biological knowledge in model architectures makes these approaches interpretable by design [70]. As interpretability is often a mandatory requirement in many research areas and especially in biology and healthcare, statistical-based techniques have been successfully adopted for multi-omics data integration [43].

If their elementary structure and the possibility of choosing prior distributions are the main reasons for their success, they are also the main limitations of statistical methods [71]. Indeed, these approaches heavily depend on the right choice of both variable factors and prior distributions to converge properly. When prior domain knowledge is scarce or the underlying biological process is highly complex, the correct design and statistical models might be challenging. In the following, some of the most relevant statistical-based methods are presented.

4.1. iCluster

Integrative Clustering (iCluster) [72–74] is a statistical-based method for dimensionality reduction. iCluster assumes the observable data distribution is generated from a fixed linear combination of latent factors. Compared to other dimensionality reduction techniques, iCluster explicitly considers a normally distributed noise matrix as an additional element in the model, accounting for all unobservable and uncertainty factors. Both the expectation–maximization algorithm [75] and Bayesian optimization procedures [76] have been used to optimize the model parameters. Finally, a k-means clustering [77] is performed over the estimated lower-dimensional representation.

4.2. Paradigm

PAthway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM[2]) [78] is a statistical-based algorithm for the analysis of cellular processes through the integration of multiple data sources. PARADIGM integrates the information coming from different omics through factor graphs representing known biological pathways [79]. For each biological entity in the factor graph, PARADIGM provides an estimate of each patient activity. The activity scores are used to find the final clusters using hierarchical clustering with centroid linkage [80]. The state of non-measured biological entities in the factor graph is estimated using the expectation–maximization algorithm [75].

[2] http://paradigm.five3genomics.com/

4.3. LRAcluster

Low-Rank Approximation based multi-omics data clustering LRAcluster [81,82] is a probabilistic approach for dimensionality reduction. The methodology was originally developed to integrate four high-dimensional omic data for the identification of different cancer subtypes. LRAcluster aims at estimating a low-rank ultrahigh-dimensional parameter matrix Θ in order to extract a common low-dimensional subspace for all the omics. The maximum rank of Θ depends on a user-defined parameter r. Once Θ is estimated, LRAcluster computes the singular value decomposition of $\Theta = V\Sigma V^T$ [83]. The common low-dimensional subspace corresponds to the first r columns of ΣV^T. The final clusters are estimated, on the reduced subspace, using k-means [84], whose number of clusters k is evaluated employing Silhouette values [85].

4.4. Fuzzy integration

One of the most substantial assumptions behind multi-omics data integration is that the information is consistent across multiple data sources. Several statistical-based techniques [86–88] have been proposed to relax this assumption, providing multiple clustering labels for each sample so that samples are allowed to be grouped in different clusters for different omics. To this aim, each omic variable's contribution to the final clusters is treated as a random variable itself, whose prior is described using a Dirichlet process [89].

5. Neural Networks based

In the very last years, neural networks and, more specifically, deep neural networks have been applied in the context of data fusion due to the incredible success they have obtained in the single-omic learning tasks [90]. Neural networks are particularly suitable for this application. First of all, as a parametric method, they do not require training data at test time. Secondly, they can deal with both structured data, like gene or protein expressions [91], and unstructured data, such as medical images [92,93]. Furthermore, they are thought to be trained and process a large amount of heterogeneous and noisy data [94]. All the above has paved the way to deep neural network adoption in bioinformatics [95], e.g. for transcription factor binding sites prediction [96–99] or DNA/RNA motif mining [100–103]. However, since lots of training data are not always available, particularly in the medical field, this is also the main weakness of this type of algorithm [104,105]. Finally, in the medical/biological context, the data fusion task has also been referred to as multi-view learning [106]. The difference between the two terms is that while the first only refers to combining the information coming from different data types, the second always includes their employment in a unique supervised/unsupervised learning task.

In the following, we will refer to tasks where only two inputs (X, Y) are given, but all of the reported methods can be extended to the case of many inputs.

Deep architectures Different types of deep architecture are generally available for unsupervised learning. In the field of multi-view learning, the most commonly used architectures are feedfor-

ward neural networks and Auto-Encoders (AE) [107,106]. Neural networks are either trained to maximize the Canonical Correlation [108] or Clustering indices [109].

5.1. Deep-CCA (DCCA)

The Canonical Correlation Analysis (CCA) and kernel variants [110,111] have been extensively employed in the field of multi-view feature learning and dimensionality reduction [112,113]. CCA allows learning an embedding in which features are maximally uncorrelated. Feature correlation can be calculated by dealing with the features learned either from each view or among views. Imposing uncorrelation among views allows learning complementary features from each view. Many works attempted to learn a CCA-like neural network model [110] but only in [114] a full DNN extension has been proposed, named Deep CCA (DCCA). In DCCA, two deep neural networks f and g are learned to extract a single non-linear representation from each input view. Then, the canonical correlation among the extracted feature representations $f(X)$ and $g(Y)$ is maximized:

$$\begin{aligned} f,g := \quad & \max_{W_f,W_g,U,V} \tfrac{1}{N} tr(U^T f(X) g(Y)^T V) \\ s.t. \quad & U^T(\tfrac{1}{N} f(X) f(X)^T + r_x I)U = I, \\ & V^T(\tfrac{1}{N} g(Y) g(Y)^T + r_y I)V = I, \\ & u_i^T f(X) g(Y)^T v_j = 0, \qquad for \quad i \neq j, \end{aligned} \tag{3}$$

where W is the set of learnable weights of each neural network, $tr()$ is the trace function, U and V are the CCA eigenvectors that project the encoding of each network, $r_{x,y}$ is the regularization parameter, and N is the number of training examples. From a theoretical perspective, the DCCA objective cannot be directly optimized since it needs to be calculated over all the input samples. However, stochastic gradient descent (SGD) methods may still be employed, as reported in [115], provided that the mini-batch on which gradients are estimated are sufficiently large and representative of all the population. At last, as shown in [106], DCCA may also work when only one input source is available at test time, with $U^T f()$ being the projection used for testing.

5.2. Split Auto-Encoders (Split-AE)

AutoEncoders (AEs) [116,117] are generally trained to find a compact representation of the input data that best allows their reconstruction. AEs are composed of two fully connected neural networks: the first one $E(x)$ (generally referred to as *encoder*) maps the input data x into a compact latent space. This representation is given as input to another network $D(E(x))$ (also called *decoder*), which projects it back to the original input space. Both networks are trained in such a way that the reconstructed data $\hat{x} = D(E(x))$ is as close as possible to the original data x. Therefore, the trained encoder $E(x)$ projects input data into a reduced space by maximally preserving relevant information (as recently demonstrated in [118]). Split-AutoEncoders [107] (Split-AEs) shift this idea to the multi-view domain. An AE is created for each view with each encoder projecting the input domain to a common latent space and each decoder projecting the data back to the starting input space. All AEs, however, share a common latent space: each decoding function D receives as input the output of all the encoding function E. Taking into consideration again the previous two-view example, the error function for a Split-AE is as follow:

$$\begin{aligned} E_{x,y}, D_{x,y} := \min_{W_E,W_D} \sum_{i=1}^{N} & ||x_i - D_x(E_x(x_i), E_y(y_i))||_2 \\ & + ||y_i - D_y(E_x(x_i), E_y(y_i))||_2, \end{aligned} \tag{4}$$

where x, y correspond to the features of the same sample in each input space, E_x, D_x, E_y, D_y are, respectively, the encoders and decoders for the first and second view, and N is the number of training data. The encoding of each view is concatenated in order to create a shared representation. For instance, in [107] Split-AEs are used to combine audio and visual information. More precisely, they train a Split-AE to reconstruct videos of people pronouncing certain words (e.g. digits) when also the corresponding audio is available. In Fig. 1 the learnt representations in terms of the most strongly correlated input features in both domains are reported for two samples. Also in this case, Split-AE works even if only one input view is available at test time: a single encoder may be used to represent all the sufficient information to reconstruct input data in all views. At last SGD, or other gradient-based method, may be employed to optimize Eq. 4, weighing more some of the terms according to the final goal.

5.3. Deep canonically correlated autoencoders (DCCAE)

Inspired by previous works [114,107], Deep canonically correlated autoencoders (DCCAE) [106] combines the maximization of the canonical correlation among the representations extracted from each view with a reconstruction error of SplitAE. More precisely, DCCAE employs the same structure as in [107] and adds to the autoencoder optimization problem a CCA regularization term on the learned representations:

$$\begin{aligned} E_{x,y}, D_{x,y} := \quad & \max_{W_E,W_D,U,V} \sum_{i=1}^{N} ||x_i - D_x(E_x(x_i), E_y(y_i))||_2 \\ & + ||y_i - D_y(E_x(x_i), E_y(y_i))||_2 \\ & - \lambda \tfrac{1}{N} tr(U^T E_x(X) E_y(Y)^T V) + \\ s.t. \quad & \text{sameconstraintsasin3} \end{aligned} \tag{5}$$

where λ is a weight parameter, which balances the contribute of the CCA in the overall optimization. From the information theory point of view [119], by minimizing the reconstruction error, the autoencoder maximizes the mutual information between the inputs and their projections into the common latent space [120], while the CCA maximizes the mutual information between the view projections [121]. The DCCAE loss function aims at finding the equilibrium between the information captured in the input-projection mapping within each view and the information collected in the projections among views.

5.4. Neural Graph Learning for data-Fusion (NGL-F) neural network

The Neural Graph Learning for data Fusion (NGL-F) is a gradient-based clustering neural network [109,122], which uncovers topological sample-to-sample relationships using multiple data

Fig. 1. Visualization of the learnt representations in the task of reconstructing videos and audio of people pronouncing words. Image taken from [107].

sources. The output of NGL-F is a set of graphs. For each input set, NGL-F aims at finding a graph where nodes represent cluster centroids while edges represent cluster topological properties. The learned topology described by such graphs is used to create the sample adjacency matrix (S). The information contained in the matrix represents all datasets.

NGL-F is composed of a set of dual multi-layer perceptrons (MLPs) [109], one for each dataset. Unlike other previous works, however, each network works on the transpose of the input matrix [122], which allows employing many hidden layers, preserving, at the same time, data topology. For instance, by working on the transpose of the data matrix, the input space is maintained through the network layers. Each MLP provides as output a set of vectors $w_i \in \mathbb{R}^d$ representing cluster centroids for the input data. The architecture of each network can be customized according to the complexity of its own dataset.

The loss function of NGL-F takes into account, at the same time, the quality of clusters found by each MLP and their underlying topology. The relationships among clusters are modeled using an adjacency matrix E, where $E(i,j)$ represents the number of samples for which w_i and w_j are the two closest centroids. The higher $E(i,j)$, the more their respective clusters are related. The loss function for each view is composed of three terms taking into account inter- and intra-cluster distances, quantization error, and parsimony in representing the underlying topology:

$$\mathscr{L}_z = \frac{\max_k d_{intra}(C_k)}{\max_{i,j} d_{inter}(C_i, C_j)} + Q + ||E|| \tag{6}$$

where $d_{intra}(C_k)$ is the intra-cluster distance, $d_{inter}(C_i, C_j)$ the inter-cluster distance, and Q the quantization error. The NGL-F loss function is the linear combination of the loss function in the different views: $\mathscr{L} = \sum_z \mathscr{L}_z$.

Once all networks terminate the training procedure, the resulting clusters are analyzed. For each input set, two samples are considered near each other if they belong to the same cluster, far from each other, if they belong to different clusters. A sample adjacency matrix S is then computed as follow: $S(i,j) = \sum_{d=1}^{n} near_d(i,j)$, where $near_d(i,j)$ is a boolean function calculating the proximity of the samples as previously explained and n is the number of datasets taken into consideration. This matrix is the result of the fusion process.

6. Benchmarks for performance evaluation and comparison

The multi-omics paradigm has been investigated to assess the clustering capabilities of state-of-the-art techniques. To this purpose, eight methods have been selected, and their performance compared on standard quality indices. Two datasets have been employed as benchmarks, one synthetic and the other biological.

Two datasets have been employed as benchmarks, one synthetic and the other biological. The synthetic dataset has been chosen to control specific conditions in the data (e.g., the number and the density of the clusters and the number of samples). Data in the synthetic dataset are very well clustered. Thus, this dataset is ideal for testing the tool performances in a controlled condition but does not fully represent the biological variation in the data. Therefore, a biological dataset has been employed, which by contrast is not controllable in terms of parameters, but it represents the typical multi-omics dataset.

In the synthetic dataset, three omics have been generated in R using the InterSIM package [123]: the mRNA raw count gene expression values (131 features), the methylation values (367 attributes), and the relative protein expressions (165 variables). Each omic is composed of 500 samples, grouped in five clusters.

The biological dataset has been downloaded from the NIH Genomic Data Commons portal [124]. The dataset is composed of two omics: mRNA and miRNA transcriptome profiling matrices of lung samples. The former is composed of raw counts gene expression values (17683 features) [125]; higher values correlate with a higher protein production rate. The second omic consists of raw counts of miRNA values (1665 features) [126]; higher values indicate a reduction in mRNA-translated protein as miRNA inhibits mRNA translation. Both datasets consist of 1250 samples extracted from either cancerous or healthy lung tissues. The data have been collected from four different projects: *TCGA-LUAD* [127] and *CPTAC-3*, with samples from Lung Adenocarcinoma (LUAD) patients; *TCGA-LUSC*, with samples from Lung Squamous cells Carcinoma (LUSC); and *TCGA-MESO* from Mesothelial neoplasm (MESO). Usually, healthy samples have been taken from non-tumoral tissues adjacent to the tumor. From the above metadata, seven different labels have been generated to check the quality of the clusters predicted by each method:

- TCGA-LUAD_healthy
- TCGA-LUAD_tumoral
- TCGA-LUSC_healthy
- TCGA-LUSC_tumoral
- CPTAC-3_healthy
- CPTAC-3_tumoral
- TCGA-MESO_tumoral

Table 2 reports the eight clustering algorithms compared in the experiments. Only techniques with publicly available software and clear documentation were selected. The first two algorithms, SNF and MultiSpC, belong to the graph-based group described in Section 2. Among the dimensionality reduction methods (see Section 3), JIVE, RGCCA, tICA, MOFA were selected. Finally, the iCluster and NGL-F techniques were tested for the statistical (Section 4) and and neural network (Section 5) categories, respectively.

6.1. Quality indices

In order to compare clustering algorithms, we selected a set of metrics that are not directly related to a specific biological problem in order to provide a fair comparison among the different techniques.

The first index used for assessing the clustering performances is the Peak Signal-to-Noise Ratio (*PSNR*) [130], which is one of the most famous and widely used measures of the fidelity of a representation (i.e., a clustering) w.r.t. the original signal. The *PSNR* is defined as:

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{7}$$

where MAX_I^2 is the squared Euclidean norm of the vector connecting the two most distant samples in the input distribution, and *MSE* is the mean squared error between each centroid weight vector and its associated data. The higher the *PSNR* value, the better the clustering.

PSNR measures only the intra-cluster compactness, but it does not take into account the inter-cluster separation. To this end, the Davies-Bouldin index (*DB*) [131] has been employed as it considers both aspects:

$$DB = \frac{1}{N}\sum_{i=0}^{N} \max_{j \neq i} \frac{RMSE_i + RMSE_j}{D_{i,j}} \tag{8}$$

Table 2
Summary of the methods selected for benchmark comparison.

Method	Type	Source	Reference
SNF	Graph based	https://cran.r-project.org/web/ packages/SNFtool/index.html	[128]
MultiSpC	Graph based	https://it.mathworks.com/help/stats/spectralcluster.html	[50]
JIVE	Dimensionality reduction based	https://cran.r-project.org/web/ packages/r.jive/index.html	[55]
RGCCA	Dimensionality reduction based	https://cran.r-project.org/web/ packages/RGCCA/index.html	[60]
tICA	Dimensionality reduction based	https://cran.r-project.org/web/ packages/tensorBSS/index.html	[63]
MOFA	Dimensionality reduction based	https://www.bioconductor.org/ packages/release/bioc/html/MOFA.html	[64]
iCluster	Statistical based	https://cran.r-project.org/web/ packages/iCluster/index.html	[72]
NGL-F	Neural network based	https://github.com/ pietrobarbiero/cola/blob/ 82f05f639bb14bdb3e65a0008f9447 ffc88bb204/fexin/_fexin.py	[122,129]

where $RMSE_i$ is the Root Mean Squared Error [132] for the *ith* cluster $D_{i,j}$ is the Euclidean distance between the *ith* and *jth* cluster centroids, and N is the number of clusters. Lower DB values indicate better clustering.

The third quality measure used in the experiments is the cluster Silhouette value (S) [85]. As the DB index, it considers both the inter-cluster and intra-cluster distances and is defined as:

$$S = \frac{1}{C}\sum_{i=1}^{C}\frac{b(i)-a(i)}{\max(a(i),b(i))} \tag{9}$$

where $a(i)$ is the average distance of the *ith* sample from the samples in the same cluster, $b(i)$ is the minimum among the mean distances of the *ith* sample from the samples in the other clusters, and C is the cardinality of the current dataset. While DB checks compactness and cluster separation, the S index estimates if, on average, samples are correctly assigned to the nearest neighbouring cluster [130]. Because of Eq. 9, $S \in [-1, 1]$, where a high value indicates a good clustering.

The last metric used in the experiments was the harmonic mean (PE) between cluster efficiency and purity [130]. The two metrics were computed, averaging their scores obtained for each predicted cluster and for each ground-truth label. The efficiency is the ratio between the number of samples with the same ground-truth label i in the same cluster over the overall number of samples labeled as i. The purity is the ratio between the number of samples with the most common ground-truth label j in the same cluster over the overall number of samples of the cluster.

The selected metrics have been chosen primarily because they are among the most used to assess clustering algorithms and also because they are complementary to each other. In fact, they can be used to efficiently summarize a wide range of information such as: the amount of information retained by cluster centroids (PNSR), clusters' compactness (DB), distance among different clusters (DB), closeness to the nearest centroid (Silhouette), the class-homogeneity of clusters (purity, PE), and the scattering of samples of the same class across different clusters (efficiency, PE).

6.2. Synthetic dataset

The first benchmark deals with data drawn from the 500 samples synthetic dataset. The output matrix of each of the eight algorithms has been clustered using k-means [84] with a number of target centroids equal to the number of expected clusters, i.e., five, to perform a fair comparison.

The *PSNR* has been computed for the output matrix of each omic (see Fig. 2a) in order to evaluate the amount of information retained by cluster centroids. The two graph-based methods behave in opposite ways. The SNF clustering is quite poor ($\approx 3dB$), while MultiSpC performs well ($\approx 25dB$). The dimensionality reduction group exhibits a common trend ($17-20dB$) with the exception of JIVE, which has the highest PSNR value ($\approx 42dB$) among all techniques. Statistical and neural-based approaches show a similar clustering performance ($27-30dB$) in identifying meaningful cluster centroids, slightly higher than MultiSpC but still much lower than JIVE.

The Davies-Bouldin index has been computed by concatenating all the three omics (i.e., *TS*). Fig. 2b illustrate the results in considering the compactness and the distance between different clusters. Conversely to the previous metric, six out of eight techniques show about the same performance ($DB = 1.1$). SNF obtains a slightly higher value ($DB = 1.8$), while MultiSpC clusters are significant worse (≈ 13).

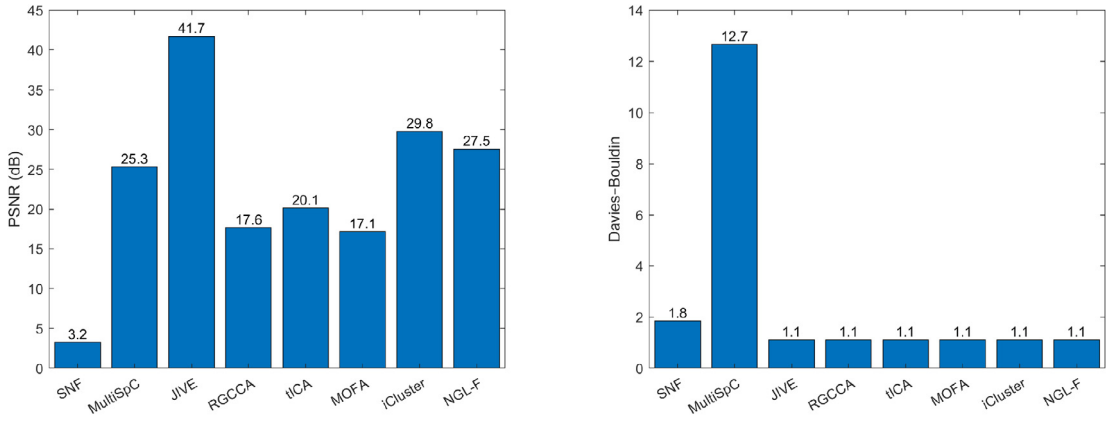
Silhouette scores are reported in Fig. 3 for each method and measures for each sample the closeness to the nearest centroid. As per the DB, this index has been computed on the concatenated omics. SNF (see Fig. 3a) groups properly the first and the fourth clusters ($S > 0.6$); the third cluster has a lower but still good Silhouette score (≈ 0.4), while SNF was not able to detect the two remaining groups. According to the Silhouette score, MultiSpC was not able to identify correctly the clusters, as shown in Fig. 3b. Dimensionality reduction based approaches performed better than the previous category. With the exception of tICA (see Fig. 3f), the other three algorithms - JIVE (Fig. 3c), RGCCA (Fig. 3d), and MOFA (Fig. 3d), and - obtained a high S score (≈ 0.7) for all clusters. Similar results have been obtained by iCluster (see Fig. 3h). Finally, the quality of NGL-F clusters (see Fig. 3i) was similar, on average, to tICA (≈ 0.5).

The last comparison for the synthetic dataset has been done by means of the harmonic mean PE of each cluster purity and efficiency. Fig. 4 shows algorithm performances (colors identify clusters). Six out of eight methods yielded a perfect clustering in terms of purity and efficiency for all classes. SNF scores were slightly worse, while MultiSpC clustering quality was the lowest.

In conclusion, it can be stated that JIVE is the best technique, w.r.t to the proposed metrics, in clustering the synthetic dataset. The second-best option is iCluster, followed by NGL-F and the remaining dimensionality reduction methods. SNF exhibits a moderate ability to cluster this benchmark, while the worst performance is shown by MultiSpC.

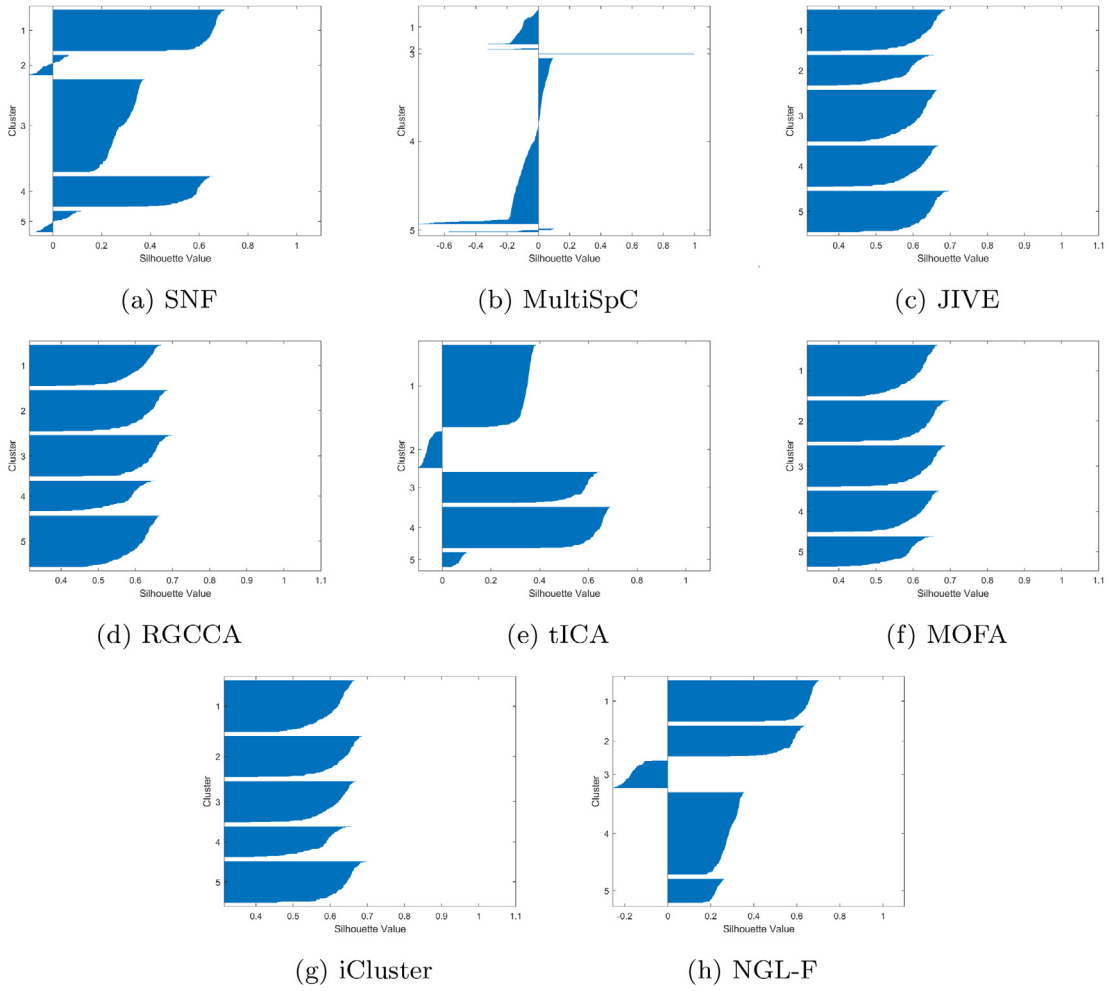
6.3. Lung dataset

The second benchmark consisted of data extracted from the NIH Genomic Data Commons lung dataset. Genes with an expression value of zero across all the samples were removed from the analy-



(a) PSNR (b) Davies–Bouldin

Fig. 2. Quality indices for the synthetic dataset: (a) *PSNR* (the higher the better) and (b) Davies-Bouldin (the lower the better). Each column yields the index value for the corresponding technique.



(a) SNF (b) MultiSpC (c) JIVE

(d) RGCCA (e) tICA (f) MOFA

(g) iCluster (h) NGL-F

Fig. 3. Silhouette index for the synthetic dataset computed for each cluster (Y-axis): graph-based (SNF and MultiSpC), dimensionality reduction based (JIVE, RGCCA, tICA and MOFA), statistical-based (iCluster) and neural network based (NGL-F). Values close to 1 are related to good clustering, while negative values imply a poor clustering quality.

sis. The mRNA matrix was normalized using a variance stabilizing transformation [133]. The miRNA matrix was scaled by taking the $log_2(exprValue + 1)$ [134] over the normalized values obtained with the DESeq2 algorithm [135]. The output matrix of each of the eight algorithms has been clustered using k-means [84] with a number of target centroids equal to the cardinality of the label set, i.e., seven. Then, PSNR, DB, Silhouette and PE indices are computed. Their meaning is summarized at the and of subSection 6.1. The *PSNR* has been computed between each multi-omics output matrix and the corresponding k-means closest centroid, see Fig. 5a. The two graph-based methods obtained opposite results. SNF clustering was quite poor ($\approx 3dB$), while MultiSpC obtained the best results overall ($PSNR = 29dB$). The dimensionality reduction group yielded similar results ($\approx 24dB$) with the exception of MOFA,

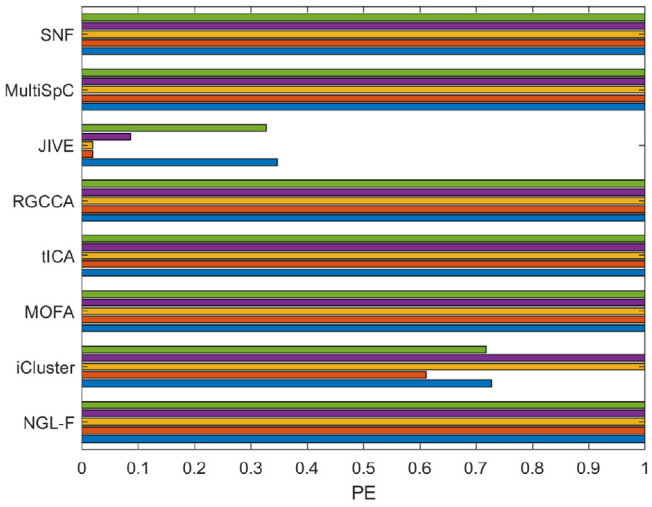


Fig. 4. PE quality indices for the synthetic dataset. Each bar (the higher the better) yields the index value for the corresponding technique and cluster (identified by colors).

whose PSNR (28.8dB) was close to MultiSpC. NGL-F showed a similar clustering performance, slightly lower than MOFA ($\approx 26dB$). Finally, iCluster obtained the worst results overall ($PSNR = 17dB$).

The Davies-Bouldin index has been computed by concatenating the three omics and the k-means Voronoi sets. Fig. 5b illustrates the results. The *DB* score varied significantly among the different algorithms without coherence within each category. The tICA technique obtained the lowest score overall ($DB = 1.1$), followed by JIVE ($DB = 2.5$), RGCCA ($DB = 4.1$), and MOFA ($DB = 9$). In the graph-based group SNF clusters obtained a result slightly above RGCCA ($DB = 4.4$), while MultiSpC had the worst Davies-Bouldin score overall ($DB = 9.2$). The statistical-based algorithm (iCluster) performance was slightly better ($DB = 3.4$) than SNF and RGCCA but worse than JIVE and tICA. Finally, NGL-F score was quite high ($DB = 5$).

Fig. 6 illustrates the Silhouette scores for each method. This index has been computed by concatenating the three omics and the k-means Voronoi sets. Based on this metric, SNF (see Fig. 6a) was able to properly identify only the sixth cluster ($S \approx 0.6$), while the remaining ones were not appropriately learned by the technique. MultiSpC was not able to identify any cluster, as shown in Fig. 6b. Dimensionality reduction approaches performed much better than the previous category. Both JIVE (see Fig. 6c) and tICA (see Fig. 6e) were able to identify the first three clusters. In addition, JIVE detected the fourth group, while tICA was able to model the remaining three clusters. RGCCA results were similar to SNF (see Fig. 6d), while MOFA (see Fig. 6f) was able to detect the third cluster only. The statistical-based approach (shown in Fig. 6g) found the third and sixth clusters ($S > 0.6$) and only partially the first one. Finally, NGL-F obtained a good score only for the sixth ($S \approx 0.8$) and the fifth clusters ($S \approx 0.4$), as reported in Fig. 6h.

The last comparison for the lung dataset has been done according to the harmonic mean *PE* of purity and efficiency for each cluster. Fig. 7 shows algorithm performances (colors identify clusters). SNF obtained a good result for five out of seven clusters ($PE > 0.5$), while MultiSpC scores were very low for all groups, with the exception of the seventh one. Among dimensionality reduction techniques, JIVE and RGCCA obtained good results for all groups, while tICA and MOFA received lower scores for some clusters. Finally, iCluster and NGL-F had a similar clustering performance, slightly worse than dimensionality reduction methods but better than graph-based ones.

The algorithms with the best *PSNR*, i.e., MultiSpC, MOFA, and NGL-F, obtained the worst *DB* score. This result is confirmed by their silhouettes. *DB* and *S* classify the algorithms in a very similar way, with tICA, JIVE and iCluster among the best techniques and MOFA and MultiSpC as the worst ones. Finally, according to the PE ranking, RGCCA looked like the best approach, followed by JIVE and tICA, while graph-based techniques obtained the worst performance.

In conclusion, the above results showed how JIVE and tICA were among the best algorithms in clustering the lung dataset with regard to the proposed metrics. The second-best option was iCluster, followed by NGL-F and the remaining dimensionality reduction methods. The lowest scores were obtained by SNF and MultiSpC.

6.4. Discussion

In this section, two experiments have been conducted to compare the quality of some of the most common algorithms for multi-omics clustering. These techniques were uniformly selected among the classes identified in Sections 2–5.

Overall, the performance of all the algorithms that explicitly implemented multi-omics clustering was good. Only MultiSpC consistently reported lower results in all quality indices on both datasets. The data fusion step in MultiSpC only consists of the concatenation of the input matrices. Sometimes, this straightforward approach is not sufficient to correctly combine highly different input datasets, as reported in the experiments. The highest performance has been obtained by dimensionality reduction-based methods (JIVE and MOFA on the synthetic dataset, tICA and JIVE on the lung dataset).

To get deeper insights on the above analysis, two lung omics manifolds have been studied to estimate their corresponding intrinsic dimensionality[3].

At first, a linear manifold was assumed, and PCA was used to obtain a lower-dimensional representation. A cumulative explained variance greater than 90% was considered as a good indicator for assessing the size of the lower-dimensional subspace. The number of principal components required to explain the 90% of the variance (i.e. the intrinsic dimensionality δ) was equal to 459 and 494 for mRNA and miRNA datasets, respectively.

The linear analysis was used only to have an initial value for the intrinsic dimensionality δ. From this starting point, a more complex non-linear technique, the Curvilinear Component Analysis [136,137], was used to refine this estimation. The Curvilinear Component Analysis is a self-organizing neural network for data projection, which maintains the input topology by means of local distance preservation. In this sense, it can be used to reduce the number of input variables without altering the shape of the original manifold. A fundamental tool associated with this neural technique is the *dx-dy* diagram, where the in-between neuron distances in the projected space (dy) are plotted against their corresponding ones in the input space (dx). The projection results for the miRNA ($\lambda = 50, proj_{Dim} = 80, epochs = 200, \alpha_0 = 0.5$) and the mRNA ($\lambda = 280, proj_{Dim} = 100, epochs = 100, \alpha_0 = 0.5$) omics are shown in Fig. 8a and Fig. 8b, respectively. Because blue points are aligned along the bisector, the input topology was preserved by the projection in both cases. This analysis suggests that the intrinsic dimensionality of the mRNA and miRNA dataset lies between 80 and 100, respectively. This can explain why multi-omics approaches reducing the input dimensionality were able to properly cluster input data.

[3] The notion of intrinsic dimensionality refers to the fact that any low-dimensional data space can trivially be turned into a higher-dimensional space by adding redundant or randomized dimensions, and, in turn, many high-dimensional datasets can be reduced to lower-dimensional ones without significant information loss.

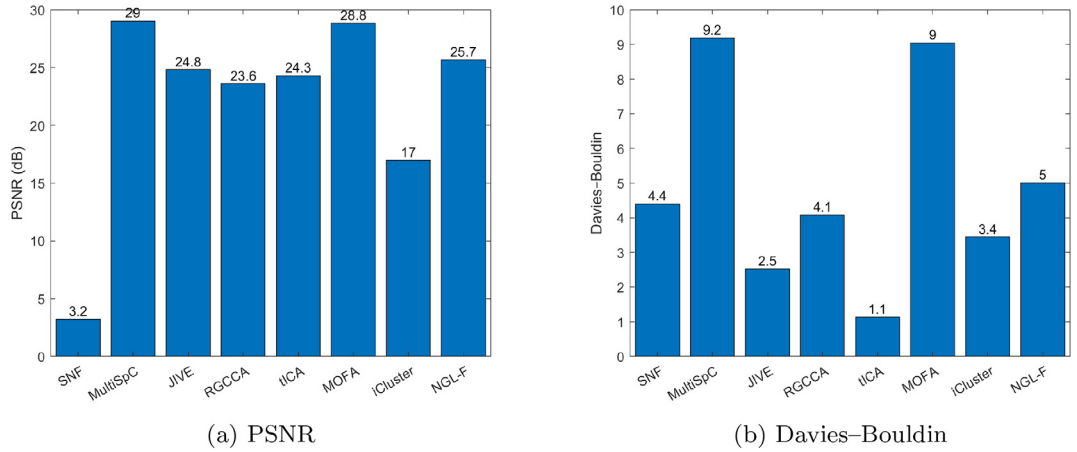


Fig. 5. Quality indices for the lung dataset: (a) *PSNR* (the higher the better) and (b) Davies-Bouldin (the lower the better). Each column yields the index value for the corresponding technique.

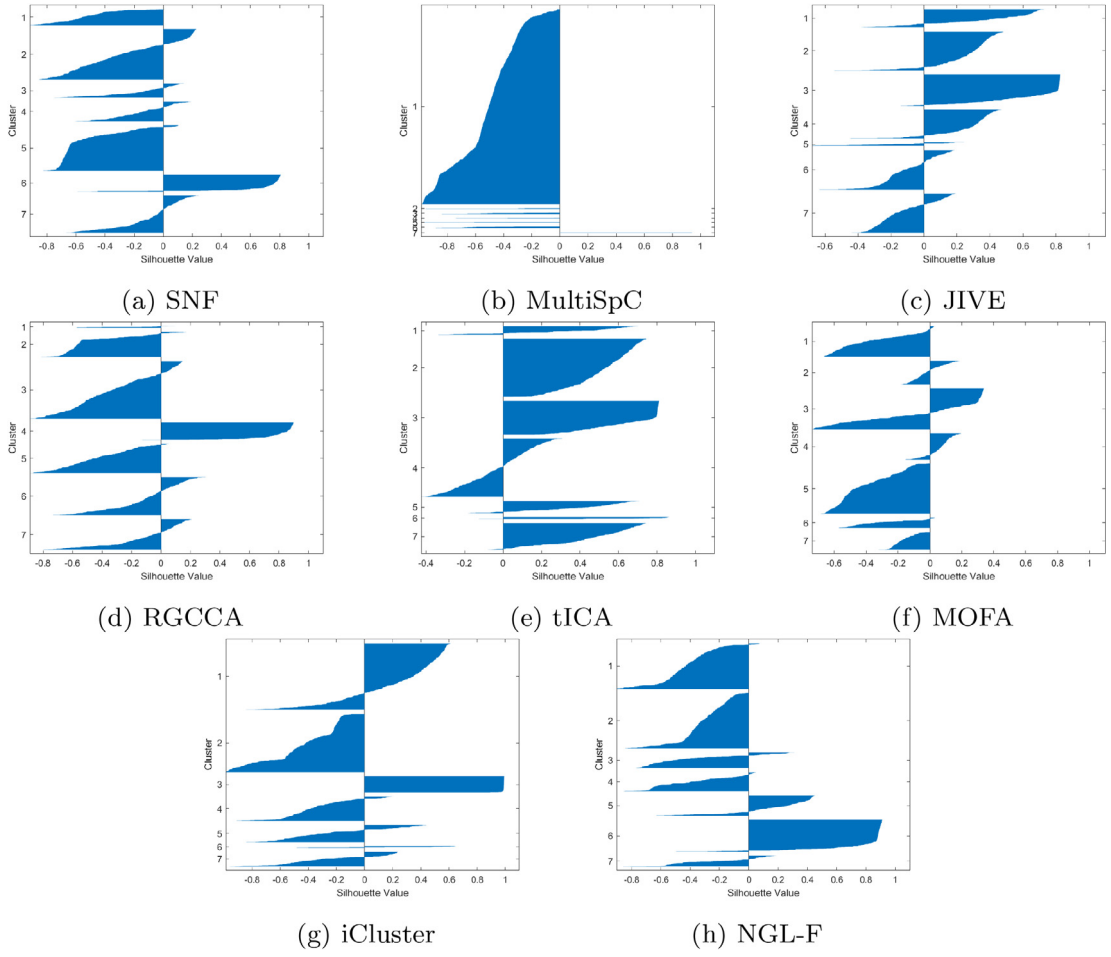


Fig. 6. Silhouette index for the lung dataset computed for each cluster (Y-axis): graph-based (SNF and MultiSpC), dimensionality reduction based (JIVE, RGCCA, tICA and MOFA), statistical-based (iCluster) and neural network based (NGL-F). Values close to 1 are related to good clustering, while negative values imply a poor clustering quality.

The dimensions of the input space of the lung dataset, as well as the dimensionality obtained by applying the methods used for the experiments, are reported in Table 3. The dimension of the input space refers to the sum of the dimensions over all the omics. Therefore, this measure is identical for all the methods and it is ~ 20000. All the selected methods allow the user to define a priori the dimension of the output space, except for RGCCA and tICA. For these methods, the output dimension has been set to 10 since this value has been optimized for the problem at hand [138].

All the reported methods generate a low-dimensional representation of the data from each omic, except for MultiSpC. RGCCA aims at maximizing the Canonical Correlation, while tICA optimizes the

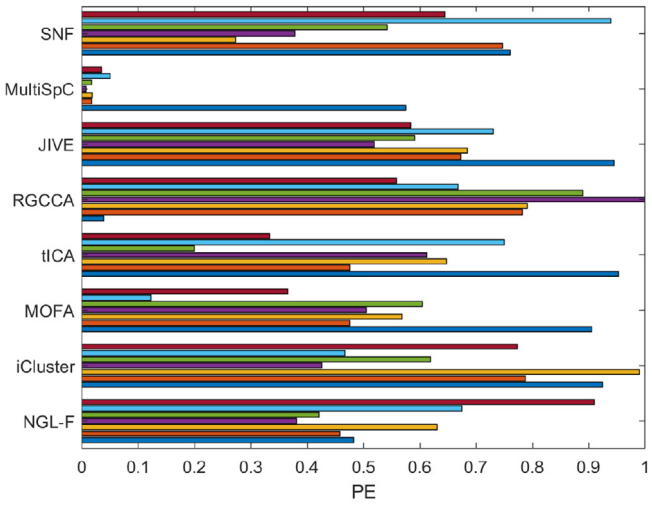


Fig. 7. PE quality indices for the lung dataset. Each bar (the higher the better) yields the index value for the corresponding technique and cluster (identified by colors).

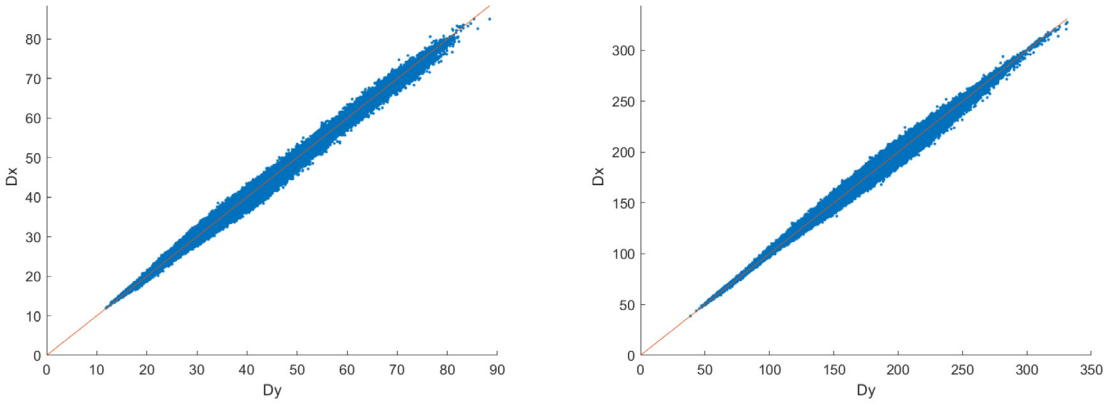
Independent Correlation and JIVE performs a variant of the PCA. Both MOFA and iCluster adopt different variants of the Factor Analysis. SNF does not perform a dimensionality reduction technique directly, but it exploits the sample similarity matrices, thus working in the sample space instead of the feature space (the sample space is usually at least 10 orders of magnitude smaller than the feature space). NGL-F implicitly performs a projection of the input data for each omic in the hidden layers, similarly as an encoder.

The choice of the best multi-omics clustering algorithm depends on data topology. If the clusters are not embedded in lower-dimensional subspaces, dimensionality reduction-based methods may lose their advantage. Other considerations can be drawn about the choice of the metrics. As the dimensionality increases, the difference between points that are close or far disappears [139] (two arbitrary vectors become orthogonal [140]). As a consequence, all nearest neighbor strategies (like k-means) may struggle. In this case, a possible solution consists in using fractional Minkowski distances [139]. All previous considerations are out of the scope of this survey. This work addresses the taxonomy of existing algorithms and compares a few representative techniques on challenging benchmarks, whose dimensionality still allows the use of Euclidean metrics.

7. Conclusions

This paper aims at providing a general overview of the major techniques for biological sample clustering, which can be divided into four groups, according to the underlying approach: graph, dimensionality reduction, statistics and neural network. The most meaningful algorithms have been tested, both on a synthetic and a real biological dataset, and their performance has been compared using four clustering evaluation scores (*PSNR*, *DB*, *S* and *PE*).

In both experiments, the dimensionality reduction-based approach seems to be the best way to tackle multi-omics clustering. On the contrary, graph-based algorithms are not able to properly deal with this kind of problem. Finally, statistical and neural network-based methods have promising performance and may deserve further improvement. As a further investigation, it would be interesting to test multi-omics approaches for controlled databases, i.e., changing the topological and statistical properties. This paradigm would address questions like finding the best method in the case of non-separable clusters, increasing noise or dimensionality, different inter-distances, so on and so forth. This analysis will be our future line of research.



(a) miRNA (b) mRNA

Fig. 8. The *dy-dx* diagrams for miRNA (left) and mRNA (right) omics: blue points are the in-between neuron distances, red line indicates the bisector.

Table 3
Summary of the dimension space before and after applying the selected methods.

Method	Type	Dimension of the input space	Dimension after applying the selected method
SNF	Graph based	≈ 20000	1250
MultiSpC	Graph based	≈ 20000	5
JIVE	Dimensionality reduction based	≈ 20000	400
RGCCA	Dimensionality reduction based	≈ 20000	10
tICA	Dimensionality reduction based	≈ 20000	10
MOFA	Dimensionality reduction based	≈ 20000	1
iCluster	Statistical based	≈ 20000	200
NGL-F	Neural network based	≈ 20000	1250

CRediT authorship contribution statement

Marta Lovino: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Project administration. **Vincenzo Randazzo:** Software, Formal analysis, Writing - original draft, Writing - review & editing. **Gabriele Ciravegna:** Methodology, Software, Writing - original draft. **Pietro Barbiero:** Methodology, Software, Formal analysis, Writing - original draft. **Elisa Ficarra:** Conceptualization, Methodology, Formal analysis, Writing - review & editing, Supervision, Funding acquisition. **Giansalvo Cirrincione:** Conceptualization, Methodology, Formal analysis, Writing - review & editing, Supervision.

Funding

This study was funded by the European Union's Horizon 2020 research and innovation programme DECIDER under Grant Agreement 965193.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Pietro Barbiero is an AI engineer and member of the founding team of Bactell Inc. However, there has been no financial support for this work that could have influenced its outcome. All the other Authors declare no conflict of interests.

Acknowledgments

The Authors thank Mattia Siviero for his comments on an earlier version of the manuscript.

References

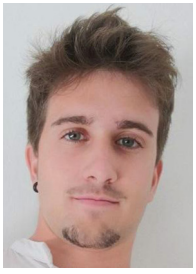
[1] P. Muir, S. Li, S. Lou, D. Wang, D.J. Spakowicz, L. Salichos, J. Zhang, G.M. Weinstock, F. Isaacs, J. Rozowsky, et al., The real cost of sequencing: scaling computation to keep pace with data generation, Genome biology 17 (1) (2016) 1–9.
[2] Y. Souilmi, A.K. Lancaster, J.-Y. Jung, E. Rizzo, J.B. Hawkins, R. Powles, S. Amzazi, H. Ghazal, P.J. Tonellato, D.P. Wall, Scalable and cost-effective ngs genotyping in the cloud, BMC medical genomics 8 (1) (2015) 1–9.
[3] I.G. Gut, New sequencing technologies, Clinical and Translational Oncology 15 (11) (2013) 879–881.
[4] C.W. Fuller, L.R. Middendorf, S.A. Benner, G.M. Church, T. Harris, X. Huang, S.B. Jovanovich, J.R. Nelson, J.A. Schloss, D.C. Schwartz, et al., The challenges of sequencing by synthesis, Nature biotechnology 27 (11) (2009) 1013–1023.
[5] M. Lovino, G. Urgese, E. Macii, S. Di Cataldo, E. Ficarra, A deep learning approach to the screening of oncogenic gene fusions in humans, International journal of molecular sciences 20 (7) (2019) 1645.
[6] M. Lovino, M.S. Ciaburri, G. Urgese, S. Di Cataldo, E. Ficarra, Deeprior: a deep learning tool for the prioritization of gene fusions, Bioinformatics 36 (10) (2020) 3248–3250.
[7] M. Lovino, G. Urgese, E. Macii, S. Di Cataldo, E. Ficarra, Predicting the oncogenic potential of gene fusions using convolutional neural networks, in: International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, Springer, 2018, pp. 277–284.
[8] A.R. Joyce, B.Ø. Palsson, The model organism as a system: integrating'omics' data sets, Nature reviews Molecular cell biology 7 (3) (2006) 198–210.
[9] C.D. Bustamante, M. Francisco, E.G. Burchard, Genomics for the world, Nature 475 (7355) (2011) 163–165.
[10] E.S. Lander, The new genomics: global views of biology, Science 274 (5287) (1996) 536–539.
[11] S. Fields, Proteomics in genomeland, Science 291 (5507) (2001) 1221–1224.
[12] Z. Wang, M. Gerstein, M. Snyder, Rna-seq: a revolutionary tool for transcriptomics, Nature reviews genetics 10 (1) (2009) 57–63.
[13] M. Esteller, Cancer epigenomics: Dna methylomes and histone-modification maps, Nature reviews genetics 8 (4) (2007) 286–298.
[14] C. Vilanova, M. Porcar, Are multi-omics enough?, Nature microbiology 1 (8) (2016) 1–2
[15] Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome biology 18 (1) (2017) 1–15.
[16] C. Meng, B. Kuster, A.C. Culhane, A.M. Gholami, A multivariate approach to the integration of multi-omics datasets, BMC bioinformatics 15 (1) (2014) 162.
[17] C. Meng, O.A. Zeleznik, G.G. Thallinger, B. Kuster, A.M. Gholami, A.C. Culhane, Dimension reduction techniques for the integrative analysis of multi-omics data, Briefings in bioinformatics 17 (4) (2016) 628–641.
[18] K.-L. Huang, S. Li, P. Mertins, S. Cao, H.P. Gunawardena, K.V. Ruggles, D. Mani, K.R. Clauser, M. Tanioka, J. Usary, et al., Proteogenomic integration reveals therapeutic targets in breast cancer xenografts, Nature communications 8 (1) (2017) 1–17.
[19] D.J. Clark, S.M. Dhanasekaran, F. Petralia, J. Pan, X. Song, Y. Hu, F. da Veiga Leprevost, B. Reva, T.-S.M. Lih, H.-Y. Chang, et al., Integrated proteogenomic characterization of clear cell renal cell carcinoma, Cell 179 (4) (2019) 964–983.
[20] A. Nakorchevsky, J.A. Hewel, S.M. Kurian, T.S. Mondala, D. Campbell, S.R. Head, C.L. Marsh, J.R. Yates, D.R. Salomon, Molecular mechanisms of chronic kidney transplant rejection via large-scale proteogenomic analysis of tissue biopsies, Journal of the American Society of Nephrology 21 (2) (2010) 362–373.
[21] P. Mertins, D. Mani, K.V. Ruggles, M.A. Gillette, K.R. Clauser, P. Wang, X. Wang, J.W. Qiao, S. Cao, F. Petralia, et al., Proteogenomics connects somatic mutations to signalling in breast cancer, Nature 534 (7605) (2016) 55–62.
[22] A. Forget, L. Martignetti, S. Puget, L. Calzone, S. Brabetz, D. Picard, A. Montagud, S. Liva, A. Sta, F. Dingli, et al., Aberrant erbb4-src signaling as a hallmark of group 4 medulloblastoma revealed by integrative phosphoproteomic profiling, Cancer cell 34 (3) (2018) 379–395.
[23] S. Rivero-Hinojosa, M. Grant, A. Panigrahi, H. Zhang, V. Caisova, C. Bollard, B. Rood, Abstract a23: Proteogenomic discovery of novel tumor proteins as neoantigens for personalized t-cell immunotherapy in pediatric medulloblastoma (2020)..
[24] I. Roberti, M. Lovino, S. Di Cataldo, E. Ficarra, G. Urgese, Exploiting gene expression profiles for the automated prediction of connectivity between brain regions, International journal of molecular sciences 20 (8) (2019) 2035.
[25] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al., The ensembl genome database project, Nucleic acids research 30 (1) (2002) 38–41.
[26] P. Legrain, R. Aebersold, A. Archakov, A. Bairoch, K. Bala, L. Beretta, J. Bergeron, C.H. Borchers, G.L. Corthals, C.E. Costello, et al., The human proteome project: current state and future direction, Molecular & cellular proteomics 10 (7) (2011).
[27] M.A. Jensen, V. Ferretti, R.L. Grossman, L.M. Staudt, The nci genomic data commons as an engine for precision medicine, Blood 130 (4) (2017) 453–459.
[28] Z. Zhang, K. Hernandez, J. Savage, S. Li, D. Miller, S. Agrawal, F. Ortuno, L.M. Staudt, A. Heath, R.L. Grossman, Uniform genomic data analysis in the nci genomic data commons, Nature communications 12 (1) (2021) 1–11.
[29] J. Zhang, J. Baran, A. Cros, J.M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, et al., International cancer genome consortium data portal-a one-stop shop for cancer genomics data, Database 2011 (2011).
[30] I.C.G. Consortium, et al., International network of cancer genome projects, Nature 464 (7291) (2010) 993..
[31] J.R. Whiteaker, G.N. Halusa, A.N. Hoofnagle, V. Sharma, B. MacLean, P. Yan, J.A. Wrobel, J. Kennedy, D. Mani, L.J. Zimmerman, et al., Cptac assay portal: a repository of targeted proteomic assays, Nature methods 11 (7) (2014) 703–704.
[32] G. Nicora, F. Vitali, A. Dagliati, N. Geifman, R. Bellazzi, Integrated multi-omics analyses in oncology: A review of machine learning methods and tools, Frontiers in oncology 10 (2020) 1030.
[33] T. Ching, X. Zhu, L.X. Garmire, Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data, PLoS computational biology 14 (4) (2018) e1006076.
[34] S.P. Couvillion, Y. Zhu, G. Nagy, J.N. Adkins, C. Ansong, R.S. Renslow, P.D. Piehowski, Y.M. Ibrahim, R.T. Kelly, T.O. Metz, New mass spectrometry technologies contributing towards comprehensive and high throughput omics analyses of single cells, Analyst 144 (3) (2019) 794–807.
[35] J. Ovesná, O. Slabý, O. Toussaint, M. Kodíček, P. Maršík, V. Pouchová, T. Vaněk, High throughput 'omics' approaches to assess the effects of phytochemicals in human health studies, British Journal of Nutrition 99 (E-S1) (2008) ES127–ES134.
[36] G. Judes, K. Rifaï, M. Daures, L. Dubois, Y.-J. Bignon, F. Penault-Llorca, D. Bernard-Gallon, High-throughput omics technologies: New tools for the study of triple-negative breast cancer, Cancer letters 382 (1) (2016) 77–85.
[37] N.P. Long, S. Park, N.H. Anh, T.D. Nghi, S.J. Yoon, J.H. Park, J. Lim, S.W. Kwon, High-throughput omics and statistical learning integration for the discovery and validation of novel diagnostic signatures in colorectal cancer, International journal of molecular sciences 20 (2) (2019) 296.
[38] L. Dalton, V. Ballarin, M. Brun, Clustering algorithms: on learning, validation, performance, and applications to genomics, Current genomics 10 (6) (2009) 430–445.
[39] N. Altman, M. Krzywinski, The curse(s) of dimensionality, Nature Methods 15 (6) (2018) 397, https://doi.org/10.1038/s41592-018-0013-3.
[40] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, L. Milanesi, BMethods for the integration of multi-omics data: mathematical aspects, BMC Bioinformatics 17 (2) (2016) S15, https://doi.org/10.1186/s12859-015-0857-9.

[41] B. Palsson, K. Zengler, The challenges of integrating multi-omic data sets, Nature Chemical Biology 6 (11) (2010) 787–789, https://doi.org/10.1038/nchembio.462. http://www.nature.com/articles/nchembio.462.
[42] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, Information Fusion 38 (2017) 43–54, https://doi.org/10.1016/j.inffus.2017.02.007. https://linkinghub.elsevier.com/retrieve/pii/S1566253516302032.
[43] I. Subramanian, S. Verma, S. Kumar, A. Jere, K. Anamika, Multi-omics Data Integration, Interpretation, and Its Application, Bioinformatics and Biology Insights 14 (Jan. 2020), https://doi.org/10.1177/1177932219899051, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7003173/.
[44] L. Cantini, P. Zakeri, C. Hernandez, A. Naldi, D. Thieffry, E. Remy, A. Baudot, Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer, Nature communications 12 (1) (2021) 1–12.
[45] B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nature methods 11 (3) (2014) 333.
[46] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering, in: Advances in neural information processing systems, 2011, pp. 1413–1421..
[47] A. Kumar, H. Daumé, A co-training approach for multi-view spectral clustering, in, in: Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 393–400.
[48] Y. Li, F. Nie, H. Huang, J. Huang, Large-scale multi-view spectral clustering via bipartite graph, in, in: Proceedings of the twenty-ninth AAAI conference on artificial intelligence, 2015, pp. 2750–2756.
[49] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition, in: Proceedings of the twenty-eighth AAAI conference on artificial intelligence, 2014, pp. 2149–2155.
[50] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on pattern analysis and machine intelligence 22 (8) (2000) 888–905.
[51] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing 17 (4) (2007) 395–416.
[52] N. Rappoport, R. Shamir, Nemo: Cancer subtyping by integration of partial multi-omic data, Bioinformatics 35 (18) (2019) 3348–3356.
[53] H. Nguyen, S. Shrestha, T. Nguyen, Pinsplus: Clustering algorithm for data integration and disease subtyping, CRAN R package (2018).
[54] H. Nguyen, S. Shrestha, S. Draghici, T. Nguyen, Pinsplus: a tool for tumor subtype discovery in integrated genomic data, Bioinformatics 35 (16) (2019) 2843–2846.
[55] E.F. Lock, K.A. Hoadley, J.S. Marron, A.B. Nobel, Joint and individual variation explained (jive) for integrated analysis of multiple data types, The annals of applied statistics 7 (1) (2013) 523.
[56] I.T. Jolliffe, Principal components in regression analysis, Principal component analysis (2002) 167–198.
[57] M.E. Wall, A. Rechtsteiner, L.M. Rocha, Singular value decomposition and principal component analysis, in: A practical approach to microarray data analysis, Springer, 2003, pp. 91–109..
[58] K. Pearson, On lines of closes fit to system of points in space, london, e dinb, Dublin Philos. Mag. J. Sci 2 (1901) 559–572.
[59] H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of educational psychology 24 (6) (1933) 417.
[60] A. Tenenhaus, M. Tenenhaus, Regularized generalized canonical correlation analysis, Psychometrika 76 (2) (2011) 257.
[61] H.D. Vinod, Canonical ridge and econometrics of joint production, Journal of econometrics 4 (2) (1976) 147–166.
[62] S.E. Leurgans, R.A. Moyeed, B.W. Silverman, Canonical correlation analysis when the data are curves, Journal of the Royal Statistical Society: Series B (Methodological) 55 (3) (1993) 725–740.
[63] A.E. Teschendorff, H. Jing, D.S. Paul, J. Virta, K. Nordhausen, Tensorial blind source separation for improved analysis of multi-omic data, Genome biology 19 (1) (2018) 76.
[64] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J.C. Marioni, F. Buettner, W. Huber, O. Stegle, Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets, Molecular systems biology 14 (6) (2018) e8124.
[65] R. De Vito, R. Bellio, L. Trippa, G. Parmigiani, Multi-study factor analysis, Biometrics 75 (1) (2019) 337–346.
[66] P. Chalise, B.L. Fridley, Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm, PloS one 12 (5) (2017) e0176278.
[67] P. Bady, S. Dolédec, B. Dumont, J.-F. Fruget, Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities, Comptes rendus biologies 327 (1) (2004) 29–36.
[68] M. Žitnik, B. Zupan, Data fusion by matrix factorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (1) (2015) 41–53, https://doi.org/10.1109/TPAMI.2014.2343973.
[69] W.J. Ewens, G.R. Grant, Statistical methods in bioinformatics: an introduction, Springer Science & Business Media, 2006.
[70] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017)..
[71] D.J. Wilkinson, Bayesian methods in bioinformatics and computational systems biology, Briefings in bioinformatics 8 (2) (2007) 109–116.
[72] R. Shen, A.B. Olshen, M. Ladanyi, Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis, Bioinformatics 25 (22) (2009) 2906–2912.
[73] Q. Mo, S. Wang, V.E. Seshan, A.B. Olshen, N. Schultz, C. Sander, R.S. Powers, M. Ladanyi, R. Shen, Pattern discovery and cancer gene identification in integrated cancer genomic data, Proceedings of the National Academy of Sciences 110 (11) (2013) 4245–4250.
[74] Q. Mo, R. Shen, C. Guo, M. Vannucci, K.S. Chan, S.G. Hilsenbeck, A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data, Biostatistics 19 (1) (2018) 71–86.
[75] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society: Series B (Methodological) 39 (1) (1977) 1–22.
[76] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N. De Freitas, Taking the human out of the loop: A review of bayesian optimization, Proceedings of the IEEE 104 (1) (2015) 148–175.
[77] S. Lloyd, Least squares quantization in pcm, IEEE transactions on information theory 28 (2) (1982) 129–137.
[78] C.J. Vaske, S.C. Benz, J.Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, J.M. Stuart, Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm, Bioinformatics 26 (12) (2010) i237–i245.
[79] F.R. Kschischang, B.J. Frey, H.-A. Loeliger, Factor graphs and the sum-product algorithm, IEEE Transactions on information theory 47 (2) (2001) 498–519.
[80] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proceedings of the National Academy of Sciences 95 (25) (1998) 14863–14868.
[81] D. Wu, D. Wang, M.Q. Zhang, J. Gu, Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification, BMC genomics 16 (1) (2015) 1022.
[82] Z. Wei, Y. Zhang, W. Weng, J. Chen, H. Cai, Survey and comparative assessments of computational multi-omics integrative methods with multiple regulatory networks identifying distinct tumor compositions across pan-cancer data sets, Briefings in Bioinformatics (2020).
[83] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (3) (1936) 211–218.
[84] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, Oakland, CA, USA, 1967, pp. 281–297..
[85] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics (1987), https://doi.org/10.1016/0377-0427(87)90125-7.
[86] P. Kirk, J.E. Griffin, R.S. Savage, Z. Ghahramani, D.L. Wild, Bayesian correlated clustering to integrate multiple datasets, Bioinformatics 28 (24) (2012) 3290–3297.
[87] E.F. Lock, D.B. Dunson, Bayesian consensus clustering, Bioinformatics 29 (20) (2013) 2610–2616.
[88] E. Gabasova, J. Reid, L. Wernisch, Clusternomics: Integrative context-dependent clustering for heterogeneous datasets, PLoS computational biology 13 (10) (2017) e1005781.
[89] T.S. Ferguson, A bayesian analysis of some nonparametric problems, The annals of statistics (1973) 209–230.
[90] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, science 313 (5786) (2006) 504–507.
[91] Y. Chen, Y. Li, R. Narayan, A. Subramanian, X. Xie, Gene expression inference with deep learning, Bioinformatics 32 (12) (2016) 1832–1839.
[92] X. Liu, L. Song, S. Liu, Y. Zhang, A review of deep-learning-based medical image segmentation methods, Sustainability 13 (3) (2021) 1224.
[93] M.P. McBee, O.A. Awan, A.T. Colucci, C.W. Ghobadi, N. Kadom, A.P. Kansagra, S. Tridandapani, W.F. Auffermann, Deep learning in radiology, Academic radiology 25 (11) (2018) 1472–1480.
[94] S. Sukhbaatar, R. Fergus, Learning from noisy labels with deep neural networks, arXiv preprint arXiv:1406.2080 2 (3) (2014) 4..
[95] S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, Briefings in bioinformatics 18 (5) (2017) 851–869.
[96] Q. Zhang, S. Wang, Z. Chen, Y. He, Q. Liu, D.-S. Huang, Locating transcription factor binding sites by fully convolutional neural network, Briefings in Bioinformatics (2021).
[97] S. Wang, Q. Zhang, Z. Shen, Y. He, Z.-H. Chen, J. Li, D.-S. Huang, Predicting transcription factor binding sites using dna shape features based on shared hybrid deep learning architecture, Molecular Therapy-Nucleic Acids 24 (2021) 154–163.
[98] A. Trabelsi, M. Chaabane, A. Ben-Hur, Comprehensive evaluation of deep learning architectures for prediction of dna/rna sequence binding specificities, Bioinformatics 35 (14) (2019) i269–i277.
[99] Z. Shen, W. Bao, D.-S. Huang, Recurrent neural network for predicting transcription factor binding sites, Scientific reports 8 (1) (2018) 1–10.
[100] Y. He, Z. Shen, Q. Zhang, S. Wang, D.-S. Huang, A survey on deep learning in dna/rna motif mining, Briefings in Bioinformatics 22 (4) (2021) bbaa229.
[101] B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of dna-and rna-binding proteins by deep learning, Nature biotechnology 33 (8) (2015) 831–838.
[102] J. Zhou, O.G. Troyanskaya, Predicting effects of noncoding variants with deep learning–based sequence model, Nature methods 12 (10) (2015) 931–934.
[103] D. Quang, X. Xie, Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences, Nucleic acids research 44 (11) (2016) e107.
[104] M. Lai, Deep learning for medical image segmentation, arXiv preprint arXiv:1505.02000 (2015)..

[105] A.Y.C. Florez, L. Scabora, S. Amer-Yahia, J.F.R. Júnior, Augmentation techniques for sequential clinical data to improve deep learning prediction techniques IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE 2020 (2020) 597–602.
[106] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, in: International conference on machine learning, 2015, pp. 1083–1092.
[107] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: ICML, 2011..
[108] H. Hotelling, Relations between two sets of variates, in: Breakthroughs in statistics, Springer, 1992, pp. 162–190..
[109] G. Cirrincione, P. Barbiero, G. Ciravegna, V. Randazzo, Gradient-based competitive learning: Theory, arXiv preprint arXiv:2009.02799 (2020)..
[110] P.L. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis, International Journal of Neural Systems 10 (05) (2000) 365–377.
[111] S. Akaho, A kernel method for canonical correlation analysis, arXiv preprint cs/0609071 (2006)..
[112] A. Vinokourov, N. Cristianini, J. Shawe-Taylor, Inferring a semantic representation of text via cross-language correlation analysis, Advances in neural information processing systems 15 (2002) 1497–1504.
[113] P. Dhillon, D.P. Foster, L. Ungar, Multi-view learning of word embeddings via cca, Advances in neural information processing systems 24 (2011) 199–207.
[114] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: International conference on machine learning, PMLR, 2013, pp. 1247–1255..
[115] A. Lu, W. Wang, M. Bansal, K. Gimpel, K. Livescu, Deep multilingual correlation for improved word embeddings, in, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 250–256.
[116] M.A. Kramer, Nonlinear principal component analysis using autoassociative neural networks, AIChE journal 37 (2) (1991) 233–243.
[117] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, Vol. 1, MIT press Cambridge, 2016.
[118] S. Lee, J. Jo, Information flows of diverse autoencoders, Entropy 23 (7) (2021) 862.
[119] Z. Ghahramani, Information theory, Encyclopedia of, Cognitive Science (2006).
[120] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, L. Bottou, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, Journal of machine learning research 11 (12) (2010).
[121] M. Borga, Canonical correlation: a tutorial, On line tutorial http://people. imt, liu. se/magnus/cca 4 (5) (2001).
[122] P. Barbiero, G. Ciravegna, V. Randazzo, E. Pasero, G. Cirrincione, Topological gradient-based competitive learning, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.
[123] P. Chalise, R. Raghavan, B.L. Fridley, Intersim: Simulation tool for multiple integrative 'omic datasets', Computer methods and programs in biomedicine 128 (2016) 69–74.
[124] National Cancer Institute, Gdc data portal, https://portal.gdc.cancer.gov/, last accessed on 2020-06-14..
[125] S. Anders, P.T. Pyl, W. Huber, Htseq-a python framework to work with high-throughput sequencing data, Bioinformatics 31 (2) (2015) 166–169.
[126] A. Chu, G. Robertson, D. Brooks, A.J. Mungall, I. Birol, R. Coope, Y. Ma, S. Jones, M.A. Marra, Large-scale profiling of micrornas for the cancer genome atlas, Nucleic acids research 44 (1) (2016) e3.
[127] K. Tomczak, P. Czerwińska, M. Wiznerowicz, The cancer genome atlas (tcga): an immeasurable source of knowledge, Contemporary oncology 19 (1A) (2015) A68.
[128] B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nature methods 11 (3) (2014) 333.
[129] P. Barbiero, M. Lovino, M. Siviero, G. Ciravegna, V. Randazzo, E. Ficarra, G. Cirrincione, Unsupervised multi-omic data fusion: The neural graph learning network, in, International Conference on Intelligent Computing, Springer (2020) 172–182.
[130] G. Cirrincione, G. Ciravegna, P. Barbiero, V. Randazzo, E. Pasero, The gh-exin neural network for hierarchical clustering, Neural Networks 121 (2020) 57–73.
[131] D.L. Davies, D.W. Bouldin, A Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence (1979), https://doi.org/10.1109/TPAMI.1979.4766909.
[132] A. Paviglianiti, V. Randazzo, E. Pasero, A. Vallan, Noninvasive arterial blood pressure estimation using abpnet and vital-ecg, in: 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), IEEE, 2020, pp. 1–5..
[133] W. Huber, A. Von Heydebreck, H. Sültmann, A. Poustka, M. Vingron, Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics 18 (suppl_1) (2002) S96–S104.
[134] S. Anders, W. Huber, Differential expression of rna-seq data at the gene level–the deseq package, 10, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, 2012, f1000research.
[135] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for rna-seq data with deseq2, Genome biology 15 (12) (2014) 550.
[136] P. Demartines, J. Hérault, Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets, IEEE Transactions on neural networks 8 (1) (1997) 148–154.
[137] J. Sun, C. Fyfe, M.K. Crowe, Curvilinear component analysis and bregman divergences, ESANN, in, 2010.
[138] M. Lovino, G. Bontempo, G. Cirrincione, E. Ficarra, Multi-omics classification on kidney samples exploiting uncertainty-aware models, in: International Conference on Intelligent Computing, Springer, 2020, pp. 32–42.
[139] C.C. Aggarwal, A. Hinneburg, D.A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: International conference on database theory, Springer, 2001, pp. 420–434.
[140] A. Rajaraman, J.D. Ullman, Mining of massive datasets, Cambridge University Press, 2011.

Marta Lovino is currently Postdoctoral Research Fellow in the EDA group. She received her B.Sc. and M.Sc. degrees in Biomedical Engineering at Politecnico di Torino in 2015 and 2017. Then, in 2021, Marta received the Ph.D. in Computer and Control Engineering cum laude offered by Politecnico di Torino. Since 2017, when she joined the EDA group as a Ph.D. candidate, she has worked on software tools and pipelines to analyze genomics and proteomics data. In 2020, she was a visiting researcher for six months at the Institute Curie Center in Paris, in the Signaling in development and brain tumors lab. Her primary research interests are cancer prognosis prediction, gene expression analysis, miRNA target prediction, gene regulatory networks, gene fusions, and multi-omics data integration.

Vincenzo Randazzo is a researcher at the Politecnico di Torino on the topic "Neural networks for telemedicine". He got his PhD cum laude in electrical, electronics and communications engineering at the Politecnico di Torino with a final thesis on "Novel Neural Approaches to Data Topology Analysis and Telemedicine". He graduated with honors in Computer Engineering from the University of Palermo. He has published on Neural Networks (Elsevier), Journal of Nephrology (Springer), IEEE Access, Electronics (MDPI); and has several chapters of Springer books to his credit. Furthermore, his works have been accepted at SAS2018, IJCNN2018, WIRN2018, WIRN2019, MEMEA2019, I2MTC2020, IJCNN2020. His current research interests include: neural networks, data analysis and intrinsic dimension estimation, pattern recognition, diagnosis of nonlinear systems, medical and biomedical applications. He also won the "Technology transfer: prototyping and business development activities in entrepreneurship start-up programs" research fellowship to promote the culture of innovation and entrepreneurship among young talents. He is the IEEE Young Professionals Vice-Chair for Italy section and the former treasurer of Politecnico di Torino IEEE Student Branch.



Gabriele Ciravegna is a PhD student at the University of Siena since 2018 under the supervision of Professor Marco Gori. In 2018 he received the master's degree in Computer Engineering with honors at the Polytechnic University of Turin. He has always been interested in the machine learning field. Nowadays, he is focused on overcoming the intrinsic limits of machine learning and neural networks, especially in the context of Explainable AI. He presented his works in several international venues such as AAAI, IJCAI, IJCNN. He also serves as reviewer in conferences and journals that are about Neural Networks, such as IEEE Transactions on Neural Networks and Learning Systems. Besides machine learning, he also likes football, volleyball and playing the piano.

Pietro Barbiero is a Ph.D. student and research assistant in the Artificial Intelligence group at the University of Cambridge (UK). He is an AI engineer and member of the founding team of Bactell Inc. He graduated in Mathematics and Engineering at the Polytechnic University of Turin. His current research interests include explainable AI applied to computational biology and precision medicine.

Elisa Ficarra got a PhD in Systems and Computer Science at Dept. of Control and Computer Engineering - Politecnico di Torino, Italy, in June 2006. She worked from 2015 to 2021 as Associate Professor in the same department. Currently, she is an Associate Professor at the Dept. of Engineering "Enzo Ferrari" in the University of Modena and Reggio Emilia, Italy. Prof. Ficarra was an invited scientist at the Computer Science Dept. of Stanford University (California, USA) in 2002 and a visiting scientist at the Integrated System Laboratory of EPFL (Lausanne, CH) in 2005 for a research project on bioimaging and functional genomics. Prof. Ficarra acted as scientific coordinator for Politecnico di Torino, and currently for the University of Modena and Reggio Emilia, of various European and Italian funded projects in the field of bioinformatics and bioimage processing focusing on genomics, fluorescence and histological imaging, and molecular data analysis exploiting machine learning/deep learning systems and data mining techniques. She was Associate Editor for IEEE Transactions on Information Technology in Biomedicine from 2011 to 2013, and she is a member of TPCs of several IEEE/ACM conferences. Prof. Ficarra's research activity led to more than 115 publications in journals (more than 50) and peer-reviewed conference proceedings.

Dr. Giansalvo Cirrincione received the <Laurea> degree in Electrical Engineering from the Politecnico of Turin in Italy in 1991 and the PhD degree (with honors) from the "Laboratoire d'Informatique et Signaux" of the 'Institut National Polytechnique de Grenoble (INPG) in 1998. In 1999 he was a post-doc in the Department SISTA, Leuven University, Belgium. In 2008 he received the HdR = Habilitation á diriger des recherché (Title demanded for becoming Full Professor in France, it is a Professorial Thesis) Since 2000 he has been an Associate Professor (Maître de Conférence) at The Department of Electrical Engineering and Informatics Engineering (GEII) at the Professional University Institute (IUP) and member of the Laboratory of Innovative Technologies (LTI), Amiens (tenured in September 2002) of the University of Picardie "Jules Verne". He has published over 150 papers in highly ranked international journals, international conferences, His current research interests include neural networks, deep learning, data analysis, computer vision, control and diagnosis of electrical systems, power electronics and electrical drives.