Briefings in Bioinformatics, 2022, pp. 1–24

doi: DOI HERE

Heterogeneous data integration methods for patient similarity networks

HETEROGENEOUS DATA INTEGRATION METHODS FOR PATIENT SIMILARITY NETWORKS

Heterogeneous data integration methods for patient similarity networks

Jessica Gliozzo,[1,2,5] Marco Mesiti,[1,5] Marco Notaro,[1,5] Alessandro Petrini,[1,5] Alex Patak,[2] Antonio Puertas-Gallardo,[2] Alberto Paccanaro,[3,4] Giorgio Valentini[1,5,6,7] and Elena Casiraghi[1,5] *

[1]AnacletoLab - Computer Science Department, Università degli Studi di Milano, Via Celoria 18, 20135, Milan, Italy, [2]European Commission, Joint Research Centre (JRC), Ispra (VA), Italy, [3]Department of Computer Science, Royal Holloway, University of London, Egham, TW20 0EX UK, [4]School of Applied Mathematics (EMAp), Fundação Getúlio Vargas, Rio de Janeiro Brazil, [5]CINI, Infolife National Laboratory, Roma, Italy, [6]DSRC UNIMI, Data Science Research Center, Milano, 20135, Italy and [7]ELLIS, European Laboratory for Learning and Intelligent Systems, Berlin, Germany

[*]Corresponding author: Elena Casiraghi - email: elena.casiraghi@unimi.it

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Patient similarity networks (PSNs), where patients are represented as nodes and their similarities as weighted edges, are being increasingly used in clinical research. These networks provide an insightful summary of the relationships among patients and can be exploited by inductive or transductive learning algorithms for the prediction of patient outcome, phenotype, and disease risk. PSNs can also be easily visualized, thus offering a natural way to inspect complex heterogeneous patient data, and providing some level of explainability of the predictions obtained by machine learning algorithms. The advent of high-throughput technologies, enabling us to acquire high-dimensional views of the same patients (e.g. omics data, laboratory data, imaging data) calls for the development of data fusion techniques for PSNs in order to leverage this rich heterogeneous information. In this article, we review existing methods for integrating multiple biomedical data views to construct PSNs, together with the different patient similarity measures that have been proposed. We also review methods that have appeared in the machine learning literature but have not yet been applied to PSNs, thus providing a resource to navigate the vast machine learning literature existing on this topic. In particular, we focus on methods that could be used to integrate very heterogeneous datasets, including multi-omics data as well as data derived from clinical information and medical imaging.

Key words: patient similarity networks, biomedical applications, multimodal data, data fusion

Introduction

In the last decades, medical research has begun to move from a population-based perspective to a personalized one, often referred to as Precision Medicine, where patients' biomedical characteristics are leveraged for diagnosis, prognosis and choice of appropriate treatment [1, 2]. In this context, it is widely accepted that if two patients share similar clinical variables and omics profiles, their clinical outcomes should also be similar. Pairwise similarities between patients have a natural representation as graphs - Patient Similarity Networks (PSN) - where nodes represent patients and edges represent the similarity between patients calculated using their clinical and/or biomolecular features. In this framework unsupervised clustering methods and supervised classification models that leverage similarities between patients have been successfully applied to stratify patients and to predict their phenotype or clinical outcome [3, 4, 5, 6, 7, 8]. Representing data as graphs provides several advantages, including interpretability and privacy [9], as patient specific information cannot be recovered from the similarity measures.

The increasing availability of high-throughput technologies able to generate high-dimensional, distributed biomedical datasets, ranging from multi-omics [8] to imaging [10], clinical and demographic data [11], calls for approaches to mine and aggregate salient information [12] with the ultimate aim of building PSNs integrating such diverse datasets. However, the majority of PSNs that have been proposed are built using only one source of information. At the same time, several methods that can integrate

heterogenous sources of information into graph structures have appeared in the past decades in the biomedical and machine learning literature.

In this article, we review existing methods for integrating multiple biomedical data views to construct patient similarity networks. Since the type of data being integrated and the specific integration method must be coupled with an appropriate choice of similarity measure, we will also discuss different similarity measures. Importantly, this paper also reviews methods for integrating information into graph structures that appeared in the machine learning literature but have not yet been used for PSNs. We believe that this will be beneficial for the reader, providing a resource to navigate the vast machine learning literature existing on this topic, and possibly inspire the use and development of novel techniques of data integration for PSNs. Moreover, unlike earlier reviews (see e.g. [13, 14, 15, 8, 16]), we focus on methods that may be used for patients' classification and clustering that integrate not only multi-omics data, but also clinical and image sources.

We propose a taxonomy that groups existing methods for building PSNs into three main categories. *PSN-fusion methods* [3, 17, 6] build different PSNs, one for each data source, that are then fused together into a single PSN. *Input data-fusion* methods [18, 19, 20, 21] combine the different data sources into a single dataset that is then used for building a single PSN. Finally, *Output-fusion methods* [22, 23, 24] build different PSNs, one for each data source, that are analysed separately, and results are then combined.

Other multimodal data-fusion surveys not specific for PSNs have been proposed in the bioinformatics field by adopting different taxonomies (schematized in Fig. 1, Appendix A). Some taxonomies focus on the type of multi-datasets being integrated, thus identifying *horizontal integration techniques* [25] (top of Fig. 1-yellow box) and *vertical integration techniques* [25] (top of Fig. 1-light blue box). While the former fuse *homogeneous multisets* [26], i.e. multimodal datasets where each view produces the same data type under different settings, the latter integrate the classic *heterogeneous* [26] multimodal datasets. Vertical integration techniques are further classified into methods applying a *hierarchical* (alias *multi-staged* [27]) *integration* flow, where ground knowledge about the relationships between the different views is considered during the integration, and methods applying a *parallel* (alias *meta-dimensional* [27]) *integration* flow (bottom of Fig. 1-red-dashed box), where each view is processed in a similar but independent way. Parallel integration methods are the most diffused in literature given their generalizability. For this reason several reviews concentrate solely on them and introduce taxonomies that distinguish, e.g., *model-agnostic* versus *model-dependent* methods [28], or exploit an *early-intermediate-late* taxonomy [29, 30, 31, 27, 32, 33] (described in detail in Appendix A).

Anyhow, each review paper focuses on different aspects of the multimodal data integration. For example, some works solely focus on integrative unsupervised clustering techniques [34] or supervised multi-omics prediction models [29, 35, 33], or survey data-fusion techniques that are either applied to multi-omics data [36, 26, 27, 25, 16], or that apply specific data-fusion techniques (e.g. integrative Bayesian models [37, 13] or multimodal neural networks [38]).

Unlike previous reviews, this work specifically focuses on integrative methods for PSN-based models integrating not only multi-omics data, but also clinical and imaging sources. Each method is critically described to highlight its main advantages and drawbacks, enabling the reader to select the most appropriate approach to answer her/his scientific questions.

Given a set of patients and their corresponding clinical and biomolecular features, the topology of the corresponding PSN depends crucially on how the similarity measure is calculated. Therefore, we begin describing the similarity measurement methods presented in the literature. Our taxonomy of existing methods for building PSNs is described in Sections PSN-fusion methods and Input data-fusion and output-fusion methods. Tables 3 - 8 summarize the most relevant methods we surveyed.

PSN construction

The construction of the PSN is a crucial step in PSN analysis models, whose effectiveness mainly depends on the available multimodal datasets from which samples are extracted and on the choice of the measure exploited for pairwise similarity computation between samples.

Several kinds of similarity measures have been adopted in literature for PSN construction: classic distance metrics tailored to the data type [39, 40]; kernel functions [41, 42] that substitute distance metrics; "kernels on graphs" methods [43]. In the remainder, we discuss their characteristics.

The usage of classic *(opportunely inverted) distances* or *similarity metrics* [39, 40] is often preferred when the data types are normalized and homogeneous. As an example, PSNs on continuous, normalized, data have been constructed by using the cosine similarity [44, 5], or the Euclidean [45] or Mahalanobis distance [45]; PSNs on discrete data types have been built by exploiting the Chi-squared distance [3, 6]; binary data has been handled by using the Jaccard distance [46] or many other distance measures (see [47] for a list of 76 metrics and measures specifically designed for binary data).

When data-blocks with heterogeneous and/or normalized variable types are available, more articulated schemas [48, 6] have been proposed to integrate different similarity metrics into a unique measure. As an example, in [48] the authors proposed a supervised Cox regression model to initially learn a weight for each variable; the learnt weights are then used to compute a similarity score as a weighted sum of individual similarities obtained on each feature by using standard metrics. In this way, different similarity metrics can be used on the different variables based on their type, and the influence of each variable to the global similarity score is weighted on the prediction (e.g. survival time when using a Cox regression model). On the other hand, when dealing with datasets composed by continuous non-normalized variable types, Pai et al. [6] propose computing the average of all the normalized similarities over each variable, where the normalization is essentially a min-max normalization.

When dealing with complex problems, literature works often rely on *Kernel functions* [49] for PSN computation. The rationale behind this choice is based on the assumption that point separability is often improved after a non-linear projection of points into a higher-dimensional space. Kernel functions are particularly appealing in this context since they express pairwise distances in a higher-dimensional space by directly using the (lower-dimensional) input samples, therefore avoiding the expensive explicit computation of a non-linear higher-dimensional mapping followed by pairwise similarity evaluation (using the well-known *kernel-trick*). Even in this case the choice of the kernel function must be tailored to the data type which is crucial to obtain reliable results. In this context, PSNs are often computed in literature methods working on biomedical data by using classic parametric *normalized linear kernels* [30, 50], *polynomial kernels*, or *Gaussian kernels* [51, 52], whose parameters are tuned to optimize performance. As an example, the prognostic approach presented in [30] obtains a set of unimodal PSNs by applying normalized linear kernels on each of the data-sources containing clinical and multi-omics data sets. In this case, the usage of the same kernel function on different sources is appropriate because they are characterized by the same data type (real-valued data type).

In a subsequent work [53], the same authors extend the data set by including categorical and integer data types; therefore, they substitute the linear kernels with a set of kernels tailored on each data type being processed. Of note, the kernels used in [30, 53] are always normalized. This is a crucial characteristic when integrating multiple kernels because

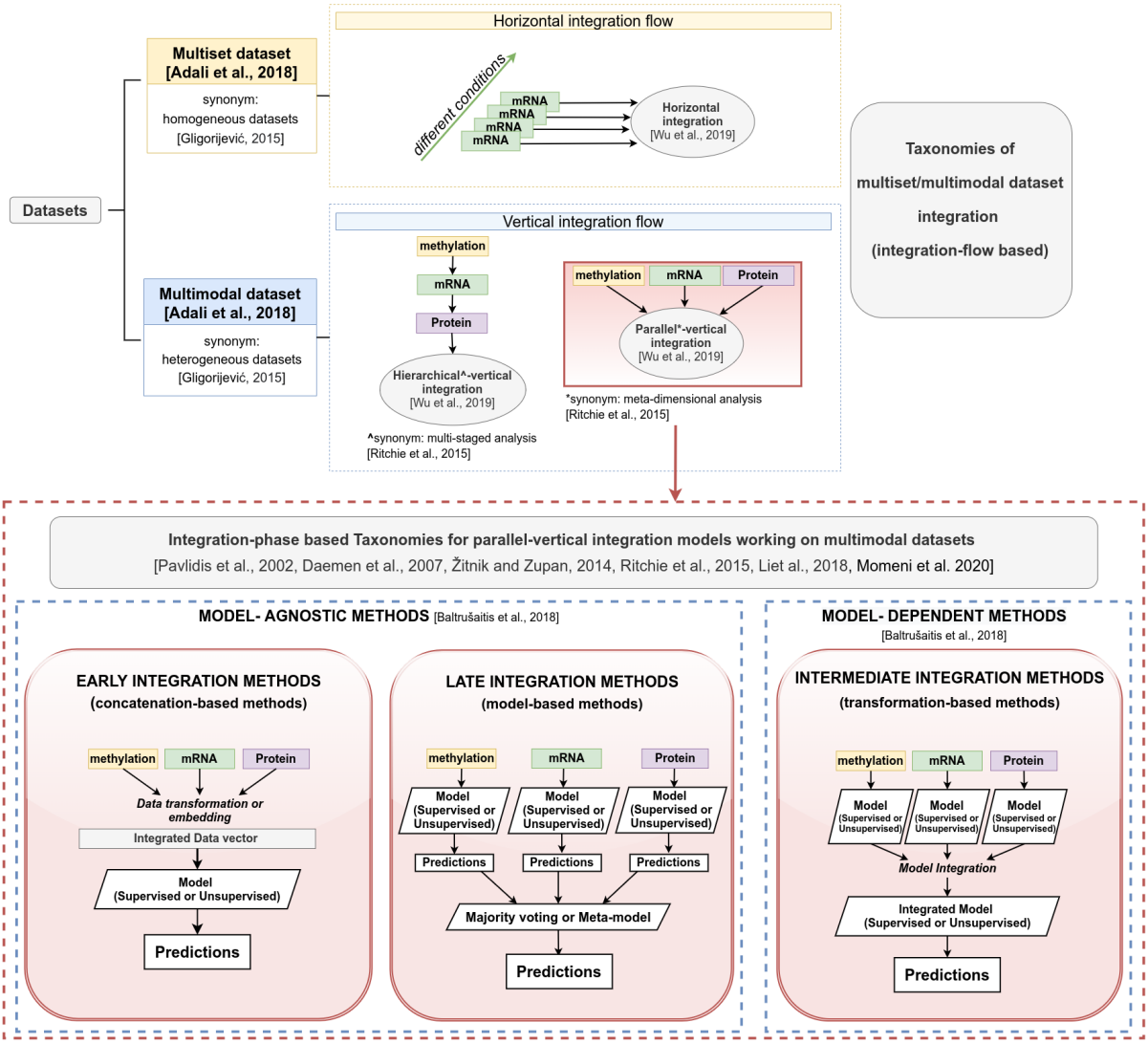


Fig. 1. Schema of the main taxonomies proposed in literature for categorizing multimodal integration methods. Considering the data integration flow, literature works identify two broad classes: *horizontal integration* approaches and *vertical integration* approaches. *Horizontal integration* approaches fuse *multisets* (i.e. datasets where each view is acquired by the same source under different conditions) by independently applying the same process on each view and then pooling the individual results. On the other hand, *vertical integration* approaches fuse *multimodal datasets* (i.e. datasets composed by semantically different views) through more complex techniques, further categorized as *hierarchical-vertical integration* methods and *parallel-vertical integration* techniques. The former fuse data views following a *hierarchy* driven by biological *a-priori* knowledge while the latter do not exploit knowledge-based dependencies between views. *Parallel-vertical integration* methods are the most diffused integration methods; they are further classified based on the phase when the data *integration-step* is performed w.r.t. the model construction (red-dashed box). Thus, methods are divided in (I) *early approaches*, which integrate the data types before model construction, (II) *late approaches*, which integrate the results of models independently built on each data view, and (III) *intermediate approaches* where intermediate models are obtained from each view and subsequently integrated. Of note, the latter class of approaches is more dependent on the exploited learning model, which is the reason why they have been also classified as *model-dependent* methods opposed to *model-agnostic* methods (blue-dashed boxes). We refer interested readers to Appendix A.

comparable kernel scales are obtained, therefore facilitating the kernel integration. Moreover, in the case of kernel-aggregation systems exploiting weighted averages of the unimodal kernels, normalization also improves the interpretability of the computed integration weights, the latest being directly related to the importance of their respective kernel [53].

A recent advance in the field of PSN analysis is provided by unsupervised methods that compute the PSN through the *scaled exponential Euclidean kernel* [3] and its modifications [54, 55]. They essentially apply a local normalization of the distance between a central node and any of its neighbors, so that distances are independent from the neighborhood scales. Their application in the context of unsupervised patient clustering through PSN analysis has obtained promising results [3] (see Section **SNF-based methods**).

Given its effectiveness, the scaled Euclidean distance has been extended in [54] to deal with heterogeneous data types containing continuous and boolean variables. More precisely, the similarity on boolean data is measured by using the weighted Hamming distance with weights computed by supervised approaches or pre-set based on existing knowledge. Further, in [55] the authors propose adopting the Chebyshev distance instead of the Euclidean distance.

Gliozzo et al. [7] extend to PSNs a previous *kernel-based* approach originally applied to the semi-supervised analysis of biomolecular networks

[56]. More precisely, the authors obtain promising outcome predictions on unimodal PSNs by firstly using the filtered Pearson correlation (by setting to zero all negative values) to measure similarities between unimodal gene expression profiles, and then applying a random walk kernel to strengthen high similarities while diminishing low ones. The neighborhoods identified in the obtained PSN are then used to compute a score for each patient, which is thresholded to obtain the desired classification. While unimodal PSNs are exploited in [7], the works proposed in [57] and [58] exploit random-walks to compute similarities in a multimodal setting.

To improve informativeness, Tables 1 and 2 sketch the similarity measures/ methods used for PSN construction by notable literature works exploiting multimodal datasets; for each paper we report the data types of the different data-sources exploited for the investigation, and the similarity measures/methods used for building the corresponding unimodal PSNs.

Even if a wide range of similarity computation methods has been proposed in literature, a consensus on which strategy performs better on specific data types and problems in the context of precision medicine is still lacking. Some tentative experiments have been conducted for determining the best performing strategies (see e.g. [59, 60]), but the lack of common benchmark datasets prevents an unbiased comparison of the different proposed approaches.

PSN-fusion methods

PSN-fusion methods have been specifically developed to process a set of unimodal PSNs and produce an integrated PSN. In Fig. 2 we sketch the generic workflow of the PSN-fusion methods. They start by building unimodal PSNs on each data source or data-type (Fig. 2-A). Mind that the choice of the similarity measure/kernel function used to build each PSN (Section PSN construction) is crucial for obtaining informative unimodal PSNs, which would otherwise hamper the achievement of successful results. Next, the aggregation of the unimodal PSNs (Fig. 2-B) is performed by either Multiple Kernel Learning methods (MKL, Section **MKL-based methods**, Table 3) which run optimization algorithms inherited from the machine learning field to find the optimal weights of an additive unimodal kernel aggregation, or approaches stemming from the seminal Similarity Network Fusion algorithm (SNF - [3], Section **SNF-based methods**, Table 4), which use different strategies to diffuse the similarity information both between neighboring nodes in each unimodal PSN and between corresponding nodes in different PSNs, or other network-based approaches (Section **Other PSN-fusion methods** and Table 5).

The integrated PSN may be finally used as input to unsupervised clustering methods aiming at, e.g., identifying patients' subtypes, or supervised classification methods predicting, e.g., patients' risk, prognosis, or outcome (Fig. 2-C).

MKL-based methods

Inheriting theories and algorithms from the machine learning fields, MKL methods [17, 65, 66, 64] view the unimodal PSNs as kernels and propose their optimal additive combination, as a weighted sum of the available unimodal kernels. In this context, "optimality" refers to either a *supervised* setting or an *unsupervised* one.

supervised MKL algorithms (e.g. simpleMKL [17]) exploit a supervised classifier model designed to work on the fused kernel. Supervision is guaranteed by the availability of a training set composed of samples whose labels are known. Such training set is used by the chosen supervised MKL method to solve a constrained optimization problem that finds the kernel weights and classifier hyper-parameters maximizing the classification accuracy on the training set. On the other side, *unsupervised MKL* methods make no use of labeled samples, but instead solve an optimization problem to find the weights that essentially lead to the maximum alignment between the integrated kernel and any of the input unimodal kernels.

Recent PSN-fusion methods exploiting a *supervised MKL* strategy are those presented by [30, 53, 64, 50, 67]. The work proposed in [50] designs specific kernels for each omic type in the TCGA cancer dataset and then computes the kernel weights by using the training set to optimize the fit of a Cox-survival model.

All the other works [30, 53, 64, 67] share the use of the kernelized Support Vector Machine (svm) classifiers [68], opportunely modified as defined in [17] and [66] to work on the kernel resulting from an optimal additive sum. In particular, the works proposed by Daemen et al. [30, 53] aggregate specific kernels on each clinical data type and uses a classic svm optimization strategy to derive the optimal weights, while the works proposed in [64] and [67] use the easyMKL algorithm to optimize an svm aggregating multiple kernels defined over multimodal datasets also including opportunely coded imaging sources. More precisely, in [64] authors use the same Gaussian kernels to process both the real CerebroSpinal Fluid (CSF) biomarkers features and the shape and texture features extracted to code Magnetic Resonance Images (MRI). On the other side, the work proposed in [67] improves upon the work presented in [69] and defines specific kernels for the multi-omics data from the TCGA cancer dataset and for the features automatically extracted from histopathological images (see Table 3). The effectiveness of the simpleMKL strategy is witnessed by its several extensions (easyMKL [70], SEMKL [71], SpicyMKL [72]).

As expected, our literature search highlighted that SVMs are the most widely used base-learner models in conjunction with MKL in the context of biomedical predictions; however, some authors have also presented MKL methods using Multiple Kernel Fisher Discriminant Analysis (MK-FDA [73]) or Kernel Regularized Discriminant Analysis (KRDA, [74]) as base learners where the single kernel is substituted by multiple kernels. Though these strategies have not been applied on patients' data, their promising results on the protein sub-cellular localization prediction task [73, 75] suggest they could be good options for developing a multimodal PSN analysis task.

Unsupervised MKL approaches are described in the works of [76, 77, 52]. The regularized MKL with Locality Preserving Projection algorithm (rMKL-LPP [76]) is an unsupervised, regularized MKL-based clustering approach for the identification of cancer subtypes from multi-omics data. It builds upon the MKL-DR model proposed in [78] to constrain the optimization problem by handling the "small-sample-size" problems caused by the high dimensionality of the input data-sources and exploits the theories at the base of the locality preserving projection algorithm [79] to find the integrated kernel in a lower-dimensional space that maintains the local neighborhoods relationships. In other words, the model minimizes a function that allows finding both the hyper-parameters of the multiple kernels and their combination weights so that patients that are similar according to "many" input sources (kernels) remain neighbors in the integrated kernel. Further, to avoid restricting the usage of only one kernel per data-source or data-type, authors add a constrained regularization that avoids overfitting, so that multiple kernels can be used for each source without risking to over-fit the data. Similar topological constraints are used by [52] to compute kernel weights such that the resulting integrated kernel maintains the neighborhood-relationship described above, and at same time maximizes the alignment (similarity) to all the input kernels.

By contrast, Liu et al. [77] leverage the standard kernel k-means clustering [80], which applies k-means in the kernel space, to a *multiple kernel k-means clustering* (*MKKM*) that considers the relationships between all the input kernels. The optimal clusters are found by minimizing a loss that measures the intra-class sample distance as a function of the cluster assignment matrix and the kernel weights. However, differently from other

Table 1. Similarity measures/methods used in literature to build PSNs. For notable works in literature the table reports: the reference of the literature work presenting a multimodal PSN analysis method (column *Reference*), the data types (column *Data type*) of the different sources (column *Data*) exploiting for the investigation, and the similarity measures/methods exploited for building the unimodal PSNs.

Reference	Data type	Data	Similarity measure/method
[46]	binary	ICD-9 diagnosis code	Jaccard similarity
[44]	continuous, categorical, discrete	clinical data	cosine similarity
[5]	continuous categorical, discrete	clinical data	cosine similarity
[61]	continuous	mRNA, PPI	Pearson correlation
[7]	continuous	mRNA	Pearson correlation
[6]	continuous	clinical variables individual gene genes in pathways/networks	mean of normalized difference normalized difference Pearson correlation
	discrete	categorical-ordinal variable (e.g. tumour stage) unbinned counts (e.g. mutation data) matrix scores (e.g. response to questionnaire)	normalized difference shared incidence in a grouped unit chi-square distance
[3]	continuous discrete binary	mRNA, miRNA, DNA methylation	scaled exponential kernel of Euclidean distance chi-squared distance agreement-based measure
[54]	continuous binary	mRNA, DNA methylation somatic mutation	scaled exponential kernel of weighted Euclidean distance scaled exponential kernel of weighted Hamming distance
[62]	continuous	mRNA, miRNA, DNA methylation	scaled exponential kernel of Euclidean distance
[63]	categorical, discrete	demographic, APOE4 allele status, MRI	squared-exponential kernel
[55]	continuous	gene expression, miRNA, isoform expression	kernel of Chebyshev distance
[48]	continuous, categorical, discrete	clinical data	weighted sum of distances with weight determined by a scaled Cox regression coefficient

Abbreviations
ICD-9: International Classification of Diseases Version 9; **CNV**: Copy Number Variation; **miRNA**: micro RNA; **MRI**: Magnetic Resonance Imaging; **mRNA**: messenger RNA; **PPI**: Protein-Protein Interaction;

multiple kernel clustering models, the MKKM loss function includes a term that promotes the choice of higher weights for uncorrelated kernels.

SNF-based methods

PSNs are similarity graphs by definition; therefore, recent promising works apply graph-based algorithms and theories to integrate them. In particular, some authors simply integrate the information from different similarity graphs by using graph kernels [57], or by averaging [58, 81].

On the other side, Similarity Network Fusion (SNF [3]) exploits a nonlinear message-passing algorithm [82] that diffuses the information between all the unimodal PSNs constructed on each data-block until they converge to the integrated PSN. The diffusion process is designed so that the similarity between any two points computed over a specific source is updated and diffused if the two points are neighbors or share common neighbours in the other modalities. SNF has proven to be successful when compared to relevant PSN-fusion methods [83] in the unsupervised clustering task on three real, complex, multi-omics datasets (murine liver -

Table 2. Similarity measures/methods used in literature to build PSNs. For notable works in literature the table reports: the reference of the literature work presenting a multimodal PSN analysis method (column *Reference*), the data types (column *Data type*) of the different sources (column *Data*) exploiting for the investigation, and the similarity measures/methods exploited for building the unimodal PSNs.

Reference	Data type	Data	Similarity measure/method
[30, 53]	continuous categorical, discrete, binary	mRNA, clinical clinical	normalized linear kernel
[64]	discrete continuous	MRI CSF	gaussian kernel
[51]	continuous discrete	mRNA, miRNA, CNV, DNA methylation, clinical	gaussian kernel
[50]	continuous, binary, discrete	mRNA, miRNA, CNV, DNA methylation, RPPA, somatic mutations, clinical data	linear kernel
[65]	continuous	mRNA, CNV, DNA methylation	normalized linear kernel, normalized polynomial kernel, normalized gaussian kernel
[53]	continuous, categorical (ordinal) categorical (nominal)	clinical variables clinical variables	absolute difference of values/ranks of two subjects compared and rescaled using variable range kernel defined using Kronecker delta function
[57]	continuous, binary	mRNA, RPPA, somatic mutation	novel graph kernel called SmSPK

Abbreviations
CSF: CerebroSpinal Fluid; **CNV**: Copy Number Variation; **miRNA**: micro RNA; **MRI**: Magnetic Resonance Imaging; **mRNA**: messenger RNA; **RPPA**: Reverse-Phase Protein Arrays.

BXD [84], platelet reactivity [85], and Breast Cancer dataset from TCGA - BRCA [86]).

Several works extended SNF in different ways, thus creating a group of algorithms (called SNF-based methods). As an example, Affinity Network Fusion (ANF) [87] has been developed to diminish the computational costs of SNF, by reducing the iterative integration strategy of SNF to a unique step. To this aim, authors design a multigraph where each layer corresponds to a source-specific PSN, and then apply the one-step random walk kernel, where user-defined parameters are the transition probabilities between different layers, and the PSN for a specific layer represents the transition probabilities between nodes in that layer. When tested on multiple TCGA datasets, AFN outperforms SNF both in terms of clustering efficacy and computational costs.

By taking into account that the Euclidean distance metric employed in SNF suffers the curse of dimensionality [88] and may affect the results, [89] presented HSNF (hierarchical SNF), which essentially runs SNF several times, where each iteration uses a set of unimodal PSNs, generated on each data-block by using a randomly sampled feature set. At each iteration, the computed PSNs are fused with the integrated network computed in the precedent steps through SNF. The method is evaluated by its capacity to identify cancer subtypes by applying spectral clustering on the integrated matrix. Though outperforming SNF on several cancer datasets, HSNF has a higher computational cost because of the iteration of SNF.

To reduce noise in the integrated network, the Similarity Kernel Fusion algorithm (SKF) [90] multiplies the PSN built by using SNF with a matrix of weights, where the weight is higher if two samples are included in each other neighbourhood. Moreover, different from SNF, a term in the iterative update function is added to control the amount of information to be retained from the integrated kernel at the preceding step. When compared to SNF and to a simple average fusion of different kernels, SKF obtains comparable or even better performance in the discovery of cancer subtypes from real cancer datasets.

The association-signal-annotation boosted similarity network fusion (ab-SNF) method [54] tries to improve SNF by considering a weighted version of distance measures with the goal to up-weight signal features and down-weight noisy ones. In this work, the weight for continuous variables consists in a p-value computed by the univariate t-test to assess the feature significance in predicting the outcome variable; the weights for binary features, such as mutation data, are obtained by considering prior knowledge from databases (e.g. 1 for features related to cancer and 0 otherwise). Given the computed weights, the unimodal PSNs are obtained by using the scaled exponential kernel [3], where the Euclidean distance is substituted by the weighted Euclidean distance, for continuous variables, or the weighted Hamming distance, for binary variables. The use of feature-level weights leads to superior performance in clustering accuracy with respect to SNF on both simulated and real data, while subtypes captured by ab-SNF are significant in terms of patient survival on real cancer data.

Other PSN-fusion methods

NetDx [6] fuses unimodal PSNs by a simple weighted network sum, where the weights for each network are identified by ridge regression to a target network constructed on the training patients in order to enforce higher

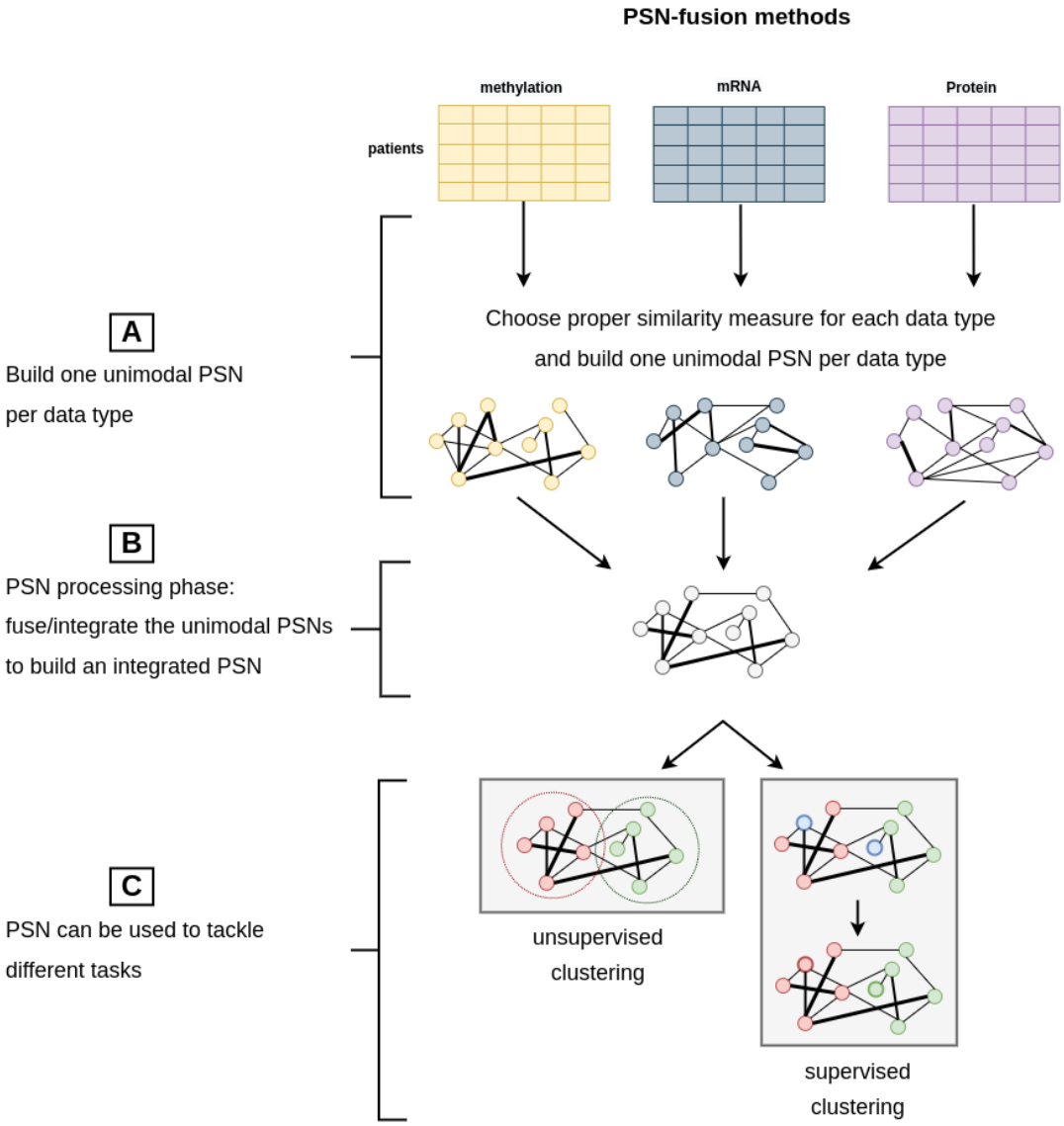


Fig. 2. High-level representation of PSN-fusion methods. (A) Given a set of matrices, each representing the patients vectors acquired from one source, proper similarity measures or kernel functions are used to build a set of unimodal PSNs (one PSN per data-source or data-type); (B) all the PSNs are then fused through either MKL methods, SNF methods or other PSN-fusion approaches; (C) the integrated PSN is processed either by unsupervised clustering algorithms for solving, e.g., patients' subtype prediction tasks, or by supervised classifier models for, e.g., patients' outcome prediction.

similarities between positive nodes and lower similarities between nodes belonging to different classes.

Some recent integration methods propose integrating the different PSNs by using a graph-based construction, and then compute integrated similarities by visiting the graph through random walk kernels. As an example, [58] propose computing similarities over a multiplex graph composed by a collection of PSNs (layers) each built on an individual data-block. The different layers share the same set of nodes [91], and corresponding nodes in different layers are connected to guarantee connectivity across multiple layers, but are considered as different entities to avoid disrupting the difference between the multiple views available for each node/sample. Then authors use the random walk kernel with restart [92] to express the similarities as the probabilities of reaching a node in a specific layer when another node in the same or in another layer is used as the starting point of the walk. To account for multimodality, that is with the presence of multiple layers, the probability of "jumping" to another layer during the walk is weighted by a parameter λ. The probabilities are computed by an iterative process that continues until a stationary point is reached. RWRNF [58] is an extension of this method that allows connecting multiple layers by also using edges between neighbourhoods of corresponding nodes. The use of many random walks, starting from all the nodes in each layer, adjusts the weights of the multiplex network taking into account its global topology. Finally, an integrated similarity network is computed by averaging corresponding weights across different layers of the network.

Table 3. MKL-based PSN-fusion methods. For each method, the table reports: the name/acronym with the corresponding reference paper; whether it requires the same set of patients across all data modalities (i.e. "Matched Samples"); the dataset used to develop and evaluate the approach in the reference paper and the corresponding sample cardinality and data types composing the dataset; the exploited integration method; the application task and the code availability (with link to the repository and programming languages for which the code is available).

Name	Matched Samples	Dataset	Sample Cardinality	Data type	Integration approach	Task	Code and Language
PAMOGK[1] [57]	x	TCGA KIRC NCI-PID at NDEXBio	361	somatic mutation mRNA RPPA	MKKM [77]	Unsupervised Clustering (Patient subtype identification)	MATLAB, Python code
[64]	x	ADNI	120	CSF features MRI	MKL [17]	Supervised Classification (HC vs MCI patients)	
[69]	x	TCGA	585	Histopathological images clinical data mRNA methy RPPA	simpleMKL [17]	Supervised Classification (Patient's Prognosis)	
[67]	x	TCGA GBM	125	Histopathological images CNV mRNA miRNA	simpleMKL [17]	Supervised Classification (Patient's Prognosis)	
MK-FDA [73] [75]	x	Protein dataset	Not provided	protein sequences	MKL	Supervised Multiclass Classification (Protein subcellular localization)	
[50]	x	14 TCGA datasets	3382	germline variants somatic mutation CNV mRNA miRNA methy	MKL	Supervised Classification (Patient's Survival)	
[52]	x	TCGA from mixOmics	989	mRNA miRNA methy	MKL	Unsupervised Clustering (Patients' subtype identification)	
rMKL-LPP [76]	x	TCGA GBM TCGA BIC TCGA KIRC TCGA LUSC TCGA COAD	213 105 122 106 92	mRNA miRNA methy	MKL	Unsupervised Clustering (Patient subtype identification)	

Abbreviations

ADNI: Alzheimer's Disease Neuroimaging Initiative; **CNV**: Copy Number Variation; **CSF**: CerebroSpinal Fluid; **HC**: Healthy Control; **MCI**: Mild Cognitive Impairment; **methy**: DNA methylation; **miRNA**: micro RNA; **MKKM**: Multiple kernel k-means clustering; **MKL**: Multiple Kernel Learning; **MRI**: Magnetic Resonance Imaging; **mRNA**: messenger RNA; **NCI-PID**: National Cancer Institute—Pathway Interaction Database; **RPPA**: Reverse-Phase Protein Arrays; **TCGA+*cancer code***: The Cancer Genome Atlas+ *link to complete cancer codes*.

The efficacy provided by the use of similarities computed across local neighborhoods is proven by its use in simpler unsupervised PSN analysis methods. As an example, NEMO (NEighborhood based Multi-Omics clustering, [93, 94]) is an unsupervised clustering approach where authors use a scaled normalized euclidean kernel to compute similarities, which are then made symmetric in a way very similar to SNF and are designed to have values equal to zero for nodes that are not neighbors. Extensive experiments on simulated and real datasets showed the competitive effectiveness and efficiency of NEMO with respect to 9 state-of-the-art methods among which one MKL-based method, a spectral clustering method, the classic k-means clustering approach, and 6 clustering methods exploiting an input data-fusion approach (Section Input data-fusion and output-fusion methods).

Finally, a noteworthy PSN-fusion method applied for unsupervised patient subtype identification in the TCGA dataset is Multi-view Spectral Clustering Based on Multi-smooth Representation Fusion (MRF-MSC)

[95]. MRF-MSC starts by individually processing each data-block to obtain a smoothed similarity matrix where strong/weak similarities are strengthened/eliminated; this is obtained by solving a regularized optimization problem that computes the similarity matrix in a feature space that minimizes the point-reconstruction error while strengthening the point groupings. Next, a fused similarity matrix that minimizes the weighted distance from all the smoothed source-similarity matrices is obtained by integrating a self-weighting method [96] into the distance minimization problem. Finally, the clusters in fused similarity networks are strengthened by applying the constrained Laplacian rank method and Spectral clustering is then applied to solve the clustering problem.

Input data-fusion and output-fusion methods

Opposite to PSN-fusion models, the input data-fusion and the output-fusion techniques reviewed in this section integrate the information available either in the multimodal input data (*input data-fusion* methods - Fig. 3) or in the output computed by a set of individual unimodal PSN-analysis models (*output-fusion* methods - Fig. 4).

Input data-fusion methods are schematized in Fig. 3. These approaches are based on the assumption that the input samples originally lied in a latent (eventually orthogonal) space from which the multiple source-views have been generated by unknown projections. This results in data-blocks being expressed into separate source-specific spaces that are characterized by: 1) an individual source-specific structure generating an individual variability within each data-block; 2) a joint sample-specific structure [18] resulting in shared variance (collinearities) between data blocks. Therefore, input data-fusion methods estimate the embedding that back-projects the input data-blocks into a shared latent space minimizing redundancy between the data-blocks while maximizing the individual data-block variability. In other words, all the methods find the *joint components* (Fig. 3) allowing to capture the greatest amount of shared variance; most of the methods also define ways to identify the *individual components* capturing the source-specific variability (Fig. 3).

Depending on the technique used to project the data into the shared latent space, we can distinguish input data-fusion methods into PCA-based techniques (Table 6) or Matrix Factorization (MF) or Blind Source Separation based methods (Table 7). One advantage of solving the information-fusion in a pre-processing phase, i.e. preceding the construction of an integrated PSN, is that a standard unimodal PSN-analysis model can be subsequently applied (Fig. 3-B) to deal with clustering or supervised classification problems (Fig. 3-C). In particular, the input data-fusion methods make the choice of the similarity measure to be used for PSN construction particularly easy, since they compute normalized, a-dimensional, integrated point representations, whose pairwise similarities could be handled by classic measures such as the cosine similarity or the inverted euclidean distance. Moreover, a side-effect of the estimated embedding is that the estimated component loadings or factors may be analyzed for uncovering hidden relationships between variables (data analysis task in Tables 6 and 7 and in Fig. 3).

The strategy applied by *output-fusion* methods is sketched in Fig. 4 and their experimental design is summarized in Table 8. They apply individual PSN pipelines on each data source to obtain individual clustering or supervised prediction results (Fig. 4-A and 4-B). All the obtained results are then fused by aggregation strategies that, acting as judges, compute a final decision by considering all the individual decisions taken by each unimodal pipeline.

Input data-fusion via PCA-based and CCA-based methods

In the bioinformatics field, Consensus PCA (CPCA [99]), hierarchical PCA (HPCA [106]), and Multiple Factor Analysis (MFA [107]), are some of the most used PCA-based integrative methods. They achieved interesting results on multimodal datasets including different types of patient data, from omics [18] to images [108, 109, 110].

Their effectiveness is due to their ability to project the data-blocks into a lower dimensional space spanned by not-correlated axis (principal components) maximizing the within-block variances and between-block covariances [111, 112]. By stretching the data along those axis, they induce a natural separability that improves the performance of the downstream algorithms, which are mostly devoted to data-exploration and unsupervised clustering, though some exceptions using supervised clustering exist [20] (Table 6).

The difference between the three approaches relies on the way the latent space is found. Indeed, while *CPCA* solves an optimization problem by an iterative algorithm in the set of nonlinear iterative partial least squares methods (NIPALS [113]), *HPCA* [106] and *MFA* [107] consecutively apply PCA on respectively: a) each block separately to derive lower-dimensional "stretched" block representations maximizing the within-block variance; b) the concatenation of the obtained block representations to derive a stretched latent space maximizing the between-block covariance.

A notable generalization of PCA for multimodal data is *JIVE* (Joint and Individual Variation Explained, [18]), which explicitly models each data-block X_i as the sum of a matrix representing the joint structure associated with X_i and shared with other sources, and a matrix representing the source-specific structure characterizing X_i, and residual noise. Given this formulation, authors apply an iterative estimation procedure that minimizes the reconstruction error, while constraining the axis of the joint and individual structures to be orthogonal (that is, the joint and individual structures must be uncorrelated). In practice the estimation iterates over the following two steps: 1) having removed the individual structure, apply a sparse Singular Value Decomposition (SVD) to estimate a lower-dimensional joint structure; 2) having removed the joint structure, apply a sparse SVD to find a lower-dimensional individual structure. Interestingly, JIVE also provides a permutation test to select the optimal ranks for the estimated structures. When experimented on multi-omics data from the glioblastoma multiforme (TCGA-GBM) dataset [18], JIVE showed its ability to effectively uncover the individual and joint data structures, thus leading to a better interpretation of interactions among data types and improving unsupervised classification results. Since the computational complexity of JIVE hampers its applicability, it has been recently reformulated (Angle Based JIVE - aJIVE [97]) by using a hierarchical strategy similar to HPCA, which also produces more intuitive interpretations of the obtained decomposition, especially in the presence of strong collinearities. The effectiveness of aJIVE is witnessed by the promising results obtained when applied to an extract of the TCGA breast cancer dataset from [101] for the (supervised) task of tumor subtype prediction [114]. In particular the estimated joint components and the first five individual components for each data block are used to compose the integrated sample views to train Random Forest classifiers [115].

Opposite to PCA-based integrative models, Canonical Correlation Analysis-based (CCA-based) integrative models, e.g. Regularized Generalized CCA (RGCCA) [104, 105] and its sparse counterpart Sparse Generalized CCA (SGCCA) [19, 105], find the latent space maximizing the correlation within and between the different data-blocks. They are generally used for exploratory variable analysis since they try to bring all the data blocks to a unique distribution, therefore uncovering hidden relationships between different sources. However, DIABLO [20] has shown that SGCCA is also effective in the context of supervised clustering for patients'

Table 4. (I) SNF-based PSN-fusion methods. For each method, the table reports: the name/acronym with the corresponding reference paper; whether it requires the same set of patients across all data modalities (i.e. "Matched Samples"); the dataset used to develop and evaluate the approach in the reference paper and the corresponding sample cardinality and data types composing the dataset; the exploited integration method; the application task and the code availability (with link to the repository and programming languages for which the code is available).

Name	Matched Samples	Dataset	Sample Cardinality	Data type	Integration approach	Task	Code and Language
SNF [3]	x	TCGA GBM	215	mRNA miRNA methy	SNF	Unsupervised Clustering (Patient subtype identification)	MATLAB code, , R code
ANF [87]	x	TCGA LUSC TCGA Adrenal TCGA Gland TCGA KIRC TCGA Uterus	2193	mRNA miRNA methy	SNF	Unsupervised Clustering (Patient subtype identification)	R code
HSNF [89]	x	TCGA BIC TCGA GBM TCGA KIRC TCGA LUSC TCGA COAD	105 215 122 106 92	mRNA miRNA methy	SNF	Unsupervised Clustering (Patient subtype identification)	
SKF [90]	x	TCGA BIC TCGA COAD TCGA KIRC TCGA LUSC TCGA Stomach	1071 426 868 981 377	mRNA miRNA isoform level	SNF	Unsupervised Clustering (Patient subtype identification)	MATLAB code
ab-SNF [54]	x	TCGA LIHC TCGA KIRP TCGA BIC	Not provided	somatic mutation mRNA methy	SNF	Unsupervised Clustering (Patient subtype identification)	R code
NEMO [93, 94]		TCGA AML TCGA BIC TCGA COAD TCGA GBM TCGA KIRC TCGA LIHC TCGA LUSC TCGA SKCM TCGA OV TCGA SARC	3168 across all datasets	mRNA miRNA methy	SNF	Unsupervised Clustering (Patient subtype identification)	R code

Abbreviations
methy: DNA methylation; **miRNA**: micro RNA; **mRNA**: messenger RNA; **SNF**: Similarity Network Fusion; **TCGA+*cancer code***: The Cancer Genome Atlas+ *link to complete cancer codes*.

subtype prediction. In practice, given a multimodal dataset containing N samples organized into Y classes, DIABLO firstly creates an extra dummy (supervising) data-block where each column is an indicator variable for the point-class $(1 \ldots Y)$. Next, it uses SGCCA to maximize the covariance between all the data-blocks, including the supervising data-block. Given this representation, supervised clusters may be identified either 1) by averaging the components across data-blocks, to obtain an integrated patient representation that is then used by any supervised clustering algorithm (such as the Maximum Centroids algorithm [116]); 2) by applying the Maximum Centroids algorithm on each projected data-block to obtain individual clustering results, subsequently aggregated via a majority voting algorithm.

Though effective in several applications, all the aforementioned PCA-based methods suffer from two main limitations: sensitiveness to outliers

Table 5. (II) Other PSN-fusion methods. For each method, the table reports: the name/acronym with the corresponding reference paper; whether it requires the same set of patients across all data modalities (i.e. "Matched Samples"); the dataset used to develop and evaluate the approach in the reference paper and the corresponding sample cardinality and data types composing the dataset; the exploited integration method; the application task and the code availability (with link to the repository and programming languages for which the code is available).

Name	Matched Samples	Dataset	Sample Cardinality	Data type	Integration approach	Task	Code and Language
netDx [6]	x	TCGA KIRC	150	mRNA	average score	Supervised Classification (Patient's Survival)	R code
		TCGA OV	252	miRNA			
		TCGA GBM	155	methy			
		TCGA LUSC	77	CNV			
				RPPA			
				clinical data			
RWRF, RWRNF [58]	x	TCGA ACC	76	mRNA	RWR	Unsupervised Clustering (Patient subtype identification)	R code
		TCGA BLCA	396	miRNA			
		TCGA HNSC	469	methy			
		TCGA UVM	80				
		TCGA PAAD	175				
		TCGA THCA	492				
MRF-MSC [95]	x	TCGA COAD	92	mRNA	maximization of alignment to all the unimodal PSNs	Unsupervised Clustering (Patient subtype identification)	
		TCGA GBM	215	miRNA			
		TCGA BRCA	105	methy			
		TCGA KIRC	122				
		TCGA LSCC	106				

Abbreviations

CNV: Copy Number Variation; **methy**: DNA methylation; **miRNA**: micro RNA; **mRNA**: messenger RNA; **PSN**: Patient Similarity Network; **RPPA**: Reverse Phase Protein Array; **RWR**: Random Walk Kernel with Restart; **TCGA+*cancer code***: The Cancer Genome Atlas+ *link to complete cancer codes*.

and inability of handling missing data. Generalized Integrative PCA (GIPCA) [100] has been recently proposed as an extension of Consensus PCA for dealing with missingness of some values and of entire views. To this aim, eigenvectors are used to explain the intra/inter-block variance by neglecting those samples/views with missing values/views.

Input data-fusion via Matrix Factorization-based methods

Matrix Factorization (MF) methods [117] embed the points into a latent space that minimizes the reconstruction error and whose components (factors) are not constrained to be orthogonal (as in PCA) [118, 31, 119]. The most effective and used MF method applied on unimodal data is Non-negative MF (NMF, [120]); it constrains both the component and loading matrices to be non-negative, which makes the approximation purely additive.

Given its effectiveness, several works proposed methods where NMF is extended to the integration of multimodal datasets (see Table 7). The most relevant example is joint NMF (jNMF [121]) where multiple NMF problems are solved subject to a shared factor matrix that contains the basis vectors of the shared latent space. However, jNMF is sensitive to

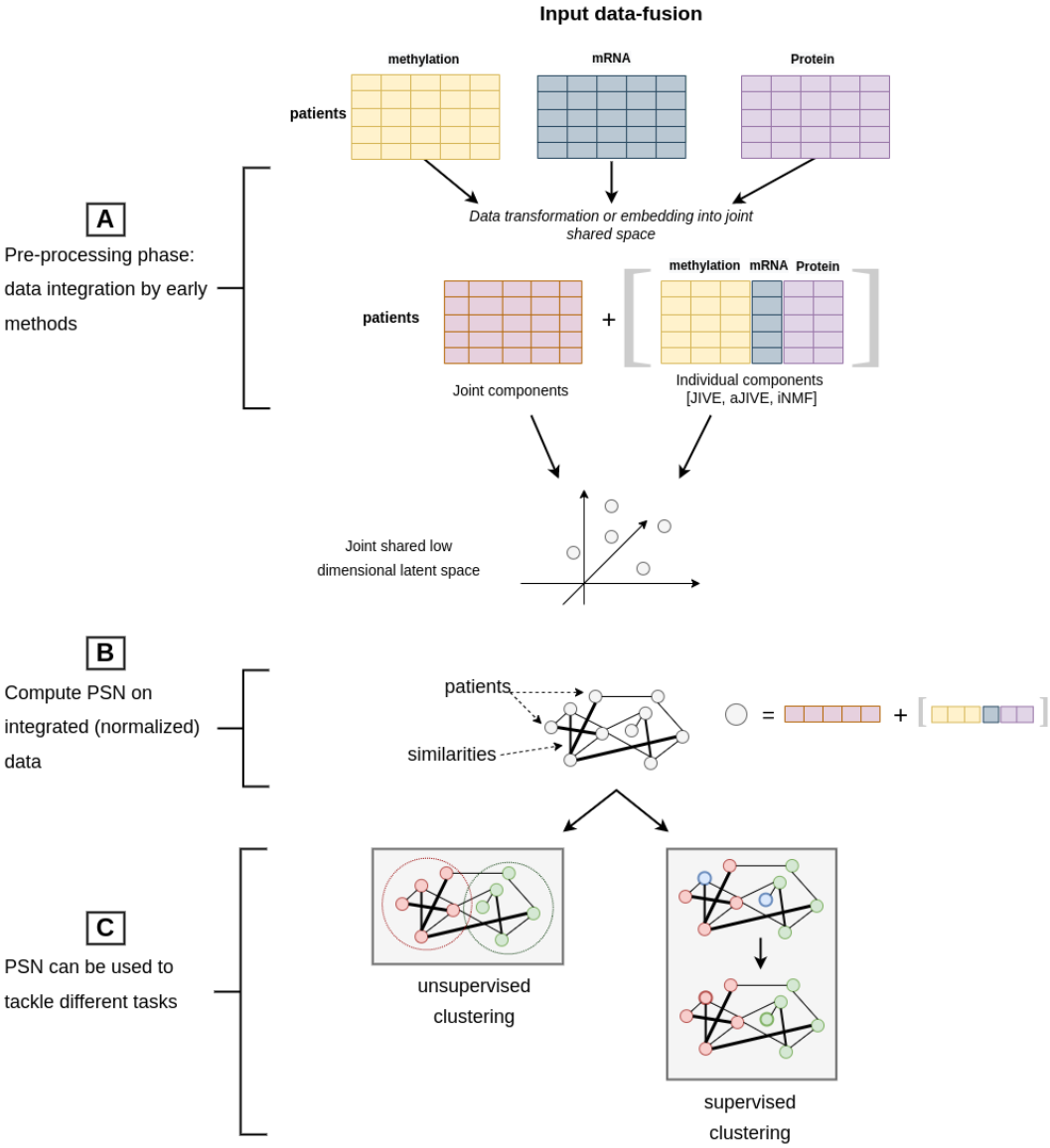


Fig. 3. Input data-fusion. (A) During the pre-processing phase the data are integrated by either a PCA-based integrative model or a MF-based model. They estimate a shared latent space where the integrated, normalized point representations express the joint structure underlying all the data blocks plus, eventually, the individual structures characterizing each data block (e.g. JIVE [18], aJIVE [97], iNMF [98]); (B) a PSN model is then constructed on the integrated profiles by using a classic similarity measure; (C) a clustering or supervised classification model is applied to the computed PSN.

random noise and confounding effects [98] that are specific to each source, and that cannot be detected if a unique shared factor matrix is estimated. This affects the accuracy of the common structure estimation computed by jNMF [98]. Therefore, integrative Non-negative Matrix Factorization (iNMF [98]) uses an approach similar to JIVE, where the factor matrices to be estimated are composed both by a shared and a source-specific structure. Unsupervised clustering experiments on the TCGA dataset [98, 122] have proven the superiority of iNMF with respect to jNMF [121], NMF [123], and to integrative Bayesian methods [124, 125].

Integrative Graph Regularized Non-Negative Matrix Factorization for Network Analysis (iGMFNA [126, 127]) proposes improving the minimization of the reconstruction error, typical of NMF, by exploiting a graph-view on each data block. Thanks to such representation, the designed iterative optimization minimizes the reconstruction error while maintaining the topology of the graph-views. When compared to jNMF and iNMF to prioritize genes associated with cancer in two TCGA datasets by an unsupervised clustering approach, iGMFNA showed its superior performance.

The popular Penalized Non-negative Matrix Tri-Factorization (NMTF, [128, 31]) starts from a relational matrix $R_{1,2}$ containing non-negative elements that represent the strengths of the relationships between objects of two different types, ϵ_1 and ϵ_2, whose respective characteristics are defined

by specific constraints, θ_1 and θ_2. NMTF finds the decomposition of $R_{1,2}$, subject to constraints θ_1 and θ_2, such that: $R_{1,2} \approx G_1 S_{1,2} G_2^T$ so that G_1 and G_2 are the low-dimensional representations of objects with types, respectively, ϵ_1 and ϵ_2, and $S_{1,2}$ is the backbone matrix linking the two types.

NMTF is exploited by [31] in *DFMF*, where the reliability of the integrated low dimensional estimates computed over a multimodal dataset is improved by considering all the relational matrices (and corresponding constraints) linking the different sources between each other and with the patient data. Given all the relational matrices, $R_{i,j}$, and respective constraints, each $R_{i,j}$ is decomposed so that each backbone matrix represents the latent structure between two data types, the generic low-dimensional data representations of objects with a specific type, G_i, is bound to be used in the reconstruction of every relational matrix involving that type. Thanks to the abundance of information, the proposed model can also handle missing data and treat sparse relational matrices. Furthermore, it does not make any assumption about the structural properties of relations, which can also be asymmetric. DFMF can also be used in a semi-supervised setting. During training, the model parameters (i.e. the factorization ranks) are learnt, and are then used in a matrix completion problem, where unobserved entries in the target matrix $R_{i,j}$ are reconstructed for elements that were not present in the training set.

DFMF has been successfully used in the Matrix trifactorization for Discovery of Data similarity and Association (MaDDA) algorithm proposed by [129] to construct PSNs for unsupervised clustering. In particular, given n patients to be partitioned into k clusters, the low-rank matrix $G \in R^{n \times k}$ estimated through DFMF is viewed as a membership matrix relating each patient to the k ranks/groups. After repeating the factorization multiple times with different initialization parameters, a final consensus matrix is obtained by element-wise averaging all membership matrices and then composing a PSN where the similarity between two patients (weight of the edge connecting them) represents how many times they ended up in the same group.

Multi-Omics Factor Analysis+ (MOFA+ [130]) is an integrative method exploiting Bayesian group factor analysis [51] with regularization to impose: (i) a view-wise and factor-wise sparsity, which shrinks to zero the loading for the m-th modality and the k-th factor if the latest does not explain any variability of the m-th view; (ii) a feature-wise sparsity, which sets to zero loading on individual features from active factors so that only a small number of features "actively" contribute to each factor. MOFA+ can handle missing values as well as entirely missing views for some samples; moreover, it can cope with heterogeneous data types, which is exactly what is needed when dealing with multimodal datasets containing multi-omics, clinical, and imaging data.

Given the successful results of MF-based integrative techniques, some authors have included them as a pre-processing step in their clustering/classification algorithms. As an example, iCluster+ [123, 131] uses NMF to fuse the heterogeneous data-blocks and then clusters the integrated views. It also exploits the obtained factor loadings to identify the relevant features in the cluster generation.

Input data-fusion via Blind Source Separation

In their original formulation, Blind Source Separation (BSS) models were defined as an extension of NMF techniques for "recovering unobservable source signals s from measurements x (i.e., data), with no knowledge of the parameters θ of the generative system $x = f(s, \theta)$" [132].

Given their documented ability [133, 134, 132] of uncovering hidden structures underlying the observed unimodal signals, several BSS models have been extended to handle multimodal datasets comprising also multisets[2] (Table 7), by a further step that estimates the mixing matrix that recombines all the estimated latent sources so as to compute an integrated, more informative signal with no redundancies [135, 136, 132, 21].

Given the lack of information about the mixing process and the source signals, BSS models often differ for the constraints they impose to counter the ill-conditioned problem and obtain essentially unique source estimates [137, 132, 134]. As an example, the well known Independent Component Analysis model (ICA [138]), and its extensions to multimodal data (joint ICA - jICA [139, 140]), to multisets (Independent Vector Analysis - IVA [141]), and to multidimensional sources (Independent Subspace Analysis - ISA [142]), assume a linear (additive) mixture with mutually independent sources and a non-Gaussian distribution of each independent component in the latent space.

All the BSS models base their computations on the existence of collinearities between the observed multimodal data components, so that unreliable results may be obtained when this assumption is not satisfied. Some authors [135] circumvent this problem by pre-processing the data with CCA (or its multimodal extension), to obtain a projected data representation along correlated components.

The most representative BSS-based multimodal data integration technique is Multidataset Independent Subspace Analysis (MISA [132, 21]), which was recently proposed to generalize all the BSS models to the fusion of any kind of multimodal-multisets. Motivated by the definition of multiset, MISA is driven by statistical independence between latent subspaces while assuming correspondence within the subspaces underlying the input multisets. In practice, it firstly removes redundancies by estimating non-orthogonal demixing matrices, projecting each multiset into a respective (intermediate) lower-dimensional space spanned by independent components. The sources from all the computed latent spaces are then combined through another demixing matrix that brings all the data-blocks into a unique shared latent space, resulting in an integrated patient view. The de-mixing matrices are estimated by minimizing the mutual information in the final space, while maximizing the mutual information in the intermediate spaces, so as to capture as much correlation as possible. When applied to the integration of the information extracted from Functional Multi-Resonance Imaging (fMRI), Structural Multi-Resonance Imaging (sMRI), and Electroencephalogram (EEG) data, MISA has proven its robustness with respect to high Signal to Noise Ratios (SNR) as well as its ability to produce effective data fusion in different ICA contexts.

Output-fusion methods

Following Fig. 4, in the context of multimodal PSN analysis the ouput-fusion methods described in this section may be applied to combine the (unsupervised clustering or supervised classification) results (Fig. 4-B) computed by individual PSN analyses applied on each data block (see Fig. 4-A). In Fig. 4-C, the combination of the unimodal results is performed either by some heuristics, or by majority voting, or by using a meta-model that learns from the predictions performed by each unimodal PSN analysis. Output-fusion techniques have been proposed for clustering samples (mainly from the TCGA datasets, see Table 8) to identify patients' subtypes [22, 144, 23] and for patients' classification [145, 63] (see Table 8).

[2] *Multisets* are multimodal datasets containing multiple views acquired by the same source under different acquisition conditions (e.g. observation times, experiments, tasks, machines). They are therefore *homogeneous* [26] in semantic, type, and dimensionality. *Multimodal-multisets* are multimodal datasets acquired by different sources, among which sources producing multisets.

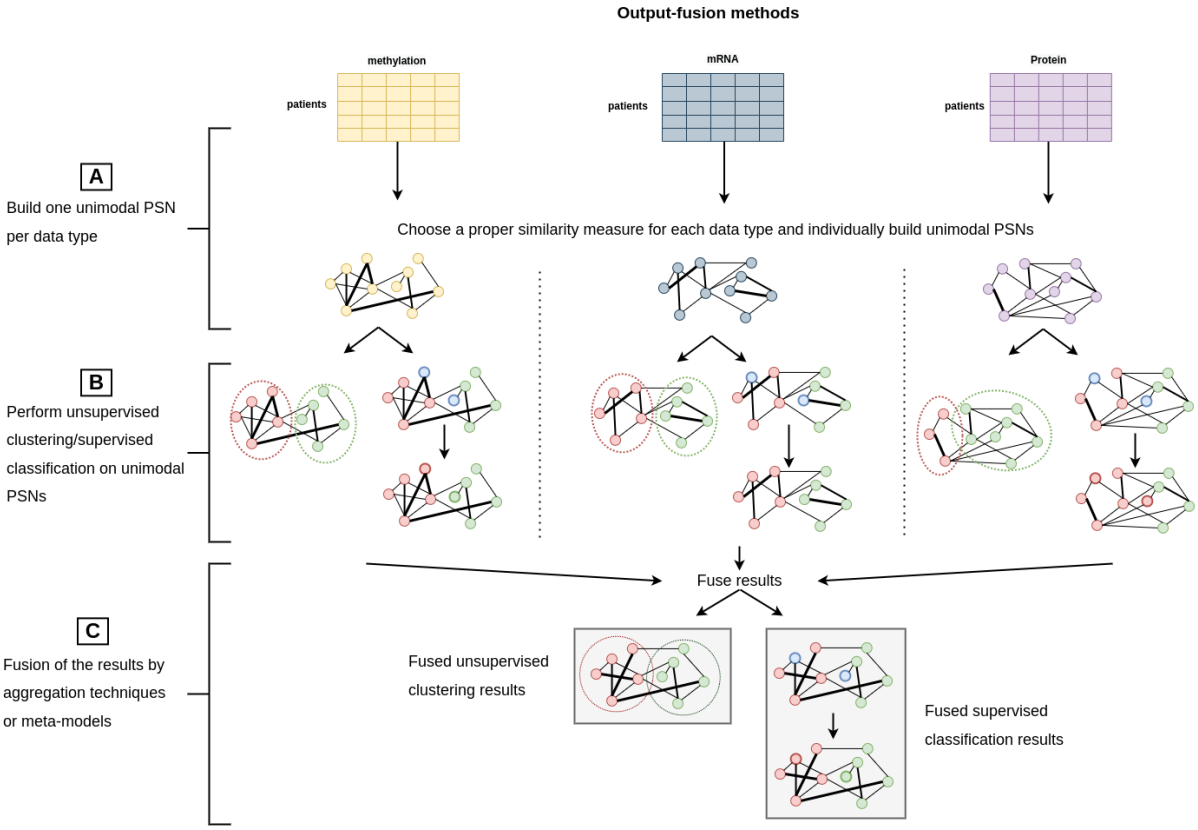


Fig. 4. Output-fusion. (A) Unimodal PSNs are constructed for each data-type or data-source, and (B) each one is individually processed to identify clusters or to classify unknown samples; subsequently, (C) a simple aggregation technique or a meta-model is used to obtain the fused/consensus clustering/classification result.

As an example, in Cluster-of-Cluster-Assignments (COCA [22]), authors combine the clustering results individually obtained by NMF [146] on each of the six data types of the TCGA datasets. To this aim, the samples are coded into vectors composed of indicator variables representing the clusters they have been assigned in each modality, so that they can be reclustered according to those vectors by Consensus Clustering Plus [147][4]

PINSPlus [144, 23] similarly exploits Consensus Clustering [148] for reaching the final partition. In practice, PINS (*Perturbation clustering for data INtegration and disease Subtyping*) starts by applying any classic unsupervised clustering algorithm (e.g., k-means) individually on each of the M-th datasets. If n is the number of patients, for the m-th dataset ($m \in 1, \ldots, M$) the clustering result is expressed by a square matrix $C_m \in \mathbb{R}^{n \times n}$, such that $C_m(i,j) = 1$ if samples i and j fall in the same cluster, and $C_m(i,j) = 0$ otherwise. All the resulting matrices are then averaged to obtain a consensus matrix $S = \frac{\sum_{m=1}^{M} C_m}{M}$. Even though matrix S may highlight that some points do not reach a strong agreement, authors consider that S itself may be used as a pairwise similarity matrix (since $S = 1$ for points for which there is a strong agreement, viewed as similarity, across all the dataset, and $S = 0$ otherwise) that is suitable for similarity/distance-based clustering algorithms such as any Hierarchical Clustering algorithm [149], Partitioning Around Medoids [150], or dynamic tree cut [151]. In their work, authors propose testing different clustering algorithms and then choose the partition that agrees the most with the partitioning of individual data types.

Consensus clustering has also been successfully applied by the recently published SUMO [62], an integrative clustering algorithm that starts by computing several unimodal PSNs by using a scaled-normalized Euclidean kernel similar to the one exploited by SNF [3]. SUMO then formulates a constrained NMTF (see Section **Input data-fusion via Matrix Factorization-based methods**) to find a sparse shared representation of all the samples in the cluster subspace by accounting for the adjacencies observed in all the data types. The NMTF optimization problem is solved by an iterative procedure that is applied several times on several sample subsets to ensure robustness with respect to the initial conditions and to the input data; consensus clustering is then exploited to pool together the clustering results. When compared to the most promising integrative clustering methods (e.g., iCluster [152], MCCA [102], NEMO [94], SNF [3], PINSPlus [23]) SUMO obtained impressive results.

The Fuzzy-Hierarchical CLUSTering - FH-Clust method [24] interestingly proposes to use fuzzy logic for identifying patients' prognostic subgroups from multiomics data, resting on the fact that in nature there is often no clear cut between subtypes. Unimodal data are separately analyzed using a fuzzy-based hierarchical clustering approach exploiting a Lukasiewicz valued fuzzy similarity and individual results are then fused through a consensus matrix. Extensive experiments on 10 cancer datasets from TCGA

[4] Given the number of clusters k, Consensus Clustering Plus works on a consensus matrix (CM_k) representing "the proportion of clustering runs in which two items are [grouped] together" [148]. Given CM_k an agglomerative hierarchical consensus clustering using distance of 1-consensus values is completed and pruned to k groups that are returned as consensus clusters.

(considering gene expression, miRNA, methylation data) show that FH-Clust is competitive with state-of-the-art methods (i.e. k-means, Spectral Clustering, LRACluster, PINS, SNF, MCCA).

Interesting output-fusion approaches aimed at patients' classification are described in [145, 63]. In [145] the authors obtain effective cancer-grade and patient-survival classifications for cancer patients represented in the TCGA renal (TCGA KIRC) and TCGA ovarian (TCGA OV) datasets by using all the data types included in TCGA, including hematoxylin and eosin (H&E) stained whole-slide images of tissue samples that are processed by digital image processing techniques to extract more that 400 features per sample. In practice authors firstly individually process each data block to apply an internal cross-validation approach to choose 1) the number of informative features to be extracted by the minimum Redundancy Maximum Relevance (mRMR) method [153] and 2) the best performing 5-fold cross classifier among SVM, logistic regression, K-nearest neighbors, and Linear Discriminant Analysis. To compose all the predictions from the different modalities authors compare the stacked generalization model [154], which essentially trains a logistic regression classifier on the obtained predictions, to the majority vote strategy. The best results are obtained by the stacked prediction model, which leverages the results obtained by any of the multimodal predictions, independent from the classifier that is used for producing them.

In [63] authors simply use the average to integrate the different prognostic classifications computed over multimodal profiles of suspected Alzheimer Disease (AD) patients, with the aim of identifying patients who are vulnerable to conversion from mild-cognitive impairment to AD. In particular, the squared-exponential kernels are firstly used to build unimodal PSNs, and, for each unimodal network, a Gaussian process is then exploited to assign labels to unknown points based on the nearest known points. Finally, the unknown patients' condition is computed as the average over all the unimodal predictions.

Discussion and Conclusion

In the context of Precision Medicine, PSNs are gaining momentum given their ability to uncover and exploit relationships among patients when applied to clustering and classification tasks [9]. According to the state-of-the-art surveys describing the application of PSNs for precision medicine or health-data processing [9, 45, 157, 158], PSN-based models benefit from several advantages; they are: (I) easy to understand, (II) interpretable by design, (III) privacy preserving, (IV) competitive or even superior to state-of-the-art clustering/classification methods, (V) potentially able to integrate different data views. In particular, the possibility of using PSN models in a multimodal setting is especially relevant in light of the increasing availability of digital technologies by means of which huge amount of multimodal data can be collected that describe each patient/sample by considering different biological/medical views. Moreover, in the past few years the increasing availability of cloud technologies allowed us to distribute data processing across multiple local servers belonging to, e.g. different institutions. In this context, the development of promising information integration models would allow the application of a Federated Learning strategy [159], where a central server collects, further integrates, and eventually processes, the (already) integrated data, or the individual PSNs, or the predictions individually computed by local servers located in the institutions where the data belong. In this way, the initial processing of the sensitive data would be demanded to the local institutions to protect patient privacy, and the central server would have access only to pre-processed information, thus hiding explicit sensitive data.

Though in the biomedical context several multimodal approaches have already shown their ability to integrate multimodal data to improve the results obtained from a single view (unimodal data) [114], and the survey literature about data integration methods for multimodal data is wide [16, 13, 15], in the field of PSN analysis only few methods have already investigated the usage of multimodal data, by building integrated PSNs that exploit both the joint and the individual information from all the available sources. Moreover, no state-of-the-art survey has focused on the role of PSN as a cornerstone for data fusion. In this survey, we aim at filling this gap with the goal of providing interested readers with a comprehensive collection of integrative methods that may be exploited to build PSN approaches efficiently handling multimodal data.

Besides an extensive literature search, the integration approaches have been organized into three broad classes on the basis of the type of data that is fused: *PSN-fusion*, *Input data-fusion* and *Output-fusion* methods. More precisely, *PSN-fusion* methods may be split into the three sub-classes of *MKL*, *SNF-based* and *other* methods while *Input data-fusion* approaches comprehend algorithms *PCA-based*, *CCA-based* and *MF-based*.

The survey has highlighted the promising results and advantages that characterize the methods belonging to the three classes of our proposed taxonomy.

Methods based on PSN-fusion techniques are particularly useful in Network Medicine applications [160], that study human diseases through "systemic" approaches in which diseases are interpreted as perturbations in complex biomolecular networks. In this context, transductive strategies working on individual PSN models [7] would benefit from the application of PSN-fusion approaches, as shown by recent promising results [3, 6].

Methods based on input-data fusion techniques rely on factor analysis models for the removal of data collinearities and the simultaneous enhancement of the individual structure characterizing each view. For this reason, we believe such techniques are particularly useful when dealing with multiview data involving follow-up examinations, where the multiple views likely contain correlated information.

Output-fusion techniques should be used when the differences between the multimodal views impose the usage of peculiar and specific unimodal PSN models for obtaining individual inferences. This is the case, for example when we need to combine data having substantially different structures, ranging from vectorial to sequence and graph-structured data.

Though being effective, our thorough review also evidenced difficulties and drawbacks that harbour from the data-fusion strategy. In particular, PSN-fusion models require to build an individual PSNs on each data-type. This raises the crucial, still open, and often overlooked problem of choosing proper individual similarity measures for building each unimodal PSNs. Indeed, only few methods [60, 161, 162] reported exhaustive comparative evaluations among few distance metrics applied to genetic data. By considering that several problems in Precision Medicine are characterized by non-linearly separable omics data, and given the experimental results we have collected during our literature search, we recommend computing PSNs by exploiting a kernel function. In this context, though several functions have been successfully proposed and used in literature, when dealing with continuous data, we suggest using the scaled exponential kernel of Euclidean distance [3, 62], due to its ability to adapt to different neighborhood sizes. This allows dealing with datasets distributed on complex manifolds where datapoints are not evenly distributed in space, as it often happens in real-world problems. On the other hand, when dealing with simpler data-types with lower dimensionalities and complexities (e.g. clinical data), simpler normalized similarities may be sufficient to appropriately capture the data structure. Clinical datasets usually contain categorical variables, often mixed with numeric features. The former situation can be appropriately addressed by averaging the normalized similarities individually computed on each variable [6], while Chi-squared distances are the most suitable for categorical data [3, 6]. Of note, the

subsequent application of a Random Walk kernel, as proposed by Gliozzo et al [7], is a promising step to refine the obtained PSN.

On the other side, input data-fusion techniques integrate the input data by projecting them into a shared space with lower dimensionality, thus making these approaches strongly dependent on the chosen final dimensionality d.

While classic approaches have been proposed to automatically set d [163, 164, 165, 100], this value is often user-defined after observation of the scree plot. However, observing that the optimal latent vector space is the one that allows to capture the intrinsic data structure, we instead suggest setting d to the intrinsic data dimensionality (*id*) [166], which is the minimum number of parameters needed to represent the data without information loss.

Finally, output-data fusion methods are often too generic or use very simple output-aggregation strategies, e.g. average or majority voting, that may produce sub-optimal results.

Generally speaking, our survey evidenced some important open issues in the context of data integration methods for PSN that call for the future research directions summarized in the following subsection.

Future Research Directions

While conducting our survey we noted the need of investigating methods for data pre-processing, with the aim of, e.g., detecting and eliminating noise with heterogeneous characteristics, collinearities between different views, and confoundings that could bias the final results (as per [27]). Indeed, only few recently proposed preliminary attempts were able to explicitly consider the presence of noise with heterogeneous characteristics [98, 122].

Moreover, future research should be devoted to the investigation of novel multimodal feature-selection algorithms. Indeed, the few methods applying a feature selection step exploit either classic univariate statistics, or algorithms, such as mRMR [153], that analyze group of features by neglecting their multimodal characteristics.

On the other side, missing data imputation needs deeper investigation to handle two types of biomedical data-missingness: 1) missingness of some data values in some views; 2) missingness of entire views for some samples. While missingness is becoming a common problem in different fields, in the bio-medical field few approaches present thorough missing data imputation studies [11]. Besides, among the approaches we have surveyed, only GIPCA [100] specifically addressed both these types of missingness. Finally, given the big-data produced by high-throughput technologies, scalability is becoming an important and often overlooked issue, nowadays hampering the applicability of several promising tools.

Though the aforementioned issues are still open, all the surveyed strategies have reported promising results that might improve knowledge in then field of Precision Medicine. Unfortunately, different similarity metrics, experimental setups, and evaluation measures are used for model assessment; this hampers an objective comparison between the different integration techniques and data analysis models. Furthermore, we found no evidence about data integration approaches that should be preferred over the others. Instead, the type and semantic of the available data type and the specific biomedical question to address should guide the choice. An additional open problem regards the identification of the most appropriate similarity/distance measure for each biological data modality. To the best of our knowledge, only few works tried to investigate this issue by comparing different metrics for specific data views and most of them are focused on gene expression data [60, 161, 162]. Comprehensive studies comparing the usage of different similarity measures in different contexts (e.g. when applied to different biological data types and in supervised and unsupervised prediction contexts) would provide fruitful insights to guide the scientific community towards effective PSN construction. We also remark that, though some algorithms are already available as open source packages/repositories (mostly coded using R, Python and Matlab) [16], many others are not, thus slowing down their diffusion and testing by the community.

Another interesting research line that should be given attention is represented by the development of Web applications extending, e.g., those presented in [167, 168], for the visual analysis of PSN models. Indeed, the graphical tools can enable the visual comparison of different PSN models realized according to any of the methods discussed in this survey. This in turn can improve the explainability of the computed results and would allow the user to choose the approach mostly suited to her/his needs.

Abbreviations

ab-SNF - Association-signal-annotation Boosted Similarity Network Fusion
AD - Alzheimer's Disease
ADNI - Alzheimer's Disease Neuroimaging Initiative
ANF - Affinity Network Fusion
aJIVE - Angle based Joint and Individual Variation Explained
BSS - Blind Source Separation
CCA - Canonical Correlation Analysis
CGH - Comparative Genomic Hybridization
CNV - DNA Copy Number Variation
COCA - Cluster of Cluster Assignment
CPCA - Consensus Principal Component Analysis
CSF - CerebroSpinal Fluid
DFMF - Data Fusion by Matrix Factorization
DILI - Drug-Induced Liver Injury
DLBCL - Diffuse Large B-Cell Lymphoma
EEG - Electroencephalography
fMRI - functional Magnetic Resonance Imaging
FH-Clust - Fuzzy-Hierarchical CLUSTering
GIPCA - Generalized Integrative PCA
GNMF - Graph Regularized Non-negative Matrix Factorization
GO - Gene Ontology
HPCA - Hierarchical Principal Component Analysis
HC - Healthy Control
HSNF - Hierarchical Similarity Network Fusion
ICA (and *DS-ICA*) - Independent Component Analysis (for Disjoint Subspaces)
id - intrinsic data dimensionality
iGMFNA - integrated Graph Regularized Non-negative Matrix Factorization for Network Analysis
iNMF - integrative Non-negative Matrix Factorization
ISA - Independent Subspace Analysis
IVA - Independent Vector Analysis
jICA - joint Independent Component Analysis
JIVE - Joint and Individual Variation Explained
jNMF - joint Non-negative Matrix Factorization
KEGG - Kyoto Encyclopedia of Genes and Genomes
LPP - Locality Preserving Projections
LRAcluster - Low-Rank Approximation based multi-omics data clustering
MaDDA - Matrix trifactorization for Discovery of Data similarity and Association
MCCA - Multiple Canonical Correlation Analysis
MCI - Mild Cognitive Impairment
methy - DNA methylation
MFA - Multiple Factor Analysis
MF - Matrix Factorization

miRNA - microRNA
MISA - Multidataset Independent Subspace Analysis
MK-FDA - Multiple Kernel Fisher Discriminant Analysis
MKKM - Multiple kernel k-means clustering
MKL - Multiple Kernel Learning
MOFA - Multi-Omics Factor Analysis
MRF-MSC - Multi-view Spectral Clustering Based on Multi-smooth Representation Fusion
MRI - Magnetic Resonance Imaging
mRMR - maximum Relevance Minimum Redundancy
mRNA - messenger RNA
MTF - Matrix Tri-Factorization
MultiNMF - Multi-View Non-negative Matrix Factorization
NCI-PID - National Cancer Institute—Pathway Interaction Database
NEMO - NEighborhood based Multi-Omics clustering
NIPALS - Nonlinear Iterative Partial Least Squares
NMF - Non-negative Matrix Factorization
NMTF - Penalized Non-negative Matrix Tri-Factorization
PAMOGK - Pathway-based MultiOmic Graph Kernel clustering
PCA - Principal Component Analysis
PINS - Perturbation clustering for data INtegration and disease Subtyping
PPI - Protein-Protein Interaction
PSN - Patient Similarity Network
RDA - Regularized Discriminant Analysis
RGCCA - Regularized Generalized Canonical Correlation Analysis
rMKL - Regularized Multiple Kernel Learning
rMKL-LPP - Regularized Multiple Kernel Learning - Locality Preserving Projections
RPPA - Reverse-Phase Protein Arrays
RW - Random Walk
RWR - Random Walk with Restart
RWRF - Random Walk with Restart for multi-dimensional data Fusion
RWRNF - Random Walk with Restart and Neighbor information-based multi-dimensional data Fusion
SCA - SpinoCerebellar Ataxia
SGCCA - Sparse Generalized Canonical Correlation Analysis
SKF - Similarity Kernel Fusion
sMRI - structural Magnetic Resonance Imaging
SNF - Similarity Network Fusion
SNR - Signal to Noise Ratio
SVD - Singular Value Decomposition
SVM - Support vector Machine
TCGA - ***study name*** - The Cancer Genome Atlas - ***link to official study abbreviations***

Appendix A - Data integration in Medicine: previous surveys and taxonomies

The abundance of multimodal data integration approaches developed in the past decade in the biomedical context has motivated many relevant surveys [29, 36, 37, 34, 35, 16], which proposed different definitions and taxonomies, schematized in Fig. 1.

In the context of Precision Medicine, multimodal sets are composed of multiple views (or data-blocks) for the same set of patients. They are either *multisets*, or *multimodal datasets* [142]. *Multisets* (top of Fig. 1-yellow box) contain multiple views acquired by the same source under different acquisition conditions (e.g. observation times, experiments, tasks, machines), and are therefore *homogeneous* [26] in semantic, type, and dimensionality. Conversely, *multimodal datasets* (alias *heterogeneous* sets [26], Fig. 1-light blue box) contain data-blocks acquired by different sources, characterized by different semantics, type and dimensionalities. Among the latter, *multimodal-multiset* are datasets acquired by different sources, some of which are used to produce multisets.

Given their different characteristics, multisets and multimodal datasets are generally fused by following different integration flows. *Horizontal integration* methods [169, 25] are usually used for multisets because they equally process all the data-blocks and then pool the obtained results by e.g. summary statistics [170]. By contrast, *vertical integration* methods are used for processing multimodal datasets, which are more articulated and are usually grouped in the *hierarchical-vertical* class and the *parallel-vertical* class.

Hierarchical-vertical [171, 172, 173] or *multi-staged analysis* methods [27] consider omics data being interrelated by regulatory mechanisms and exploit such prior knowledge during the integration procedure. Since these methods are tailored for the treatment of specific data types and applications and cannot be generalized to different research contexts, they will not further considered in this survey. *Parallel-vertical* integration techniques, alias *meta-dimensional analysis* methods [27], are the most diffused and generalizable ones because dependencies between data-blocks injected by prior information are not considered. To categorize parallel-vertical approaches several interrelated taxonomies have been defined. The categorization reported in the red-dashed box in Fig. 1 is the one adopted by several authors [29, 30, 31, 27, 32, 33] that relies on the processing stage (early, intermediate, late) in which the data fusion happens, which also influences the kind of information that is fused.

Early integration techniques (also called *concatenation-based models* [27, 33]) are applied on the input data-blocks in an early stage to compose the integrated input vectors subsequently used in the analysis by either using a simple data-concatenation, or by exploiting joint latent space estimation models. The evident advantage of early methods relies on their ability to uncover the individual information characterizing each of the different sources as well as the hidden relationships between them. Another advantage is brought by the fact that early methods solve the integration problem in the first stage, so that any unimodal analysis process may be subsequently applied.

Intermediate integration approaches (also named *transformation-based models*) [27, 33] individually transform the data-blocks into intermediate (unimodal) models that are subsequently integrated to produce a unique fused model to be analyzed. In the taxonomy proposed by [28] (blue-dashed box in Fig. 1), these methods have been classified as *model-dependent* approaches for highlighting their dependency from the data analysis model, which guarantees the ability to retain the original data structure by explicitly addressing the fusion task in the construction of the predictive model itself.

Late integration approaches (also named *model-based approaches* in [27, 33]) separately analyze each of the incoming data-blocks to produce individual results, subsequently integrated in a late phase by some meta-learner acting as the final judge or by simple techniques such as majority voting. These approaches along with the early integration ones are classified as *model-agnostic* in the taxonomy proposed by [28] (blue-dashed box in Fig. 1) and are contrasted with the model-independent approaches previously discussed. They are named "agnostic" because they are independent from the specific algorithm applied in the preceding unimodal analysis, which can be therefore tailored to the processed type.

Even if the aforementioned early/intermediate/late taxonomy is the most diffused in literature, other taxonomies have been defined in the context of integrative (multi-omics) methods for Precision Medicine. As an example, [174, 25, 175] consider three classes: 1) *statistical-based methods*, most of which can be considered instances of the class of early integrative methods; 2) *unsupervised methods* neglecting the outcome variable during

the integration phase, that may be applied in any (early, intermediate, or late) phase and are mainly devoted to unsupervised clustering; 3) *supervised integration methods* fusing the available information to maximize the outcome prediction performance by mainly using an intermediate Multiple Kernel Learning (MKL) integration approach or a late fusion approach.

Other taxonomies [14, 32, 8, 16] consider the specific algorithm used for the integration; they recognize network-based approaches (among which deep-network based approaches, not treated in this survey), feature transformation models mainly applying an early integration approach (e.g., Principal Component Analysis - PCA, Canonical Correlation Analysis - CCA), integrative models exploiting Matrix Factorization (MF) techniques in an early integrative fashion, MKL models belonging to the class of intermediate methods, and Bayesian techniques applied in an early phase. Note that Bayesian models are not considered in this work since they have been exhaustively described in a dedicated survey [13].

Finally, the relevant survey by [16] is focused on the description of publicly available multimodal datasets in the context of multi-omics and in the critical analysis of open source integrative models. After a thorough study, the authors conclude that an objective comparison between different models is difficult, and highlight the lack of an easy-to-use multi-omics data fusion model providing a "biologist-friendly" visualization and interpretation.

Competing interests

There is NO Competing Interest.

Author contributions statement

J.G., G.V. and E.C. conceived the work, J.G. collected the literature papers, J.G. and E.C. studied the literature, selected the most relevant works and drafted them; J.G, M.M., M.N., A.Pac., G.V., E.C. wrote the paper; all the authors validated the work.

Acknowledgments

Funding

This work was supported by "Piano di sostegno alla ricerca" PSR2015-17 funded by the University of Milan and by the Transition Grant "UNIMI partenariati H2020" PSR2015-1720GVALE_01.

Key Points

- Patients similarity networks (PSN) are explainable and privacy preserving representations of patients that leverage the similarity of their clinical/biomolecular profiles to construct graphs of patients.
- Network Medicine algorithms on PSNs for patient stratification, phenotype and outcome prediction and disease risk assessment represent novel tools for Genomic and Precision Medicine
- The combination of clinical, omics and imaging bio-medical data can lead to novel PSNs able to leverage the synergy of multiple views of the patients.
- Several reviews about data integration methods in Bioinformatics and bio-medical applications have been proposed but no specific reviews about the emerging field of heterogeneous data integration methods for patient similarity networks are actually available.
- We provide a thorough review and propose a taxonomy of heterogeneous data integration methods for PSNs, together with the different patient similarity measures proposed in literature.
- We also review methods that have appeared in the machine learning literature but have not yet been applied to PSNs, thus providing a resource to navigate the vast machine learning literature existing on this topic.
- Strengths and limitations of the proposed methods are discussed to both assist researchers in the design and development of novel methods and to guide the selection of PSN integration methods for specific applications, focusing on methods that could be used to integrate very diverse datasets, including multi-omics data as well as data derived from clinical information and medical imaging.

References

1. Inke R Koenig, Oliver Fuchs, Gesine Hansen, Erika von Mutius, and Matthias V Kopp. What is precision medicine? *European respiratory journal*, 50(4), 2017.
2. Samuel J Aronson and Heidi L Rehm. Building the foundation for genomics in precision medicine. *Nature*, 526(7573):336–342, 2015.
3. Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014.
4. Dokyoon Kim, Je-Gun Joung, Kyung-Ah Sohn, Hyunjung Shin, Yu Rang Park, Marylyn D Ritchie, and Ju Han Kim. Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. *Journal of the American Medical Informatics Association*, 22(1):109–120, 2015.
5. Li Li, Wei-Yi Cheng, Benjamin S Glicksberg, Omri Gottesman, Ronald Tamler, Rong Chen, Erwin P Bottinger, and Joel T Dudley. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, 7(311):311ra174–311ra174, 2015.
6. Shraddha Pai, Shirley Hui, Ruth Isserlin, Muhammad A Shah, Hussam Kaka, and Gary D Bader. netdx: interpretable patient classification using integrated patient similarity networks. *Molecular Systems Biology*, 15(3):e8497, 2019.
7. Jessica Gliozzo, Paolo Perlasca, Marco Mesiti, Elena Casiraghi, Viviana Vallacchi, Elisabetta Vergani, Marco Frasca, Giuliano Grossi, Alessandro Petrini, Matteo Re, et al. Network modeling of patients' biomolecular profiles for clinical phenotype/outcome prediction. *Scientific reports*, 10(1):1–15, 2020.
8. Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman, and Riccardo Bellazzi. Integrated multi-omics analyses in oncology: A review of machine learning methods and tools. *Frontiers in Oncology*, 10:1030, 2020.
9. Shraddha Pai and Gary D Bader. Patient similarity networks for precision medicine. *Journal of molecular biology*, 430(18):2924–2938, 2018.
10. Noël Malod-Dognin, Julia Petschnigg, and Nataša Pržulj. Precision medicine—a promising, yet challenging road lies ahead. *Current Opinion in Systems Biology*, 7:1–7, 2018.
11. Elena Casiraghi, Dario Malchiodi, Gabriella Trucco, Marco Frasca, Luca Cappelletti, Tommaso Fontana, Alessandro Andrea Esposito, Emanuele Avola, Alessandro Jachetti, Justin Reese, et al. Explainable machine learning for early assessment of covid-19 risk prediction in emergency departments. *IEEE Access*, 8:196299–196325, 2020.

12. Thirunavukarasu Ramkumar, Shanmugasundaram Hariharan, and Shanmugam Selvamuthukumaran. A survey on mining multiple data sources. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):1–11, 2013.
13. Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17(2):167–177, 2016.
14. Vladimir Gligorijević, Noël Malod-Dognin, and Nataša Pržulj. Integrative methods for analyzing big data in precision medicine. *Proteomics*, 16(5):741–758, 2016.
15. Chen Meng, Oana A Zeleznik, Gerhard G Thallinger, Bernhard Kuster, Amin M Gholami, and Aedín C Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics*, 17(4):628–641, 2016.
16. Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.
17. Alain Rakotomamonjy, Francis R Bach, Stéphane Canu, and Yves Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9(Nov):2491–2521, 2008.
18. Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523, 2013.
19. Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583, 2014.
20. Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, and Kim-Anh Lê Cao. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019.
21. Rogers F Silva, Sergey M Plis, Tülay Adalı, Marios S Pattichis, and Vince D Calhoun. Multidataset independent subspace analysis with application to multimodal fusion. *IEEE Transactions on Image Processing*, 30:588–602, 2020.
22. Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.
23. Hung Nguyen, Sangam Shrestha, Sorin Draghici, and Tin Nguyen. Pinsplus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, 35(16):2843–2846, 2019.
24. Angelo Ciaramella, Davide Nardone, and Antonino Staiano. Data integration by fuzzy similarity-based hierarchical clustering. *BMC bioinformatics*, 21(10):1–15, 2020.
25. Cen Wu, Fei Zhou, Jie Ren, Xiaoxi Li, Yu Jiang, and Shuangge Ma. A selective review of multi-level omics data integration using variable selection. *High-throughput*, 8(1):4, 2019.
26. Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015.
27. Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.
28. Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.
29. Paul Pavlidis, Jason Weston, Jinsong Cai, and William Stafford Noble. Learning gene functional classifications from multiple data types. *Journal of computational biology*, 9(2):401–411, 2002.
30. Anneleen Daemen, Olivier Gevaert, and Bart De Moor. Integration of clinical and microarray data with kernel methods. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5411–5415. IEEE, 2007.
31. Marinka Žitnik and Blaž Zupan. Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):41–53, 2014.
32. Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, 19(2):325–340, 2018.
33. Zahra Momeni, Esmail Hassanzadeh, Mohammad Saniee Abadeh, and Riccardo Bellazzi. A survey on single and multi omics data mining methods in cancer data classification. *Journal of Biomedical Informatics*, page 103466, 2020.
34. Wei Tang, Zhengdong Lu, and Inderjit S Dhillon. Clustering with multiple graphs. In *2009 Ninth IEEE International Conference on Data Mining*, pages 1016–1021. IEEE, 2009.
35. Martin H Van Vliet, Hugo M Horlings, Marc J Van De Vijver, Marcel JT Reinders, and Lodewyk FA Wessels. Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PloS one*, 7(7):e40358, 2012.
36. Gert RG Lanckriet, Tijl De Bie, Nello Cristianini, Michael I Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.
37. Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.
38. Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C Collins, and Martin Ester. Moli: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509, 2019.
39. Shihyen Chen, Bin Ma, and Kaizhong Zhang. On the similarity metric and the distance metric. *Theoretical Computer Science*, 410(24-25):2365–2376, 2009.
40. Lluís Belanche and Jorge Orozco. Things to know about a (dis) similarity measure. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 100–109. Springer, 2011.
41. Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel Principal Component Analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.
42. Nils M Kriege, Fredrik D Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 5(1):1–42, 2020.
43. François Fouss, Kevin Francoisse, Luh Yen, Alain Pirotte, and Marco Saerens. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural networks*, 31:53–72, 2012.
44. Joon Lee, David M Maslove, and Joel A Dubin. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PloS one*, 10(5):e0127428, 2015.
45. Anis Sharafoddini, Joel A Dubin, and Joon Lee. Patient similarity in prediction models based on health data: a scoping review. *JMIR medical informatics*, 5(1):e7, 2017.
46. Ping Zhang, Fei Wang, Jianying Hu, and Robert Sorrentino. Towards personalized medicine: leveraging patient similarity and

drug similarity analytics. *AMIA Summits on Translational Science Proceedings*, 2014:132, 2014.

47. Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.
48. Sebastian Klenk, Jürgen Dippon, Peter Fritz, and Gunther Heidemann. Determining patient similarity in medical social networks. In *Proceedings of the First International Workshop on Web Science and Information Exchange in the Medical Web*, pages 6–14, 2010.
49. Bernhard Schölkopf. The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307, 2001.
50. Bin Zhu, Nan Song, Ronglai Shen, Arshi Arora, Mitchell J Machiela, Lei Song, Maria Teresa Landi, Debashis Ghosh, Nilanjan Chatterjee, Veera Baladandayuthapani, et al. Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Scientific reports*, 7(1):1–13, 2017.
51. Ya Zhang, Ao Li, Chen Peng, and Minghui Wang. Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5):825–835, 2016.
52. Jérôme Mariette and Nathalie Villa-Vialaneix. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6):1009–1015, 2018.
53. Anneleen Daemen, Dirk Timmerman, Thierry Van den Bosch, Cecilia Bottomley, Emma Kirk, Caroline Van Holsbeke, Lil Valentin, Tom Bourne, and Bart De Moor. Improved modeling of clinical data with kernel methods. *Artificial intelligence in medicine*, 54(2):103–114, 2012.
54. Peifeng Ruan, Ya Wang, Ronglai Shen, and Shuang Wang. Using association signal annotations to boost similarity network fusion. *Bioinformatics*, 35(19):3718–3726, 2019.
55. Shuhao Li, Limin Jiang, Jijun Tang, Nan Gao, and Fei Guo. Kernel fusion method for detecting cancer subtypes via selecting relevant expression data. *Frontiers in Genetics*, 11, 2020.
56. Giorgio Valentini, Giuliano Armano, Marco Frasca, Jianyi Lin, Marco Mesiti, and Matteo Re. RANKS: a flexible tool for node label ranking and classification in biological networks. *Bioinformatics*, 32(18):2872–2874, 06 2016.
57. Yasin Ilkagan Tepeli, Ali Burak Ünal, Furkan Mustafa Akdemir, and Oznur Tastan. Pamogk: A pathway graph kernel based multi-omics approach for patient clustering. *Bioinformatics*, 2020.
58. Yuqi Wen, Xinyu Song, Bowei Yan, Xiaoxi Yang, Lianlian Wu, Dongjin Leng, Song He, and Xiaochen Bo. Multi-dimensional data integration algorithm based on random walk with restart. *BMC bioinformatics*, 22(1):1–22, 2021.
59. Pablo A Jaskowiak, Ricardo JGB Campello, and Ivan G Costa. Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis. *IEEE/ACM transactions on computational biology and bioinformatics*, 10(4):845–857, 2013.
60. Pablo A Jaskowiak, Ricardo JGB Campello, and Ivan G Costa. On the selection of appropriate distances for gene expression data clustering. *BMC bioinformatics*, 15(2):1–17, 2014.
61. Chihyun Park, Jaegyoon Ahn, Hyunjin Kim, and Sanghyun Park. Integrative gene network construction to analyze cancer recurrence using semi-supervised learning. *PLOS ONE*, 9(1):1–9, 01 2014.
62. Karolina Sienkiewicz, Jinyu Chen, Ajay Chatrath, John T Lawson, Nathan C Sheffield, Louxin Zhang, and Aakrosh Ratan. Detecting molecular subtypes from multi-omics datasets using sumo. *Cell Reports Methods*, page 100152, 2022.
63. Hongjiu Zhang, Fan Zhu, Hiroko H Dodge, Gerald A Higgins, Gilbert S Omenn, Yuanfang Guan, and Alzheimer's Disease Neuroimaging Initiative. A similarity-based approach to leverage multi-cohort medical data on the diagnosis and prognosis of alzheimer's disease. *GigaScience*, 7(7):giy085, 2018.
64. Fayao Liu, Luping Zhou, Chunhua Shen, and Jianping Yin. Multiple kernel learning in the primal for multimodal alzheimer's disease classification. *IEEE journal of biomedical and health informatics*, 18(3):984–990, 2013.
65. Mingxin Tao, Tianci Song, Wei Du, Siyu Han, Chunman Zuo, Ying Li, Yan Wang, and Zekun Yang. Classifying breast cancer subtypes using multiple kernel learning based on omics data. *Genes*, 10(3):200, 2019.
66. Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.
67. Ya Zhang, Ao Li, Jie He, and Minghui Wang. A novel mkl method for gbm prognosis prediction by integrating histopathological image and multi-omics data. *IEEE journal of biomedical and health informatics*, 24(1):171–179, 2019.
68. Nello Cristianini and Bernhard Scholkopf. Support vector machines and kernel methods: the new generation of learning machines. *Ai Magazine*, 23(3):31–31, 2002.
69. Dongdong Sun, Ao Li, Bo Tang, and Minghui Wang. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Computer methods and programs in biomedicine*, 161:45–53, 2018.
70. Fabio Aiolli and Michele Donini. Easymkl: a scalable multiple kernel learning algorithm. *Neurocomputing*, 169:215–224, 2015.
71. Zenglin Xu, Rong Jin, Haiqin Yang, Irwin King, and Michael R Lyu. Simple and efficient multiple kernel learning by group lasso. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 1175–1182. Citeseer, 2010.
72. Taiji Suzuki and Ryota Tomioka. Spicymkl: a fast algorithm for multiple kernel learning with thousands of kernels. *Machine learning*, 85(1-2):77–108, 2011.
73. Fei Yan, Josef Kittler, Krystian Mikolajczyk, and Atif Tahir. Non-sparse multiple kernel fisher discriminant analysis. *The Journal of Machine Learning Research*, 13(1):607–642, 2012.
74. Gaston Baudat and Fatiha Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.
75. Cheng Soon Ong and Alexander Zien. An automated combination of kernels for predicting protein subcellular localization. In *International Workshop on Algorithms in Bioinformatics*, pages 186–197. Springer, 2008.
76. Nora K Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, 2015.
77. Xinwang Liu, Yong Dou, Jianping Yin, Lei Wang, and En Zhu. Multiple kernel k-means clustering with matrix-induced regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.
78. Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh. Multiple kernel learning for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1147–1160, 2011.
79. Xiaofei He and Partha Niyogi. Locality preserving projections. *Advances in neural information processing systems*, 16(16):153–160, 2004.
80. Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.
81. Giorgio Valentini, Alberto Paccanaro, Horacio Caniza, Alfonso E. Romero, and Matteo Re. An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artificial Intelligence in Medicine*, 61(2):63–78,

2014.

82. Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.
83. Giulia Tini, Luca Marchetti, Corrado Priami, and Marie-Pier Scott-Boyer. Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in bioinformatics*, 20(4):1269–1279, 2019.
84. Evan G Williams, Yibo Wu, Pooja Jha, Sébastien Dubuis, Peter Blattmann, Carmen A Argmann, Sander M Houten, Tiffany Amariuta, Witold Wolski, Nicola Zamboni, et al. Systems proteomics of liver mitochondria function. *Science*, 352(6291), 2016.
85. Anne Zufferey, Mark Ibberson, Jean-Luc Reny, Séverine Nolli, Domitille Schvartz, Mylène Docquier, Ioannis Xenarios, Jean-Charles Sanchez, and Pierre Fontana. New molecular insights into modulation of platelet reactivity in aspirin-treated patients using a network-based approach. *Human genetics*, 135(4):403–414, 2016.
86. DCFR Koboldt, Robert Fulton, Michael McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua McMichael, Lucinda Fulton, David Dooling, Li Ding, Elaine Mardis, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.
87. Tianle Ma and Aidong Zhang. Integrate multi-omic data using affinity network fusion (anf) for cancer patient clustering. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 398–403. IEEE, 2017.
88. Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli. Novel fisher discriminant classifiers. *Pattern Recognition*, 45(10):3725–3737, 2012.
89. Shuhui Liu and Xuequn Shang. Hierarchical similarity network fusion for discovering cancer subtypes. In *International Symposium on Bioinformatics Research and Applications*, pages 125–136. Springer, 2018.
90. Limin Jiang, Yongkang Xiao, Yijie Ding, Jijun Tang, and Fei Guo. Discovering cancer subtypes via an accurate fusion strategy on multiple profile data. *Frontiers in genetics*, 10:20, 2019.
91. Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.
92. Alberto Valdeolivas, Laurent Tichit, Claire Navarro, Sophie Perrin, Gaelle Odelin, Nicolas Levy, Pierre Cau, Elisabeth Remy, and Anaïs Baudot. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35(3):497–505, 2019.
93. Nimrod Rappoport and Ron Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, 46(20):10546–10562, 2018.
94. Nimrod Rappoport and Ron Shamir. Nemo: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35(18):3348–3356, 2019.
95. Jian Liu, Shuguang Ge, Yuhu Cheng, and Xuesong Wang. Multi-view spectral clustering based on multi-smooth representation fusion for cancer subtype prediction. *Frontiers in Genetics*, page 1574, 2021.
96. Feiping Nie, Jing Li, Xuelong Li, et al. Self-weighted multiview clustering with multiple graphs. In *IJCAI*, pages 2564–2570, 2017.
97. Qing Feng, Meilei Jiang, Jan Hannig, and JS Marron. Angle-based joint and individual variation explained. *Journal of multivariate analysis*, 166:241–265, 2018.
98. Zi Yang and George Michailidis. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1–8, 2016.
99. Johan A Westerhuis, Theodora Kourti, and John F MacGregor. Analysis of multiblock and hierarchical pca and pls models. *Journal of chemometrics*, 12(5):301–321, 1998.
100. Huichen Zhu, Gen Li, and Eric F Lock. Generalized integrative principal component analysis for multi-type data with block-wise missing structure. *Biostatistics*, 21(2):302–318, 2020.
101. Giovanni Ciriello, Michael L Gatza, Andrew H Beck, Matthew D Wilkerson, Suhn K Rhie, Alessandro Pastore, Hailei Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519, 2015.
102. Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1), 2009.
103. Georg Lenz, George W Wright, NC Tolga Emre, Holger Kohlhammer, Sandeep S Dave, R Eric Davis, Shannon Carty, Lloyd T Lam, AL Shaffer, Wenming Xiao, et al. Molecular subtypes of diffuse large b-cell lymphoma arise by distinct genetic pathways. *Proceedings of the National Academy of Sciences*, 105(36):13520–13525, 2008.
104. Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257, 2011.
105. Imene Garali, Isaac M Adanyeguh, Farid Ichou, Vincent Perlbarg, Alexandre Seyer, Benoit Colsch, Ivan Moszer, Vincent Guillemot, Alexandra Durr, Fanny Mochel, et al. A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. *Briefings in bioinformatics*, 19(6):1356–1369, 2018.
106. Age K Smilde, Johan A Westerhuis, and Sijmen de Jong. A framework for sequential multiblock component methods. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17(6):323–337, 2003.
107. M. De Tayrac, L^ E.S. Aubry, M. Mosser, and J.F. Husson. Simultaneous analysis of distinct omics data sets with integration of biological knowledge: Multiple factor analysis approach. *BMC Genomics*, 10:32, 2009.
108. NE Kucukboyaci, N Kemmotsu, KM Leyden, HM Girard, ES Tecoma, VJ Iragui, and CR McDonald. Integration of multimodal mri data via pca to explain language performance. *NeuroImage: Clinical*, 5:197–207, 2014.
109. Maxime Chamberland, Erika P Raven, Sila Genc, Kate Duffy, Maxime Descoteaux, Greg D Parker, Chantal MW Tax, and Derek K Jones. Dimensionality reduction of diffusion mri measures for improved tractometry of the human brain. *NeuroImage*, 200:89–100, 2019.
110. Bryce Landon Geeraert, Maxime Chamberland, R Marc Lebel, and Catherine Lebel. Multimodal principal component analysis to identify major features of white matter structure and links to reading. *bioRxiv*, 2020.
111. Bradley Worley and Robert Powers. A sequential algorithm for multiblock orthogonal projections to latent structures. *Chemometrics and Intelligent Laboratory Systems*, 149:33–39, 2015.
112. Li Zhang, Chenkai Lv, Yaqiong Jin, Ganqi Cheng, Yibao Fu, Dongsheng Yuan, Yiran Tao, Yongli Guo, Xin Ni, and Tieliu Shi. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Frontiers in genetics*, 9:477, 2018.
113. Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.
114. Erica Ponzi, Magne Thoresen, Therese Haugdahl Nøst, and Kajsa Møllersen. Integrative, multi-omics, analysis of blood samples improves model predictions: applications to cancer. *bioRxiv*, pages 2020–10, 2021.
115. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
116. Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixomics: An r package for 'omics feature selection and multiple data

integration. *PLoS computational biology*, 13(11):e1005752, 2017.

117. Sini Isokäántä, Eetu Kari, Angela Buchholz, Liqing Hao, Siegfried Schobesberger, Annele Virtanen, and Santtu Mikkonen. Comparison of dimension reduction techniques in the analysis of mass spectrometry data. *Atmospheric Measurement Techniques*, 13(6):2995–3022, 2020.
118. Nicolas Gillis. Sparse and unique nonnegative matrix factorization through data preprocessing. *The Journal of Machine Learning Research*, 13(1):3349–3386, 2012.
119. Yifeng Li. Advances in multi-view matrix factorizations. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3793–3800. IEEE, 2016.
120. Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(9), 2004.
121. Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic acids research*, 40(19):9379–9391, 2012.
122. Cécile Chauvel, Alexei Novoloaca, Pierre Veyre, Frédéric Reynier, and Jérémie Becker. Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in Bioinformatics*, 21(2):541–552, 2020.
123. Ronglai Shen, Adam B Olshen, and Marc 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):2906–2912, 2009.
124. Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616, 2013.
125. Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, 2012.
126. Ying-Lian Gao, Mi-Xiao Hou, Jin-Xing Liu, and Xiang-Zhen Kong. An integrated graph regularized non-negative matrix factorization model for gene co-expression network analysis. *IEEE Access*, 7:126594–126602, 2019.
127. Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1548–1560, 2010.
128. Fei Wang, Tao Li, and Changshui Zhang. Semi-supervised clustering via matrix factorization. In *Proc. SIAM Int. Conf. on Data Mining*, 2008.
129. F Vitali, S Marini, D Pala, A Demartini, S Montoli, A Zambelli, and R Bellazzi. Patient similarity by joint matrix trifactorization to identify subgroups in acute myeloid leukemia. *JAMIA open*, 1(1):75–86, 2018.
130. Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):1–17, 2020.
131. Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.
132. Rogers F Silva, Sergey M Plis, Jing Sui, Marios S Pattichis, Tülay Adalı, and Vince D Calhoun. Blind source separation for unimodal and multimodal brain networks: A unifying framework for subspace modeling. *IEEE journal of selected topics in signal processing*, 10(7):1134–1149, 2016.
133. Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
134. Guoxu Zhou, Qibin Zhao, Yu Zhang, Tülay Adalı, Shengli Xie, and Andrzej Cichocki. Linked component analysis from matrices to high-order tensors: Applications to biomedical data. *Proceedings of the IEEE*, 104(2):310–331, 2016.
135. Tulay Adali, Matthew Anderson, and Geng-Shen Fu. Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. *IEEE Signal Processing Magazine*, 31(3):18–33, 2014.
136. Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.
137. Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.
138. Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.
139. V Calhoun, Tulay Adali, and Jingyu Liu. A feature-based approach to combine functional mri, structural mri and eeg brain imaging data. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3672–3675. IEEE, 2006.
140. Matthias Moosmann, Tom Eichele, Helge Nordby, Kenneth Hugdahl, and Vince D Calhoun. Joint independent component analysis for simultaneous eeg–fmri: principle and simulation. *International Journal of Psychophysiology*, 67(3):212–221, 2008.
141. Taesu Kim, Torbjørn Eltoft, and Te-Won Lee. Independent vector analysis: An extension of ica to multivariate components. In *International conference on independent component analysis and signal separation*, pages 165–172. Springer, 2006.
142. Tülay Adali, MABS Akhonda, and Vince D Calhoun. Ica and iva for data fusion: An overview and a new approach based on disjoint subspaces. *IEEE sensors letters*, 3(1):1–4, 2018.
143. Marinka Žitnik and Blaž Zupan. Matrix factorization-based data fusion for drug-induced liver injury prediction. *Systems Biomedicine*, 2(1):16–22, 2014.
144. Tin Nguyen, Rebecca Tagett, Diana Diaz, and Sorin Draghici. A novel approach for data integration and disease subtyping. *Genome research*, 27(12):2025–2039, 2017.
145. John H Phan, Ryan Hoffman, Sonal Kothari, Po-Yen Wu, and May D Wang. Integration of multi-modal biomedical data to predict cancer grade and patient survival. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 577–580. IEEE, 2016.
146. Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.
147. Matthew D Wilkerson and D Neil Hayes. Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573, 2010.
148. Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1):91–118, 2003.
149. Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.
150. Leonard Kaufmann and Peter Rousseeuw. Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416, 01 1987.
151. Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r.

Bioinformatics, 24(5):719–720, 2008.

152. Dingming Wu, Dongfang Wang, Michael Q Zhang, and Jin Gu. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC genomics*, 16(1):1022, 2015.
153. Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.
154. David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.
155. Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.
156. Simon Lovestone, Paul Francis, Iwona Kloszewska, Patrizia Mecocci, Andrew Simmons, Hilkka Soininen, Christian Spenger, Magda Tsolaki, Bruno Vellas, Lars-Olof Wahlund, et al. Addneuromed—the european collaboration for the discovery of novel biomarkers for alzheimer's disease. *Annals of the New York Academy of Sciences*, 1180(1):36–46, 2009.
157. Sherry-Ann Brown. Patient similarity: emerging concepts in systems and precision medicine. *Frontiers in physiology*, 7:561, 2016.
158. Leyu Dai, He Zhu, and Dianbo Liu. Patient similarity: methods and applications. *arXiv preprint arXiv:2012.01976*, 2020.
159. Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021.
160. Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68, 2011.
161. Raffaele Giancarlo, Giosuè Lo Bosco, and Luca Pinello. Distance functions, clustering algorithms and microarray data analysis. In *International Conference on Learning and Intelligent Optimization*, pages 125–138. Springer, 2010.
162. Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang, Jean Yee Hwa Yang, and Pengyi Yang. Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in bioinformatics*, 20(6):2316–2326, 2019.
163. Ian T Jolliffe. Principal component analysis: a beginner's guide—ii. pitfalls, myths and extensions. *Weather*, 48(8):246–253, 1993.
164. Yunjin Choi, Jonathan Taylor, and Robert Tibshirani. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *The Annals of Statistics*, pages 2590–2617, 2017.
165. Gen Li and Irina Gaynanova. A general framework for association analysis of heterogeneous data. *The Annals of Applied Statistics*, 12(3):1700–1726, 2018.
166. P Campadelli, E Casiraghi, C Ceruti, and A Rozza. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015, 2015.
167. Paolo Perlasca, Marco Frasca, Cheick Ba, Marco Notaro, Alessandro Petrini, Elena Casiraghi, Giuliano Grossi, Jessica Gliozzo, Giorgio Valentini, and Marco Mesiti. Unipred-web: a web tool for the integration and visualization of biomolecular networks for protein function prediction. *BMC Bioinformatics*, 20, 12 2019.
168. Paolo Perlasca, Marco Frasca, Cheick Tidiane Ba, Jessica Gliozzo, Marco Notaro, Mario Pennacchioni, Giorgio Valentini, and Marco Mesiti. Multi-resolution visualization and analysis of biomolecular networks through hierarchical community detection and web-based graphical tools. *PLOS ONE*, 15(12):1–28, 12 2020.
169. Sylvia Richardson, George C Tseng, and Wei Sun. Statistical methods in integrative genomics. *Annual review of statistics and its application*, 3:181–209, 2016.
170. Qing Zhao, Xingjie Shi, Jian Huang, Jin Liu, Yang Li, and Shuangge Ma. Integrative analysis of '-omics' data using penalty functions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):99–108, 2015.
171. Wenting Wang, Veerabhadran Baladandayuthapani, Jeffrey S Morris, Bradley M Broom, Ganiraju Manyam, and Kim-Anh Do. ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159, 2013.
172. Ruoqing Zhu, Qing Zhao, Hongyu Zhao, and Shuangge Ma. Integrating multidimensional omics data for cancer outcome. *Biostatistics*, 17(4):605–618, 2016.
173. Cen Wu, Qingzhao Zhang, Yu Jiang, and Shuangge Ma. Robust network-based analysis of the associations between (epi) genetic measurements. *Journal of multivariate analysis*, 168:119–130, 2018.
174. Sijia Huang, Kumardeep Chaudhary, and Lana X Garmire. More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, 8:84, 2017.
175. Elad Noor, Sarah Cherkaoui, and Uwe Sauer. Biological insights through omics data integration. *Current Opinion in Systems Biology*, 15:39–47, 2019.

Jessica Gliozzo. Jessica Gliozzo is a PhD student in Computer Science enrolled in the Collaborative Doctoral Partnership program between the University of Milan and the Joint Research Center of European Commission. Her latest research works comprehend the development of a multi-modal semi-supervised method based on patient similarity networks for patients' outcome prediction; the application of deep neural networks to predict the tissue-specific activity status of cis-regulatory regions in the genome (i.e. promoters and enhancers); the use of compression methods to obtain compact representations of convolution neural networks in the biological domain (e.g. ki67 and TIL-index prediction), showing their adavantages when limited computational resources are available. She is author of a few works in the fields of Machine Learning and Bioinformatics.

Marco Mesiti. He is associate professor at the Department of Computer Science Giovanni degli Antoni, Università degli Studi di Milano. He has got a master and phD degree from the University of Genova in 1998 and 2003. His research interest are in the integration, querying and visualization of different kinds of information (structured and semi-structured) according to different data models (relational, graph, and nosql). Moreover, he has involved in different projects for protein network integration, protein function prediction, and protein networks visualization. On these topics is ha published more than 100 articles in international conferences and journals. He is associate editor for the Springer Data Science and Engineering Journal and MDPI Applied Sciences.

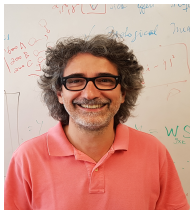
Marco Notaro. He is a postdoctoral fellow at the Computer Science Department of Milan University. His research interests touch the fields of Bioinformatics, Computational Biology, Biological Network and Machine Learning. His main expertise is the analysis and construction of complex biomolecular networks and the design and implementation of output-structured

learning algorithms to discover novel gene-disease associations or to predict novel protein function. His Ph.D. paper was awarded by International Medical Informatics Association as one of the best five papers of 2017 in the field of Medical Informatics.

Alessandro Petrini. He is a postdoctoral researcher at the department of Computer Science of Università degli Studi di Milano. He is currently a member of the laboratory of Bioinformatics and Computational Biology - AnacletoLab - and his main research is focused on High Performance Computing and Machine Learning. He is author of more than 30 articles in international journals and conferences. He designed and developed parallel and accelerated ML algorithms for image and video processing / encoding / compression, omics analisys, graph modeling and analysis, data visualization, neural network compression, MRI volumes processing and analysis.

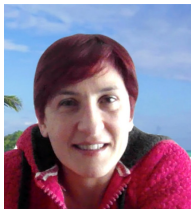
Alex Patak. Alex Patak, PhD MD, graduated in Medicine and Surgery at the School of Medicine at "Universidad Autónoma de Barcelona", Barcelona (Spain) and holds a Master in Medical Bioengineering from the "Universidad Politécnica de Cataluña". At the Instituto Municipal de Investigación Médica (Barcelona) he has been working on Expert Systems for medical diagnostic and did his PhD on computer-assisted medical education after a stage at Dartmouth Medical College in Vermont (USA). Since 1994 works at the Joint Research Centre in Ispra (Italy) where he has been working on Three-dimensional medical imaging, and from 2003 to 2017 was responsible for the Bioinformatics team at the Molecular Biology and Genomics Unit of the Institute for Health and Consumer Protection in Ispra. He is now a team leader at Knowledge for Health & Consumer Safety and is responsible for the Collaborative Doctoral Partnership programme in Genomics and Bioinformatics, working on the application of artificial intelligence to omics data and Microbiome.

Antonio Puertas-Gallardo. Antonio Puertas Gallardo is an IT Project Manager at the Joint Research Center (JRC) of the European Commission. He provides High Performance Computing (HPC) support to bioinformatics members of the Knowledge for Health and Consumer Safety Unit at JRC, and he has recently begun to collaborate with the unit's data scientists on Natural Language Processing and Machine Learning.



Alberto Paccanaro. Alberto Paccanaro is full Professor in Machine Learning and Computational Biology at the School of Applied Mathematics of the Fundação Getúlio Vargas in Rio de Janeiro and at the Department of Computer Science at Royal Holloway University of London, where he is also Director of the Centre for Systems and Synthetic Biology. He completed his undergraduate studies in Computer Science at the University of Milan and received his PhD from the University of Toronto in 2002. His research interests are in applying and developing machine learning algorithms for solving problems in molecular biology, medicine and pharmacology and he has led a number of international research projects in this area.

Giorgio Valentini. Giorgio Valentini is a full Professor at the Department of Computer Science, University of Milan (UNIMI). Director for UNIMI of the European doctorate in Genomics and Bioinformatics in collaboration with the Joint Research Center of the European Union. Director of AnacletoLab, Computational Biology and Bioinformatics Laboratory of the Department of Computer Science of the University of Milano. He has been PI in several national and international research projects funded by public and private institutions in the area of Bioinformatics, Machine learning and Big Data analytics. He is author of over 150 scientific publications with peer-review in collaboration with several research groups in Europe and America in the field of Bioinformatics, Computational Biology and Machine Learning.



Elena Casiraghi. She is associate professor at the Department of Computer Science Giovanni degli Antoni, Università degli Studi di Milano. She is co-lead of the AnacletoLab, Computational Biology and Bioinformatics Laboratory of the Department of Computer Science of the University of Milano. Her research interests are focused on the development of applications that process biomedical data (images, multi-omics, clinical) for causal inference, inductive inference, machine learning, and pattern recognition. She currently cooperates with several hospitals and reserach centers in Italy (Istituto Nazionale dei Tumori, Ospedale Policlinico e Regina Margherita, Ospedale San Raffaele, Humanitas), Europe (Berlin Institute of Health at Charité, grupo espanol de investigacion de sarcomas, groupe sarcomes francais), and United States (Berkeley Laboratories, Jackson Laboratories) and she is the scientific supervisor for UNIMI of all the researches in cooperation with the N3C Enclave (USA, `https://covid.cd2h.org/`) created and funded by the NIH.

Table 6. PCA-based and CCA-based input data-fusion methods. For each method, the table reports: the name/acronym with the corresponding reference paper; whether it requires the same set of patients across all data modalities (i.e. "Matched Samples"); the dataset used to develop and evaluate the approach in the reference paper and the corresponding sample cardinality and data types composing the dataset; the exploited integration method; the application task and the code availability (with link to the repository and programming languages for which the code is available).

Name	Matched Samples	Dataset	Sample Cardinality	Data type	Integration approach	Task	Code and Language
CPCA [99]	x	Simulated	Not provided	Numeric	PCA	Data Analysis	
CPCA for missing data [100]	x	Human Mortality Database (Italy + Switzerland)	143	exposure-to-risk	PCA	Data Analysis	
JIVE [18]	x	TCGA BIC	348	mRNA miRNA methy RPPA	PCA	Unsupervised Clustering (Patient subtype identification)	R code
aJIVE [97]	x	TCGA extract from [101]	616	mRNA miRNA somatic mutation CNV RPPA	PCA	Data Analysis and Unsupervised Clustering (Patient subtype identification)	R code
MCCA [102]	x	DLBCL Dataset [103]	203	mRNA array CGH measurements	CCA	Data analysis	R code
RGCCA [104] SGCCA [19] [105]	x	SCA Dataset Private	67 SCA + 35 Healthy	pons volume metabolic features	CCA	Data Analysis	RGCCA/SGCCA R code
DIABLO [20]	x	TCGA COAD TCGA KIRC TCGA GBM TCGA LUSC TCGA BRCA	92 122 213 106 989	mRNA miRNA methy	(SG)CCA	Data Analysis and Supervised Clustering (Patient's Survival)	R code

Abbreviations

CCA: Canonical Correlation Analysis; **CGH**: Comparative Genomic Hybridization; **CNV**: Copy Number Variation; **DLBCL**: Diffuse Large B-Cell Lymphoma; **methy**: DNA methylation; **miRNA**: micro RNA; **mRNA**: messenger RNA; **PCA**: Principal Component Analysis; **RPPA**: Reverse Phase Protein Array; **SCA**: SpinoCerebellar Ataxia; **(SG)CCA**: (Sparse Generalized) Canonical Correlation Analysis; **TCGA+*cancer code***: The Cancer Genome Atlas+ *link to complete cancer codes*.

Table 7. MF-based input data-fusion methods. For each method, the table reports: the name/acronym with the corresponding reference paper; whether it requires the same set of patients across all data modalities (i.e. "Matched Samples"); the dataset used to develop and evaluate the approach in the reference paper and the corresponding sample cardinality and data types composing the dataset; the exploited integration method; the application task and the code availability (with link to the repository and programming languages for which the code is available).

Name	Matched Samples	Dataset	Sample Cardinality	Data type	Integration approach	Task	Code and Language
MFA [107]		Brain Cancer Dataset Private	Not provided	multi-omics	MF	Data Analysis	R code
jNMF [121]	x	TCGA OV	385	mRNA miRNA methy	NMF	Data Analysis	R code
iNMF [98]	x	TCGA OV	592	mRNA miRNA methy	NMF	Unsupervised Clustering (Patient subtype identification)	Python code
iGMFNA [126]	x	TCGA CHOL TCGA PAAD	45 180	mRNA methy CNV	NMF	Data Analysis	
MOFA+ [130]		Private	Not provided	multi-omics	NMF	Data Analysis	Python and R code
iCluster [123] and iCluster+ [131]	x	TCGA CRC	189	exome sequence mRNA methy CNV	NMF	Unsupervised Clustering (Patient subtype identification)	R code
DFMF^3 [31] [143]				GO terms GO annotations drugs tissue samples DILI potentials	MTF	Unsupervised Clustering (hepatotoxic risk associated with individual drugs)	Python code
MaDDa [129]		TCGA BRCA, BioGRID KEGG Disease Ontology DisGeNET	200	gene-gene interactions, gene-pathway associations disease-disease relationships, disease-gene associations, disease-pathway relations	MTF	Unsupervised Clustering (Patient subtype identification)	Matlab code
DS-ICA [142]	x	Private	38	features from EEG and fMRI images	ICA	Data Analysis	
MISA [21]	x	Private	1001	EEG images sMRI and fMRI images	BSS	Data Analysis	MATLAB code

Abbreviations

BSS: Blind Source Separation; **CNV**: Copy Number Variation; **DILI**: Drug-Induced Liver Injury; **EEG**: Electroencephalography; **fMRI**: functional Magnetic Resonance Imaging; **GO**: Gene Ontology; **ICA**: Independent Component Analysis; **KEGG**: Kyoto Encyclopedia of Genes and Genomes; **methy**: DNA methylation; **miRNA**: micro RNA; **MF**: Matrix Factorization; **mRNA**: messenger RNA; **MTF**: Matrix Tri-Factorization; **NMF**: Non-negative Matrix Factorization; **sMRI**: structural Magnetic Resonance Imaging; **TCGA+*cancer code***: The Cancer Genome Atlas+ *link to complete cancer codes*.

Table 8. Output-fusion methods. For each method, the table reports: the name/acronym with the corresponding reference paper; the dataset used to develop and evaluate the approach in the reference paper and the corresponding sample cardinality and data types composing the dataset; the exploited integration method; the application task and the code availability (with link to the repository and programming languages for which the code is available).

Name	Dataset	Sample Cardinality	Data type	Integration approach	Task	Code and Language
COCA [22]	TCGA AML TCGA BIC TCGA COAD TCGA READ TCGA GBM TCGA KIRC TCGA LUSC TCGA OV TCGA UCEC TCGA BLCA TCGA LUAD TCGA HNSC	161 834 182 73 195 475 238 329 345 120 270 305	DNA sequence mRNA miRNA methy CNV RPPA	Consensus Clustering	Unsupervised Clustering (Patient subtype identification)	ConsensusClusterPlus R code
PINS/PINSPlus [23]	34 TCGA datasets 2 Metabric datasets	12158	mRNA miRNA methy	Consensus Clustering	Unsupervised Clustering (Patient subtype identification)	R code
SUMO [62]	TCGA extract from NEMO [94] (Table 4)	3168 across	mRNA miRNA methy	Consensus Clustering	Unsupervised Clustering (Patient subtype identification)	Python code
FH-Clust [24]	TCGA AML TCGA BIC TCGA COAD TCGA GBM TCGA KIRC TCGA LIHC TCGA LUSC TCGA SKCM TCGA OV TCGA SARC	170 621 220 274 183 367 341 448 287 257	mRNA miRNA methy	Consensus Clustering	Unsupervised Clustering (Patients' clusters related to known Survival)	R code
[145]	TCGA KIRC TCGA OV TCGA KIRC TCGA OV	418 250 220 160	histopathological image RNA-seq data	Stacked Generalization: linear regression of unimodal classifiers	Supervised Classification (Cancer grade < 3 vs Cancer grade >= 3) Supervised Classification (Known Survival < 5 vs Known Survival >= 5)	
[63]	ADNI Phase 1 [155] AddNeuroMed study [156]	628 for training 94 for validation 88 for testing	demographic data APOE e4 allele information anatomical brain features from 1.5T MRI scans	Average	Supervised Multiclass classification (HC vs MCI vs AD)	Python code

Abbreviations

AD: Alzheimer's Disease; **ADNI**: Alzheimer's Disease Neuroimaging Initiative; **CNV**: Copy Number Variation; **HC**: Healthy Control; **MCI**: Mild Cognitive Impairment; **methy**: DNA methylation; **miRNA**: micro RNA; **MRI**: Magnetic Resonance Imaging; **mRNA**: messenger RNA; **RPPA**: Reverse-Phase Protein Arrays; **TCGA+*cancer code***: The Cancer Genome Atlas+ *link to complete cancer codes*.