Computational and Structural Biotechnology Journal 19 (2021) 3735–3746

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Integration strategies of multi-omics data for machine learning analysis

Milan Picard [a], Marie-Pier Scott-Boyer [a], Antoine Bodein [a], Olivier Périn [b], Arnaud Droit [a,*]

[a] Molecular Medicine Department, CHU de Québec Research Center, Université Laval, Québec, QC, Canada
[b] Digital Sciences Department, L'Oréal Advanced Research, Aulnay-sous-bois, France

ARTICLE INFO

Article history:
Received 30 March 2021
Received in revised form 17 June 2021
Accepted 21 June 2021
Available online 22 June 2021

Keywords:
Multi-omics
Multi-view
Integration strategy
Machine learning
Deep learning
Network

ABSTRACT

Increased availability of high-throughput technologies has generated an ever-growing number of omics data that seek to portray many different but complementary biological layers including genomics, epigenomics, transcriptomics, proteomics, and metabolomics. New insight from these data have been obtained by machine learning algorithms that have produced diagnostic and classification biomarkers. Most biomarkers obtained to date however only include one omic measurement at a time and thus do not take full advantage of recent multi-omics experiments that now capture the entire complexity of biological systems.

Multi-omics data integration strategies are needed to combine the complementary knowledge brought by each omics layer. We have summarized the most recent data integration methods/ frameworks into five different integration strategies: early, mixed, intermediate, late and hierarchical. In this mini-review, we focus on challenges and existing multi-omics integration strategies by paying special attention to machine learning applications.

Contents

* Corresponding author.
E-mail address: Arnaud.Droit@crchudequebec.ulaval.ca (A. Droit).

https://doi.org/10.1016/j.csbj.2021.06.030

1. Introduction

The advent of powerful and inexpensive screening technologies [1] recently produced huge amounts of biological data that opened the way to a new era of therapeutics and personalized medicine [2]. Treatment efficiency and adverse effects can differ vastly between individuals due to differences in age, sex, genetics and environmental factors (e.g., anthropometric and metabolic statuses; dietary and lifestyle habits [3,4]). The aim of precision medicine is thus to design the most appropriate intervention based on the biological information of each individual [5].

Clinical information and omics data can be directly retrieved from databases or collected with screening technologies for disease [6], class prediction [7], biomarkers discovery [8], disease subtyping [6], improved system biology knowledge [9], drug repurposing and so on. Each type of omics data is specific to a single "layer" of biological information such as genomics, epigenomics, transcriptomics, proteomics, metabolomics, and provides a complementary medical perspective of a biological system or an individual [1]. In the past, single-omics studies were done in hope of discovering the causes of pathologies and helping select an appropriate treatment. We now realize that such approaches are overly simplistic. Most diseases affect complex molecular pathways where different biological layers interact with each other. Hence the need for multi-omics studies that can encompass several layers at once and draw a more complete picture of a given phenotype [10]. With multiple omics, faint patterns in gene expression data can be reinforced with epigenomics [11] for example. Complementary information can be exploited to better explain classification results [12], improve prediction performances [13,14] or understand complex molecular pathways [15] that would be out of grasp for single-omics studies. However, multi-omics studies include data that differ in type, scale and distribution, with often thousands of variables and only few samples. Additionally, biological datasets are complex, noisy, with potential errors due to measurement mistakes or unique biological deviations. Discovering pertinent information and integrating the omics into a meaningful model is therefore difficult and a great number of methods and strategies have been developed in recent years to tackle this challenge [6,16]. If the integration is not done correctly, adding more omics might not result in a significant increase of performance, but will increase the complexity of the problem along with computational time.

A way to classify existing approaches is therefore needed in order to select appropriate methods and find good practices. Zitnik *et al. (2019)* [17] differentiated two types of integration. Either horizontal integration, which studies the same omics across different groups of samples or vertical integration, which examines multiple omics variables on the same samples. In this mini-review, we focus on vertical integration, where each omics dataset (or omics block) has the same rows (samples), but different variables (omics features). We also assume that the datasets are already processed, normalized or scaled depending on the omics's type. In the current literature, general reviews on vertical integration [18–21] often classify methods based on mathematical aspects, usually whether or not they are Bayesian, Network-based, deep learning-based, kernel-based, or matrix factorization-based methods. In this review, we are also interested in presenting general integration strategies for multi-omics datasets and sorting methods based on how they are used, which we believe to be more intuitive and practical than using their underlying mathematical basis as only classification. For that purpose, Ritchie *et al. (2015)* [22] introduced three integration approaches for vertical integration (called *meta*-dimensional in the review): concatenation-based integration combines datasets before analysis, transformation-based integration performs mapping or data transformation of each datasets before analysing the transformed datasets and model-based integration that performs analysis separately on each dataset before combining the results. Our work continues and extends those three categories to five distinct integration strategies and rename them to early (concatenation-based), mixed (transformation-based), late (model-based), intermediate and hierarchical. The names for the integration strategies are inspired from Zitnik *et al. (2019)* [17], but it is important to clarify that their classification (early, late and intermediate) is mainly different from ours.

This review presents the most recent advances in multi-omics analysis, with a special focus on the integration strategy. It is intended for computational biologists looking for general approaches and ideas in handling their omics datasets. For more specific reviews on multi-omics integration, we suggest several other reviews either methodologically oriented on subject such as variable selection [23], dimensionality reduction [24], autoencoders [25], clustering [26,27] or network-based methods [10,11,28,29], or biologically oriented on subjects like metabolomics [30,31], phosphoproteomics [32], toxicology [33], host interactions [34,35] and others [36–39].

2. Challenges

Multiple challenges arise when integrating multi-omics datasets. Some are more general to machine learning analysis such as the presence of missing values or class imbalance and existing reviews already cover those subjects: Song *et al.* (2020) [40] and Mirza *et al.* (2019) [41].

Some are more specific and include the noisiness and complexity of omics datasets, which naturally occurs in biological data. Relevant patterns are sometimes subtles and involve many molecules from different omics layers. Finding those patterns across multiple datasets is therefore a difficult task. Moreover, when conducting omics or multi-omics experiments, the gathering of large amounts of biomedical data can often be done only on a small sample of patients for economical reasons, scarcity of the phenotype of interest, lack of volunteers, etc. This results in datasets with a number of variables greatly exceeding the number of samples. This issue is called *the curse of dimensionality* and machine learning algorithms tend to overfit these highly dimensional datasets, which decreases their generalizability on new data [42]. Another challenge is their heterogeneity which must be handled correctly as omics can have different data distribution or data types (e.g., numerical, categorical, continuous, discrete, etc.). Additionally, omics datasets can differ vastly in size (number of features), as a typical gene expression dataset will have tens of thousands of variables, while a metabolomics dataset can have a few thousands. Those discrepancies between omics can hinder their integration and produce an imbalance in the learning process. The different integration strategies presented in this mini-review address those problems differently by either reducing the number of variables, transforming the input data into a more exploitable representation, integrating at the end of the analysis, etc. More details about the strategies and tools available will be given in the next section of this review.

Class imbalance occurs when the distribution of classes in the learning data is biased, which can be a significant problem when working on rare events, such as an uncommon trait in a population. Several methods can be used to resolve this problem [14,15], such as sampling and cost-sensitive learning. Sampling tries to balance the dataset before the integration process, where either the majority class is randomly under sampled, or the minority class is oversampled by creating new artificial observations, or a combination of both methods. Cost-sensitive learning is directly integrated in the algorithm and balances the learning process by giving more weight to misclassified minority observations.

Missing data can take many forms ranging from variables with missing values to sample with missing omics data. If enough samples are available, removing the rows with missing data, namely listwise deletion, may be acceptable. If not, different statistical methods can be used to impute the missing values. A comprehensive review can be found in Song et al. (2020) [16]. Moreover, some machine learning methods can directly handle missing values like Random Forest 17 or K-Nearest Neighbor 18 or more recent methods [19,20].

3. Main integration strategies

From multiple omics datasets, each having the same rows representing samples (patients, cells) and different columns representing biological variables grouped by omics (gene expression, copy number variation, miRNA expression, etc.), different goals could be achieved such as sample classification, disease subtyping, biomarker discovery, etc. Machine learning (ML) models are commonly used to analyze complex data, but the integration of multiple noisy and highly dimensional datasets is not straightforward. Hence, multiple integration strategies have been developed, each one of them having pros and cons. Assuming each dataset has been pre-processed according to its omics data, the datasets could simply be assembled with sample wise concatenation and the resulting matrix used as input to ML models (Early integration, section 3.1). But in practice, most ML models will struggle to learn on such a complex dataset, particularly if the number of samples is low. Other strategies rely on transforming or mapping the datasets to reduce their complexity, either independently (Mixed integration, section 3.2) or jointly (Intermediate integration, section 3.3). An opposite strategy can also be adopted (Late integration, section 3.4), which does not combine data and analyzes each omics dataset separately. The prediction of each model is assembled afterward for a final decision. Finally, the hierarchical strategy (section 3.5) integrates the omics datasets by taking into account the known regulatory relationships between omics as presented by the central dogma of molecular biology [43]. In the next sections, we will first introduce dimensionality reduction methods (Section 3.0) as a powerful tool and secondary processing step and then present the different integration strategies in more detail (Section 3.1–3.5).

3.1. Dimensionality reduction for multi-omic integration

A sometimes necessary step in multi-omics analysis is dimensionality reduction, that is the process of reducing the number of variables in order to decrease the dimensionality and noise of a dataset. It is an optional simplification step and can be used regardless of the chosen integration strategy, but some (early and intermediate integration) often require prior dimensionality reduction to be more effective.

Two distinct approaches exist: feature selection which simply removes noisy and redundant variables and feature extraction, which combines the original variables into new and more meaningful variables. With an early integration, dimensionality reduction should be done on the concatenated matrix in order to take into account all the omics during the process. If dimensionality reduction is carried out separately on each dataset, a potential loss of information could ensue by not including every feature, the approach would then fall under one of the other integration strategies. In the next two sections we will quickly outline the most commonly used methods in both approaches, specific reviews on the subject can be found here [23,24].

3.1.1. Feature selection

Most omics datasets possess a high dimensionality which is in itself difficult to handle, but the problem is accentuated in multi-omics studies due to the number of datasets. One solution is to apply feature selection in order to simplify the integration process. Feature selection determines a smaller set of features which supposedly keeps most of the relevant information while reducing the dimensionality of the dataset. In addition to improving computing efficiency, removing features decreases complexity and noise which often results in higher performances and a reduced risk of overfitting for ML models. A low number of variables also makes the resulting models more interpretable. When a lot of variables are removed, feature selection can also deal on its own with the block scaling problem by evening out the number of features in each omics block [44].

Feature selection (FS) methods are organized into three classes, filter-based, wrapper-based and embedded methods. Filter-based methods are independent of any machine learning models and usually implement statistical analysis to find the most relevant variables while avoiding redundant features. They can be based on correlation (e.g., CFS [45], RCA [46]), distance (e.g., ReliefF [47]) or information gain [48] (e.g., mRMR [49]). Wrapper methods repeatedly apply a predictive ML model on different sets of features and those that improve the overall quality of the model are kept. Recursive feature elimination is the most common one, it starts by fitting a model with all the variables and gradually removes those which do not contribute to the model performances. These methods are focused on predictive power and can be used with any supervised ML models, but are limited by their computing efficiency if the dataset is large, which is often the case when studying omics. Finally, embedded methods are algorithms with feature selection built directly in the classifier. Among those embedded methods, the two most widely used are tree-based feature importance [50] and regularization. Regularization methods combine a loss function which evaluates the goodness of fit of the model, with a penalization function that punishes its complexity by favoring a smaller number of features. Compared to wrapper methods, they also resort to ML models, but are less computationally expensive. Due to the vast extent of regularization methods, we will not discuss it further and invite our reader to Wu *et al.* (2019) [23] and Vinga (2021) [51] for more information.

Feature selection can be applied to the separate omics datasets followed by concatenation. By definition it would no longer be an early integration as models would have been applied to each omics block independently. However, one might want to select variables while considering all omics together as it takes into account the redundancy of features across omics and might find more relevant features that single-omics studies will miss. The most straightforward way is to apply feature selection on the concatenated omics datasets. This strategy faces some of the same challenges described in the early integration section, that is balancing the influence of the different omics blocks, the increased complexity as well as additional computing time which would preclude the use of wrapper methods that are too computationally expensive.

3.1.2. Feature extraction

Feature extraction (FE) methods aim to transform the input features into another set of variables, that are linear or non-linear combinations of the original features. Their objective is to extract features in a way that the new variables keep the relevant information, while being less noisy and redundant. Learning from a smaller set of features also decreases complexity and improves computing efficiency. FE methods are often used in an exploratory manner to visualize data and expose important features, but they can also reduce the interpretability of a model as the extracted features are no longer biological measurements.

The most widely used FE method is Principal Component Analysis (PCA) [52]. PCA builds new variables called principal components, uncorrelated linear combinations of the original features

that maximize the description of variance in the dataset. It is however sensitive to outliers and cannot handle non-linear trends in the data. Several extensions have thus been developed to correct those problems including Kernel PCA [53] or Bayesian PCA [54]. Other similar methods include Principal Coordinates Analysis (PCoA) [55], Correspondence Analysis (CA) [56] and Independent Component Analysis (ICA) [57] may answer some of the shortcomings of PCA. Most FE methods are also being developed with sparsity constraints, often integrating regularization methods such as LASSO or elastic net in order to remove useless or redundant features. Sparse FE methods can be used for feature selection and include Sparse PCA (sPCA) [58], Sparse Canonical Correlation Analysis (CCA) [59], Sparse Non-Negative Matrix Factorization (Sparse NMF) [60], Sparse CA [61], etc. For example, Park *et al.* (2020) [62] used sPCA on each omics dataset and concatenated the retrieved PCs as a new dataset used as input to a Cox regression analysis.

FE methods can be used separately on each omics dataset to facilitate integration and for block scaling [24,63] in a mixed integration fashion, or applied on the concatenated multi-omics datasets (early integration). The extracted features are then useful as input to ML models or for clustering. However, those approaches often lead to unwanted redundancy and suboptimal results [64,65]. Intermediate methods solve this problem by jointly analysing the datasets, resulting in FE methods capable of taking into account all variables simultaneously, more in Section 3.3.

However, those methods will struggle to explore multi-omics datasets as applying them on the concatenated omics usually gives poor results. Thus, feature extraction methods are often used on each omics dataset separately for either block scaling [24,63], or after concatenation of the extracted features for clustering or other downstream analysis.

3.2. Early integration

The early integration is based on the concatenation of every datasets into a single large matrix. This process increases the number of variables, but the number of observations stays the same. Consequently, several integration challenges are exacerbated by this process resulting in a more complex, noisy and high dimensional matrix, which makes learning difficult. Additionally, the size difference between omics datasets can promote a learning imbalance as the algorithm spent more time learning on the omics with the biggest number of variables, overlooking the other omics [66,67]. Early integration also ignores the specific data distribution of each omics, which can potentially misguide ML models into finding irrelevant patterns that simply reflect the features' membership to the same omics. A conclusion found for example in a comparison study done by Spicker *et al.* (2008) [66]. Nevertheless, early integration is still commonly used as it has some clear advantages including its simplicity, easy implementation and mostly, combining variables from each omics allows ML models to directly uncover interactions between the different layers. It is also not known to what extent the aforementioned drawbacks influence the downstream analysis and it is possible that the performances of some ML models are not significantly lessened.

Approaches using the early integration strategy need at least to address the complexity of the composite matrix, often by reducing its number of variables through feature selection or dimensionality reduction methods. Then, most ML models can be used for analysis, but in recent years, Deep Learning (DL) [68] has been commonly used as it is flexible and powerful enough to accurately detect relevant patterns even from the concatenated data. For instance, Xie *et al. (2019)* [69] fed both multi-omics and clinical data to the input layer of an artificial neural network, itself linked to a Cox Proportional hazard model (Cox-PH) in order to predict survival of patients with cancer. For a similar goal, Chaudhary *et al (2018)* [70] implemented instead of the common fully connected neural network an autoencoder to reduce the dimensionality of the multi-omics matrix and extract meaningful and compact DL-based variables on which clustering was done with the k-mean algorithm.

Although they are highly adaptive and often achieve superior performances with big datasets, one of the most challenging issues with neural networks is their black box nature, that is their lack of interpretability. Particularly in biomedical studies, having a good predictive model is not enough and an understanding on how genes and other molecules are implicated in the underlying biological process is necessary. The transparency of machine learning results may also lead to new biological discoveries. As several methods have been developed for interpreting neural networks models, we invite our reader to other reviews [71,72]. Most of those strategies focus on explaining the final decision of the algorithm and identifying biomarkers, but some DL models [73,74] can directly find relevant biological pathways during the learning process (Fig. 1). The basic idea behind this approach is to utilize known biological pathways to define the architecture of the neural network. The input layer representing biological entities (molecules, genes, proteins, etc.) is connected to a second layer where each node is a known molecular pathway. Connections between nodes are made only if the molecule is known to take part in the pathway, resulting in a sparse interaction between the first two layers. The pathway layer is then fully connected to hidden layers. When training the model, all the connections are updated and the final prediction of the network is directly interpretable by looking at which nodes are activated. The drawback of such methods is that they cannot discover new interactions or make use of little studied proteins or genes if their implication in a pathway has not yet been discovered.

Additionally, the early strategy allows the inference of heterogeneous networks using methods such as Mixed Graphical Models (MGM) [75,76], which expand from Gaussian Graphical Models that assume normal distribution of variables to a mixed model. MGM regresses each variable against every other using either linear regression or logistic regression depending on the type of variable (continuous or discrete/categorical). Another method, based on decision trees, is Graphical Random Forest [44,77], which computes a Random Forest on each variable using every other feature as predictors. Features that are ranked as important by the importance measure of Random Forest are considered to interact with the selected variable. MGM as well as Graphical Random Forest can integrate prior knowledge [78,79]. Additionally, Zhong *et al.* (2019) [80] developed mixed Directed Acyclic Graph (mDAG) which can infer causal interactions based on variables with different distributions and can potentially be used in multi-omics studies. More information on inferring heterogeneous networks from multi-omics data can be found in the reviews [76,81].

3.3. Mixed integration

The mixed integration strategy addresses the shortcomings of the early integration by transforming independently each omics dataset into a simpler representation. The new representation can be less dimensional and less noisy which facilitates analysis. Moreover, most heterogeneities between omics datasets such as the data's type or size differences are removed in the new representation. The combined representation can then be analysed by classical ML models. We will present three transformation methods, kernel-based methods, graph-based methods and Deep Learning (DL).

3.3.1. Kernel learning

Kernel models are powerful ML models able to implicitly operate in a high dimensional space in which linear relationships

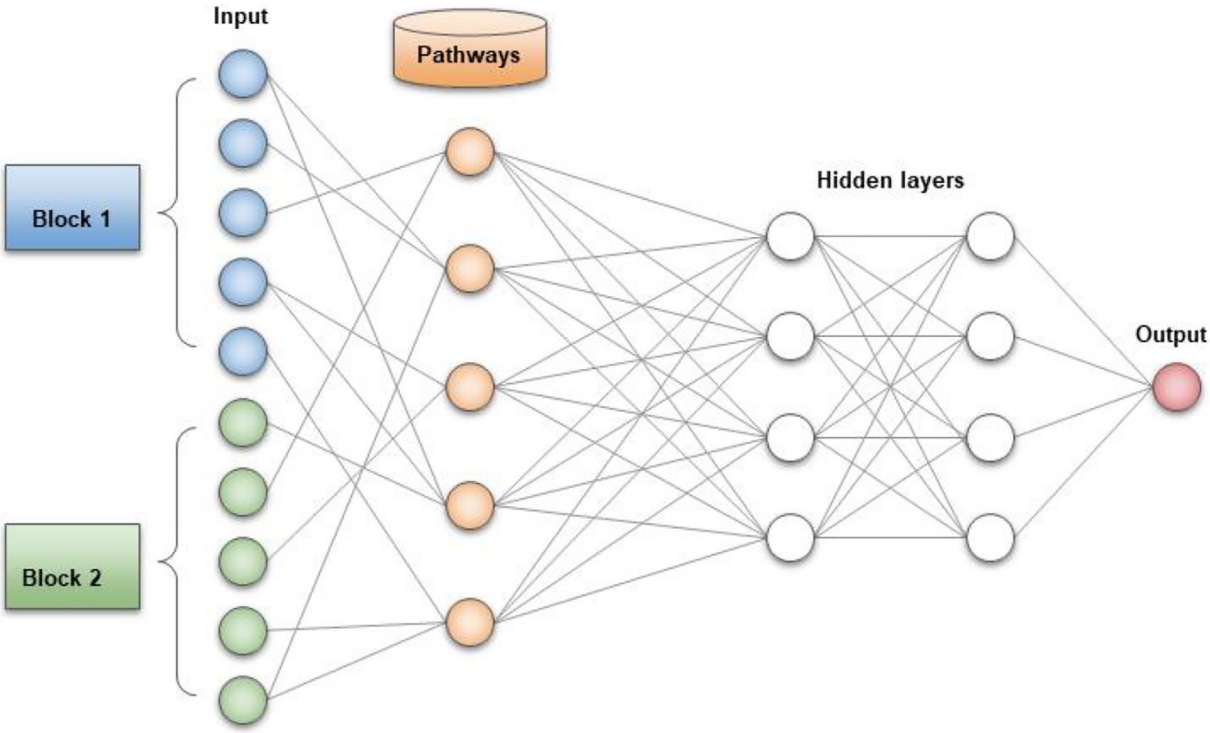

Fig. 1. Structure of an interpretable artificial neural network. The input layer is followed by an additional pathway layer, where each node corresponds to a known molecular pathway. If a molecule is known to be involved in a pathway, a connection is made between the two. Hence, important pathways implicated in the outcome are activated with bigger weights during training. Figure inspired from Deng *et al. (2020)* [73].

between observations can be found. Consequently, kernels can naturally be used to infer pairwise similarities of samples [82], taking the form of a similarity matrix that can be used for downstream analysis using ML models such as Support Vector Machines (SVM) [83], Partial Least Squares (fKPLS) [84] or a Cox model (kernel-fusion Cox) [85]. Different types of kernels exist such as linear, gaussian, polynomial, sigmoid, etc., thus providing different similarity measures for the same data. Because one of them can be more suitable than the other depending on the type of omics and its data distribution, the right kernel is often found with cross-validation. Multiple Kernel learning (MKL) [86] can integrate different omics blocks by first computing a kernel for each dataset (the kernel can be of different types), and combining them to produce a global similarity matrix which describes samples across all multi-omics datasets. The best way to combine the different kernels is found experimentally by learning their appropriate weights. Two different approaches exist [87], either using a wrapper strategy such as SimpleMKL [88] or using an optimization algorithm such as SpicyMKL [89] or SMO-MKL [90].

Recently, Zhang *et al.* (2016) [91] and He *et al.* (2021) [92] both used SimpleMKL on five different omics datasets, with a preliminary feature selection using mRMR [49] to increase performances. However, Zhang *et al.* applied for each dataset the same kernel type (Gaussian), whereas He *et al.* used two different types (Gaussian and Polynomial). On the other hand, SIMLR [93] and its multi-omics extension CIMLR [94] compute several gaussian kernels with varying hyperparameters for each omics and assume that the global kernel matrix better reflects similarity between samples and naturally reflects possible clusters. Wang *et al. (2017)* [93] shows that the constructed similarity measure outperforms standard similarity measures.

MKL can also be used in an unsupervised and exploratory manner [95,96]. The resulting similarity space produced by the multi-omics kernel can be used as input by well-known algorithms such as PCA (kPCA) [53] or k-means (kernel Power k-means) [97]. Since the samples are no longer described by their input features, but by the new feature space describing their similarities, interpretation of the unsupervised model can be more challenging. For example, Speicher and Pfeifer (2015) [96] used unsupervised MKL on gene expression, methylation and copy number data for cancer subtype discovery, but could only assess for each cancer type the relative importance of each omics. Ideally, one would want to retrieve the importance of specific variables to better explain the results. Thus, a method based on random permutation and kPCA was introduced by Mariette and Villa-Vialaneix (2017) [95] to address this problem. The importance of a variable is measured by randomly permuting its value between samples and its influence can be estimated in the PCs space obtained with kPCA, in the same way as regular PCA.

3.3.2. Graph-based

A mixed strategy based on graphs consists of modelling each omics into a separate graph before analysis. Three approaches can consequently be used. The first one is to combine them into a single homogeneous network through fusion. The second approach is to build a multi-layer (or multiplex) network with inter-layer connections. Once a unique network is obtained, utilizing the principle of guilt-by-association, nodes close to each other are assumed to share related biological functions. Therefore, finding pertinent modules is often done in order to classify nodes with unknown functions or reveal activated pathways. In contrast to modules, random walks can connect distant nodes within the network to reveal potential long-distance interactions. Random walks simulate an imaginary particle starting at a seed node and moving randomly to other nodes in order to explore the network's topology. After several iterations, a stationary probability distribution is obtained, which depicts the topological properties of the seed node and can be used to reveal its importance or its similarity to

other nodes. Finally, the third approach does not rely on integrating several networks, but on learning graph-based variables from each of them, which can be used as input to other ML models.

The first approach of building homogeneous networks often relies on creating patient similarity networks for each omics, in which patients are nodes and weighted edges describe their similarities. Then, all networks are combined using a fusion method such as Similarity Network Fusion (SNF) [98] or its variation Affinity Fusion (ANF) [99] which is implemented with block normalization. Recently, Wen *et al.* (2021) [100] introduced Random Walk with Restart for multi-dimensional data Fusion (RWRF) which authors say is a more effective fusion method. One advantage of such methods is that the network doesn't get more complex with additional omics as their overall size is based on the number of samples, not the number of features. The resulting integrated graph can then be used as input for ML models [99,101,102] for clustering, subtype discovery or survival prediction.

The second approach relies on building multi-layer networks, where each layer represents an omics and interactions between omics are either inferred or retrieved from interaction or pathways databases. We invite our readers to the review of Lee *et al.* (2019) [29] for more information on multi-layer networks inference. Studying the overall topology of the network can reveal important molecules and perturbed pathways leading to specific phenotypes. Several methods can explore the network's topology including shortest paths, random walks and other variations for multi-layer networks. In order to do gene reprioritization, Valdeolivas *et al. (2019)* [103] for example recently developed two new algorithms based on random walks for multiplex networks that can explore the different layers of physical and functional interactions between omics. For a similar goal, Shang and Liu (2020) developed iRANK [104], a variation of the PageRank algorithm [105] that they utilized on a multi-layer network composed of epigenomics data, gene expression data and protein–protein interactions network. A different approach was used by Murodzhon *et al.* (2017) [15] who developed OmicsNet, a weighted multi-layer network made of omics layers as well as a biological concept layer and phenotype layers. Specific nodes can then be associated with phenotypes by calculating along the weighted path an integrated score which can model a plausible signaling cascade. The results can be used for biomarker discovery or predictive analysis. Moreover, Liu *et al.* (2013) [106] proposed an interesting method to inspect a multi-layer network, recently improved by others [107,108], where the activity of known molecular pathways present in the network is assessed by random walks. Active pathways are predicted by their concentration of important nodes (differentially expressed, associated with disease, topological importance, etc.). New features are then constructed based on the activity values of pathways and used with ML models for subtype classification or survival prediction tasks.

The third approach utilizes graph embedding methods which can learn low dimensional representation of the nodes and their surroundings from each network. The new graph-based features are then combined and fed to other ML models for prediction, classification, etc. Existing reviews on graph embedding methods [109,110] go into more details, we will only present some recent cases. For example, DTINet [111] used random walk-based embedding on multiple interaction and similarity networks of proteins, drugs and diseases in order to detect similar nodes and predict new drug-target interactions (DTI). DTInet uses a dimensionality reduction method (DCA [112]) to reduce the multiple embeddings to a unique embedding for drug features and target features. Xuan *et al* (2019) [113] used a similar approach, but showed that the dimensionality reduction might remove too much information and instead used an ensemble learning approach with Gradient Boosting Decision Tree, which also deals with class imbalance as the number of unknown DTIs is higher than the number of known DTIs. Additionally, the graph-based variables can be compacted even further using autoencoders [114,115] (Deep Learning , section 3.2.5). DeepDR [115] for example combined the different features obtained for each graph with a multi-modal autoencoder. The bottleneck layer containing the integrated information was later fed to a collective Variational Autoencoder [116] (cVAE) for DTIs prediction.

Multiple Kernel Learning (MKL) (Section 3.2.4) can also be used as graph embedding methods. In order to establish which molecular pathways are involved in breast cancer, Manica *et al.* (2019) [117] developed PIMKL which combines MKL with prior knowledge in the form of interaction networks and annotated genes or pathways. Kernel functions were designed to encode the topological information of known pathways, which can also be combined with the multi-omics datasets. The overall process allows the mapping of samples from the omics space to an interaction space (edge space), which explicitly reveals the underlying biological mechanisms. Additionally, Tepeli *et al.* (2021) [118] developed PAMOGK which introduced a graph kernel to determine sample similarity from graph data and predict subgroups of patients.

Additionally, Graph Neural Networks (GNN) [119] and its upgrade Graph Convolutional Neural Networks (GCN), are specifically designed to receive graph data as input. GCNs have also been developed for DTIs prediction [120] or node classification [121]. Going further than simple link prediction, Zitnik *et al.* (2018) [122] have used a GCN called Decagon to predict the presence of side effects between two drugs as well as the side effects' type. GCN can also embed entire networks, which is mostly applied on drugs (which can be considered as graphs of atoms) and the resulting molecular embedding can be combined with multi-omics datasets to increase prediction performances [123]. More information on deep learning for biological networks can be found in Muzio *et al.* (2021) [124]. General integration of multi-omics data with deep learning is the subject of the next section.

3.3.3. Artificial neural networks

Artificial neural networks (NN) are powerful ML models made of many neurons organized in layers. They can be used directly on the concatenated omics (Section 3.1 Early integration) or separately on each omics (Section 3.4 Late integration), but it can also be utilized to learn meaningful latent representations (deep learning-based features) from each datasets by processing them in separate layers. The latent representations can be seen as new deep variables learned by the different layers of the model, which can be easily concatenated or connected to other neural networks for more analysis. Thus, the hidden layers of a NN can be considered as successive feature extraction layers, while only the final output layer can produce a prediction. Some NN architectures are specialized in learning a pertinent latent representation, such as Autoencoders (AE) and Restricted Boltzmann Machines (RBM). Both models are unsupervised neural networks that reproduce the original data from a compressed representation encoded within the central or bottleneck layer of the network. The fewer neurons in this layer, the more compact the representation is. The new representation is then useful as input for other ML models and particularly for clustering [125].

Among the deep learning models developed for multi-omics integration, we can present MOLI [13,127], which retrieved DL-based features using subnetworks on each omics dataset and concatenate the obtained deep features. Then, a final neural network is used on the concatenated deep features for prediction of drug activity. Using a similar approach, Islam *et al.* (2020) [128] predicted breast cancer subtypes using the concatenated features, but they learned them through convolution neural networks applied on gene expression and copy number variation datasets.

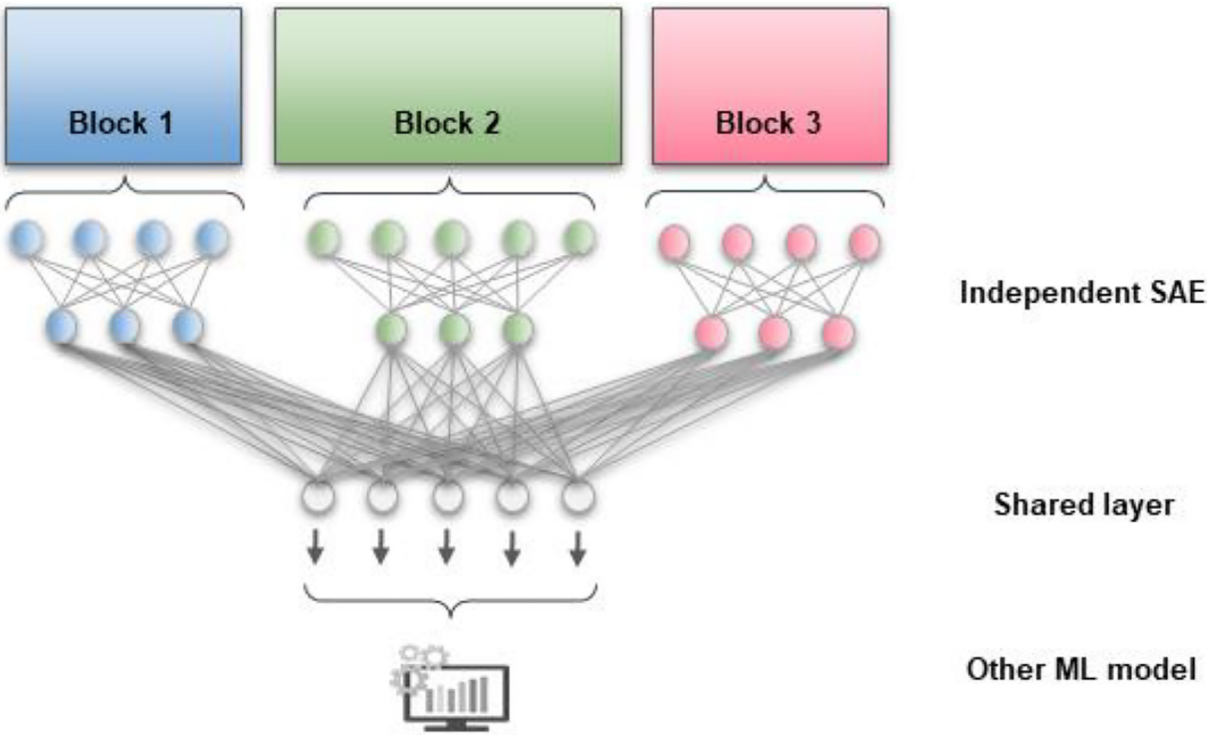

Fig. 2. Example of a mixed artificial neural network. Each omics block is first reduced to a latent representation using independent Stacked Sparse Autoencoders (SAE). The new representations learned are integrated in a final shared layer. The common representation is used for downstream analysis such as prediction or clustering. Figure inspired from Xu et al. (2019) [126].

Instead of concatenating the deep features obtained for each omics block, another approach is to simply connect them to a shared layer. For example Xu *et al.* (2019) [126] (Fig. 2) first used stacked Sparse Autoencoders (SAE) on each omics dataset and integrated the results into a final layer. The resulting shared representation was fed to a deep flexible neural forest for predicting cancer subtypes. Similarly, Yang *et al.* (2021) [129] developed a multimodal autoencoder capable of taking all omics datasets as input, compacting them into the central layer, or bottle-neck layer of the autoencoder and reconstructing them at the end. The resulting deep features generated in that common layer take into account all omics and were also used for the discovery of cancer subtypes.

3.4. Intermediate strategies

We describe as intermediate integration any methods capable of jointly integrating the multi-omics datasets without needing prior transformation and without relying on a simple concatenation. They generally output new constructed representations, one common to all omics and some omics-specific, on which further analysis can be done. This step reduces the dimensionality and complexity of the multi-omics datasets. However, they are most often used after feature selection and robust pre-processing as the heterogeneity between datasets can prevent them from working correctly. Only a few methods were developed with the ability to find semi-shared structures, that is patterns shared between some omics but not all. Such methods include SLIDE [130] for example. Ideally, to limit the loss of information that occurs by selecting features independently for each omics, an intermediate feature selection method could be used such as the extension of mRMR developed by EL-Manzalawy *et al.* (2018) [65]. It selects features by taking into account their complementarity within and across omics blocks. The same results cannot be achieved by simply applying mRMR on each dataset, nor by applying mRMR on the concatenated dataset. Some intermediate strategies are also designed as multi-block feature extractions methods and can be utilized for exploratory purposes or as basis for downstream analysis, in the same way as a regular feature extraction (Section 3.0).

Intermediate methods are often formulated with the assumption that the different datasets share a common latent space, which can reveal the underlying biological mechanisms. Among those methods, extensions of the widely used Non-negative Matrix Factorization (NMF) [131] have been developed including joint NMF [132] and integrative NMF [133,134]. Both methods infer a common matrix depicting the latent relationships between every omics dataset, but while joint NMF uses the common space to identify modules of correlated multi-omics data, integrative NMF implements sample clustering and subtype discovery. Other similar methods are presented in Table 1. The main advantage of such intermediate methods is their ability to discover the joint inter-omics structure, while also highlighting the complementary information contained in each omics. We won't go into further details as the number of multi-block dimensionality reduction methods is substantial and still increasing, for more information, specific reviews have been written on the subject [14,26,135].

Other methods originally developed for two datasets were extended to multi-omics including Canonical Correlation Analysis (CCA) [136,137] or Co-Inertia Analysis (CIA) [138]. The difference between them and the other presented in Table 1 is that they do not construct a common space, but infer omics specific factors while maximizing some joint measure such as correlation of co-inertia. We won't go into further details as the number of multi-block dimensionality reduction methods is substantial and still increasing, for more information, specific reviews have been written on the subject [14,26,135].

Table 1
A non-exhaustive list of multi-block dimensionality reduction methods for multi-omics datasets. NMF: Non-negative Matrix Factorization, MOFA: Multi-Omics Factor Analysis, JIVE: Joint and Individual Variation Explained, MO: multi-omic.

Method	Principle	Purpose	Recent applications
jNMF/intNMF/nNMF [132,133,139]	Matrix factorization	Disease subtyping, module detection, biomarker discovery	jNMF found biomarkers in MO and pharmacological data connected to drug sensitivity in cancerous cell lines [140]. intNMF identified Glioblastoma and breast cancer subtypes from MO and clinical data [134].
MOFA/MOFA+ [141,142]	Bayesian Factor Analysis	biomarker discovery, systemic knowledge	MOFA found new biomarkers and pathways associated with Alzeihmer's disease based on MO data including proteomics, metabolomics, lipidomics [143].MOFA + found predictive biomarkers from DNA methylation and gene expression data in cardiovascular disease [144].
iCluster [145]	Gaussian latent variable model Generalized linear regression Bayesian integrative clustering	Disease subtyping, biomarker discovery	iCluster was used to identify subtypes of esophageal carcinoma from genomic, epigenomic and transcriptomic data [148].
iClusterPlus [146]			iClusterPlus was used to identify subtypes of non-responsive samples with ovarian cancer from different omics datasets [149].
iClusterBayes [147]			iClusterBayes was used to identify predictive biomarkers and clinically relevant subtypes on MIB cancer from 5 different omics [150].
JIVE/aJIVE [151,152]	Matrix factorization	Disease subtyping, systemic knowledge, module detection	JIVE was used as a dimension reduction technique to improve survival prediction of patients with glioblastoma from mRNA, miRNA and methylation data [153].
Integrated PCA [64]	Generalized PCA	Visualization, prediction	iPCA was used as a dimension reduction technique to improve prediction of outcome on lung cancer from CpG methylation data, mRNA and miRNA expression [154].
SLIDE [130]	Matrix factorization	Disease subtyping, module detection, biomarker discovery	SLIDE was used on DNA methylation data and gene, protein and miRNA expression for subtyping patients with breast cancer [130].

3.5. Late integration

For handling multi-omics datasets, the most straightforward integration strategy is to apply machine learning models separately on each dataset and then combine their respective predictions, namely Late integration. Its strength relies on its capacity to use readily available tools designed specifically for each omics type, and compared to the other strategies, it does not suffer the challenges of trying to assemble different kinds of data. For example, Sun *et al.*(2019) [155] built neural networks for each dataset consisting of gene expression, copy number data and clinical information and linearly aggregated their predictions into a single final prediction for cancer prognosis. A more complex aggregation function was used by Wang *et al.* (2020) [156] where the authors trained Graph Convolutional Neural Networks on each omics (and their respective patient similarity networks) to recover initial classification predictions. The single-omics predictions were then utilized to construct a cross-omics tensor, which was forwarded to a View Correlation Discovery Network (VCDN) that makes a final class prediction based on the individual omics predictions and the latent cross-correlation between omics.

The shortcoming of such integration strategy is that it cannot capture inter-omics interactions and at no point in the learning process can the different machine learning models share knowledge and utilize the complementarity information between omics. Combining predictions is simply not enough to accurately exploit multi-omics data and understand the underlying biological mechanisms of diseases. For that reason, and because it boils down to multiple single-omics analysis, we will not discuss it further.

3.6. Hierarchical integration

A challenge in system biology is to understand the modular organization structured at the molecular level. A new trend is to incorporate these regulatory effects in the integration strategy to better reflect the nature of multidimensional data. Hierarchical strategy bases the multi-omics integration on the inclusion of the prior knowledge of regulatory relationships between the different layers. For example, a strategy for genotype-phenotype integration based on existing knowledge of cellular subsystems could follow this logic: genotypic variations in nucleotides can give rise to change in gene expression or functional changes in proteins which in turn could ultimately affect the phenotype. Therefore, hierarchical integration strategies often use external information from interaction databases and scientific literature. Moreover, because omics are organized in sequential fashion, the challenges of multi-omics integration are not exacerbated and can be dealt with separately for each dataset.

Some methods for supervised hierarchical integration include Bayesian analysis of genomics data (iBAG) [157], linear regulatory modules (LRMs) [158] and Assisted Robust Marker Identification (ARMI) [159] and Robust Network [160]. Hierarchical integration methods are often designed to study specific regulatory relationships. For example, iBAG has been developed to investigate associations between epigenetic and gene expression regulation. The framework uses hierarchical modeling to combine the data from methylation and gene expression to study the associations with patient survival. Robust Network has developed an approach for modeling the gene expression (GE) and copy number variation (CNV) regulation that describe the dominant *cis*-acting CNV effects compare to *trans*-acting CNVs. This approach could be extended to other regulation relationships such as gene expression by methylation and microRNAs. Additionally, hierarchical integration can be used to infer gene regulatory networks (GRN) from multi-omics datasets. For instance, Zarayeneh *et al.* (2016) [11] developed a GRN inference method by taking into account interaction effects between gene expression, Copy Number Variations (CNV) and DNA methylation. Their method utilized the epigenomic data to better predict regulatory interactions and achieved significantly better results on simulated data, compared to two other GRN inference methods SGRN [161] and DCGRN [162]. For more information on inference methods for GRN, we recommend Wani and Raza (2019) [28].

Finally, Fortelny and Bock (2020) [163] developed a neural network model in which each node corresponds to a biological entity such as a protein or a gene and each edge with a known interaction. The layout of the network follows the flow of information in the cell, with the input layer being gene expression and the following layers being transcription factors, signalling proteins, receptors, etc. The advantage of this DL model is that it is directly interpretable by looking at the activated nodes.

4. Summary and outlook

In this mini-review, we presented the different strategies available to handle multi-omics datasets integration. Most integration approaches developed in recent years tend to first modify and transform each dataset using different machine learning models known as Mixed integration, in order to reduce their complexities and heterogeneities and facilitate their subsequent integration and analysis. While it can give informative results, each dataset is transformed independently, potentially resulting in a loss of information and a final model that can still suffer from noise or redundant information. Ideally, at any point of the learning process, each omics dataset should be assessed while considering the other datasets, so that the complementary information could be best exploited. The early and intermediate integration strategies do solve this problem by integrating all datasets beforehand, but the large matrix resulting from an early integration is difficult to exploit by most ML models and intermediate integration often relies on unsupervised matrix factorization, which has difficulty incorporating the considerable amount of pre-existing biological knowledge. Another methodology, hierarchical integration, is explicitly designed with the prior understanding of how the different omics layers interact with each other. However, only few such methods have been developed and are often tailored for specific omics types, which makes them less generalizable than other approaches. Additionally, they are dependent on prior data, which prevents them from exploring and discovering new biological mechanisms and pathways.

Another issue to tackle is whether or not "*More is better*", quoting from Huang *et al.* (2017)[19]. Adding omics datasets for the only sake of adding more data might not always be a good idea. They carry more information and can potentially reveal pathways from different biological layers, but the additional data could also bring more noise, redundancy and an increased computational time than relevant information. Additionally, while multi-omics integration often leads to better results [126,154,164], some have shown that it is not always the case [26,165]. Worse performances could arise if the model is not suited for a particular goal or for particular multi-omics datasets. Some models cannot handle massive matrices, outliers, highly correlated variables, noise, etc., issues that are exacerbated in multi-omics studies. It is also possible that the omics are not correctly integrated [165]. We believe that if the machine learning model and the integration strategy are chosen judiciously, multi-omics should always surpass single-omics performances, but knowing in advance the proper integration strategy is not always feasible. There are not yet any general rules of thumb to foresee which method will achieve better results and most benchmarks generally conclude that the best approaches have to be chosen depending on the initial data. Nevertheless, based on recent trends, we can begin to notice effective approaches. For example, the early integration is being more and more outperformed when compared to other integration strategies [13,65,66,164]. We believe that an early integration cannot handle too many datasets, especially when their heterogeneity is great and we suggest instead the mixed strategy which is tailored to deal with such challenges. The complementarity of datasets and their relative pertinence should also be taken into account, as some omics will contain less or possibly no useful information [153,166]. Depending on the studied pathology, some omics will be more appropriate than others. In the same way, specific combinations of omics (metabolomics + proteomics, genetics + epigenet ics, etc.) should be more fruitful than others and an understanding of those interactions is necessary. Based on the literature and biological knowledge collected on a specific topic, one could infer which omics layer should be retained or disregarded. Otherwise, the influence of each omics block should be determined during the analysis. For example, the mixed integration generally assumes equal importances and reshape the omics into similar representations. Thus in order to adjust their influences, their appropriate weights can usually be learned during training in the case of supervised learning. For unsupervised learning, it remains a challenge. This issue can be elegantly tackled by intermediate methods as they produce multi-omics and omics specific outputs which reveal the complementarity and benefits of each omics.

Furthermore, progress is continuously being made and new tools for multi-omics integration are continually being proposed. Network-based integrations are very promising, particularly for their ability to use pre-existing interaction networks and known molecular pathways as well as their straightforward interpretation. The exploration of multi-layer heterogeneous networks is just beginning and will surely continue to grow and gain in predictive and explanatory power as most of the tools currently used were designed for single-omics layers. The advances of deep learning are also quite compelling. Their flexible architectures facilitate the integration of multiple omics datasets, which can also be combined with biomedical images or other types of data, offering a better grasp of a patient's pathology. However, due to its large number of parameters, DL models are hard to train, must be tuned precisely and often experience overfitting. Their performances rely heavily on the availability of samples, which is still limited. An interesting way to deal with this issue is transfer learning [167], machine learning models are pretrained on larger and general datasets on which they can learn basic patterns and are then fine-tuned on the more specific dataset of interest. Transfer learning is widely used in image recognition, but is not yet regularly used in multi-omics studies. Deep learning also suffers from the reputation of not being easily interpretable, which is a major obstacle in biomedical studies. We have presented in this mini-review some examples of interpretable DL models, but more research must be done in order to confirm their capabilities and whether or not they can be adapted for different purposes.

With the ever-growing access to biological data, multi-omics research will be performed more and more often, and it is urgent that we identify the best practices, tools and strategies for their integration. In that aspect, benchmark studies are also particularly useful and should be done more frequently. With the notable exception of Herrmann *et al.* (2020) [168] which focused on survival prediction methods for multi-omics data, most benchmarks focus on clustering and dimensionality reduction methods [14,26,27,135,169–171]. Thorough comparisons of other ML models have not been made for multi-omics datasets, and we have yet to know if the deep learning prowess made in other fields of pattern recognition can be reproduced in bioinformatics [172].

4.1. Search strategy

This mini-review presents methods and strategies for multi-omics integration. The goal is not to produce an extensive list of articles and tools currently being used in the bioinformatic community, as the subject is too wide and not fit for a mini-review. The goal is more to display general trends and interesting ideas

about the subject, with a particular focus on new and original methods developed in the past three years.

Articles and reviews were searched on PubMed with the keywords multi-omics / multiomics / machine learning multi-omics / deep learning multi-omics / network multi-omics / multi-omics integration / multi-block omics. Reviews published after 2015 were prioritized as well as methods and tools published after 2017, unless an older method was recently improved upon in a recent publication. This preliminary search produced approximately 1,000 results. Then, publications were kept only if the tool presented could manage at least two omics (exception was made for the hierarchical integration which is more specific). Additionally, we focused mostly on popular publications developing new approaches rather than on publications using existing tools to answer biological problems. In order to restrain the length of the mini-review and promote an easier reading experience, an effort was made not to include too similar articles. Altogether, these criteria resulted in the current publications presented in the mini-review.

5. Author statement

MP, AB and MPSB wrote the manuscript. AB and MPSB designed the figures. MP, MPSB, AB, OP revised the manuscript. AD supervised research.

Funding

This work was supported by Research and Innovation chair L'Oréal in Digital Biology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated omics: tools, advances, and future approaches. J Mol Endocrinol 2018. https://doi.org/10.1530/JME-18-0055.
[2] Ahmed Z. Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. Hum. Genomics 2020;14.
[3] Burney IA, Lakhtakia R. Precision Medicine: Where have we reached and where are we headed?. Sultan Qaboos Univ. Med. J. 2017;17.
[4] Jaccard E, Cornuz J, Waeber G, Guessous I. Evidence-based precision medicine is needed to move toward general internal precision medicine. J Gen Intern Med 2018;33.
[5] Tebani A, Afonso C, Marret S, Bekri S. Omics-based strategies in precision medicine: toward a paradigm shift in inborn errors of metabolism investigations. Int J Mol Sci 2016;17.
[6] Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput Struct Biotechnol J 2021;19:949–60.
[7] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18:83.
[8] Sun YV, Hu Y-J. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. Adv Genet 2016;93:147–90.
[9] Dahal S, Yurkovich JT, Xu H, Palsson BO, Yang L. Synthesizing systems biology knowledge from omics using genome-scale models. Proteomics 2020;20:e1900282.
[10] Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief. Bioinform. 2018;19:1370–81.
[11] Zarayeneh N et al. Integration of multi-omics data for integrative gene regulatory network inference. Int. J. Data Mining Bioinformatics 2017;18:223.
[12] Rappoport N, Safra R, Shamir R. MONET: Multi-omic module discovery by omic selection. PLoS Comput Biol 2020;16:e1008182.
[13] Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 2019;35:i501–9.
[14] Tini G, Marchetti L, Priami C, Scott-Boyer M-P. Multi-omics integration—a comparison of unsupervised clustering methodologies. Briefings Bioinf. 2019;20:1269–79.
[15] Murodzhon A, Alberto A, Montemanni R, Francesco B, Ivo K. OmicsNet: Integration of Multi-Omics Data using Path Analysis in Multilayer Networks. (2017).
[16] Higdon R et al. The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. OMICS 2015;19:197–208.
[17] Zitnik M et al. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Inf. Fusion 2019;50:71–91.
[18] Bersanelli M et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinf 2016;17(Suppl 2):15.
[19] Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet 2017;8:84.
[20] Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief. Bioinform. 2018;19:325–40.
[21] Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. Front Oncol 2020;10:1030.
[22] Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet 2015;16:85–97.
[23] Wu C et al. A selective review of multi-level omics data integration using variable selection. High Throughput 2019;8.
[24] Meng C et al. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform. 2016;17:628–41.
[25] Franco EF et al. Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. Cancers 2021;13.
[26] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucl Acids Res 2018;46:10546–62.
[27] Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief. Bioinform. 2020;21:541–52.
[28] Wani N, Raza K. Integrative Approaches to Reconstruct Regulatory Networks From Multi-Omics Data: A Review of State-of-the-Art Methods. doi:10.20944/preprints201804.0352.v1.
[29] Lee B, Zhang S, Poleksic A, Xie L. Heterogeneous multi-layered network model for omics data integration and analysis. Front Genet 2019;10:1381.
[30] Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. Brief. Bioinform. 2016;17:891–901.
[31] Eicher T et al. Metabolomics and multi-omics integration: a survey of computational methods and resources. Metabolites 2020;10.
[32] Mantini G, Pham TV, Piersma SR, Jimenez CR. Computational analysis of phosphoproteomics data in multi-omics cancer studies. Proteomics 2021;21:e1900312.
[33] Canzler S et al. Prospects and challenges of multi-omics data integration in toxicology. Arch Toxicol 2020;94:371–88.
[34] Culibrk L, Croft CA, Tebbutt SJ. Systems biology approaches for host-fungal interactions: an expanding multi-omics frontier. OMICS 2016;20:127–38.
[35] Khan MM et al. Multi-omics strategies uncover host-pathogen interactions. ACS Infect Dis 2019;5:493–505.
[36] Jamil IN et al. Systematic multi-omics integration (MOI) approach in plant systems biology. Front Plant Sci 2020;11:944.
[37] Leon-Mimila P, Wang J, Huertas-Vazquez A. Relevance of multi-omics studies in cardiovascular diseases. Front Cardiovasc Med 2019;6:91.
[38] Labory J et al. Multi-omics approaches to improve mitochondrial disease diagnosis: challenges, advances, and perspectives. Front Mol Biosci 2020;7:590842.
[39] Morello G, Salomone S, D'Agata V, Conforti FL, Cavallaro S. From multi-omics approaches to precision medicine in amyotrophic lateral sclerosis. Front Neurosci 2020;14:577755.
[40] Song M et al. A review of integrative imputation for multi-omics datasets. Front Genet 2020;11:570255.
[41] Mirza B et al. Machine learning and integrative analysis of biomedical big data. Genes 2019;10.
[42] Domingos P. A few useful things to know about machine learning. Commun ACM 2012;55:78–87.
[43] Crick F. Central dogma of molecular biology. Nature 1970;227:561–3.
[44] Zierer J et al. Exploring the molecular basis of age-related disease comorbidities using a multi-omics graphical model. Sci Rep 2016;6:37646.
[45] Hall MA. Correlation-based feature selection for machine learning. http://www.cs.waikato.ac.nz/~ml/publications/1999/99MH-Thesis.pdf.
[46] Wosiak A, Zakrzewska D. Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis. Complexity 2018;2018.
[47] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. in Machine Learning: ECML-94 171–182 (Springer Berlin Heidelberg, 1994).
[48] Raileanu LE, Stoffel K. Theoretical comparison between the Gini index and information gain criteria. Ann. Math. Artif. Intell. 2004;41:77–93.
[49] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226–38.
[50] Scornet E. Trees, forests, and impurity-based variable importance. arXiv [math.ST] (2020).

[51] Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. Brief. Bioinform. 2021;22:77–87.
[52] Ringnér M. What is principal component analysis?. Nat Biotechnol 2008;26:303–4.
[53] Schölkopf B, Smola A, Müller K-R. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 1998;10:1299–319.
[54] Nounou MN, Bakshi BR, Goel PK, Shen X. Bayesian principal component analysis. J Chemom 2002;16:576–95.
[55] Xie Y-L et al. Robust principal component analysis by projection pursuit. J Chemom 1993;7:527–41.
[56] Beh EJ. Simple correspondence analysis: a bibliographic review. Int. Stat. Rev. 2007;72:257–84.
[57] Sompairac N et al. Independent component analysis for unraveling the complexity of cancer omics datasets. Int J Mol Sci 2019;20.
[58] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. J. Comput. Graph. Stat. 2006;15:265–86.
[59] Hardoon DR, Shawe-Taylor J. Sparse canonical correlation analysis. Mach. Learn. 2011;83:331–53.
[60] Peharz R, Pernkopf F. Sparse nonnegative matrix factorization with $\ell 0$-constraints. Neurocomputing 2012;80:38–46.
[61] Liu R, Niang N, Saporta G, Wang H. Sparse Correspondence Analysis for Contingency Tables. arXiv [stat.ME] (2020).
[62] Park M, Kim D, Moon K, Park T. integrative analysis of multi-omics data based on blockwise sparse principal components. Int J Mol Sci 2020;21.
[63] De Tayrac M, Lê S, Aubry M, Mosser J, Husson F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. BMC Genomics 2009;10:1–17.
[64] Tang TM, Allen GI. Integrated Principal Components Analysis. arXiv [stat.ME] (2018).
[65] EL-Manzalawy Y, Hsieh T-Y, Shivakumar M, Kim D, Honavar V. Min-Redundancy and Max-Relevance Multi-view Feature Selection for Predicting Ovarian Cancer Survival using Multi-omics Data. doi:10.1101/317982.
[66] Spicker JS, Brunak S, Frederiksen KS, Toft H. Integration of clinical chemistry, expression, and metabolite data leads to better toxicological class separation. Toxicol Sci 2008;102:444–54.
[67] Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets: Multiple factor analysis. Wiley Interdiscip Rev Comput Stat 2013;5:149–79.
[68] Grossi E, Buscema M. Introduction to artificial neural networks. Eur J Gastroenterol Hepatol 2007;19:1046–54.
[69] Xie G et al. Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features. Genes 2019;10.
[70] Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res 2018;24:1248–59.
[71] Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. Briefings Bioinf 2020. https://doi.org/10.1093/bib/bbaa177.
[72] Martorell-Marugán J. et al. Deep Learning in Omics Data Analysis and Precision Medicine. in Computational Biology (ed. Husi, H.) (Codon Publications, 2019).
[73] Deng L et al. Pathway-guided deep neural network toward interpretable and predictive modeling of drug sensitivity. J Chem Inf Model 2020;60:4497–505.
[74] Hao J, Kim Y, Mallavarapu T, Oh JH, Kang M. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. BMC Med Genomics 2019;12.
[75] Lee JD, Hastie TJ. Learning the structure of mixed graphical models. J Comput Graph Stat 2015;24:230–53.
[76] Altenbuchinger M, Weihs A, Quackenbush J, Grabe HJ, Zacharias HU. Gaussian and Mixed Graphical Models as (multi-)omics data analysis tools. Biochim Biophys Acta (BBA) – Gene Regulatory Mech 2020;1863:194418.
[77] Fellinghauer B, Bühlmann P, Ryffel M, von Rhein M, Reinhardt JD. Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. Comput Stat Data Anal 2013;64:132–52.
[78] Manatakis DV, Raghu VK, Benos PV. piMGM: incorporating multi-source priors in mixed graphical models for learning disease networks. Bioinformatics 2018;34:i848–56.
[79] Petralia F, Wang P, Yang J, Tu Z. Integrative random forest for gene regulatory network inference. Bioinformatics 2015;31:i197–205.
[80] Zhong W et al. Inferring regulatory networks from mixed observational data using directed acyclic graphs. Front Genet 2020;11.
[81] Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data. Front Genet 2019;10.
[82] Lanckriet GRG. Learning the kernel matrix with semideønite programming. https://www.jmlr.org/papers/volume5/lanckriet04a/lanckriet04a.pdf (2004).
[83] Joachims T. Support Vector Machines. Learning to Classify Text Using Support Vector Machines 35–44 (2002) doi: 10.1007/978-1-4615-0907-3_3.
[84] Yang H, Cao H, He T, Wang T, Cui Y. Multilevel heterogeneous omics data integration with kernel fusion. Briefings Bioinf 2018. https://doi.org/10.1093/bib/bby115.
[85] Zhu B et al. Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers. Sci Rep 2017;7.
[86] Gönen M, Alpaydın E. Multiple kernel learning algorithms. J. Mach. Learn. Res. 2011;12:2211–68.
[87] Wilson CM, Li K, Kuan P-F, Wang X. Multiple-kernel learning for genomic data mining and prediction. doi: 10.1101/415950.
[88] Rakotomamonjy A, Bach F, Canu S, Grandvalet Y. SimpleMKL. J. Mach. Learn. Res. 2008;9:2491–521.
[89] Suzuki T, Tomioka R. SpicyMKL: a fast algorithm for Multiple Kernel Learning with thousands of kernels. Mach. Learn. 2011;85:77–108.
[90] Tao M et al. Classifying breast cancer subtypes using multiple kernel learning based on omics data. Genes 2019;10.
[91] Zhang Y, Li A, Peng C, Wang M. Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning. IEEE/ACM Trans Comput Biol Bioinf 2016;13:825–35.
[92] He Z, Zhang J, Yuan X, Zhang Y. Integrating somatic mutations for breast cancer survival prediction using machine learning methods. Front Genet 2021;11.
[93] Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods 2017;14:414–6.
[94] Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat Commun 2018;9.
[95] Mariette J, Villa-Vialaneix N. Unsupervised multiple kernel learning for heterogeneous data integration. doi: 10.1101/139287.
[96] Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics 2015;31:i268–75.
[97] Paul D, Chakraborty S, Das S, Xu J. Kernel k-Means, By All Means: Algorithms and Strong Consistency. arXiv [stat.ML] (2020).
[98] Wang B et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11:333–7.
[99] Ma T, Zhang A. Affinity network fusion and semi-supervised learning for cancer patient clustering. Methods 2018;145:16–24.
[100] Wen Y et al. Multi-dimensional data integration algorithm based on random walk with restart. BMC Bioinf 2021;22.
[101] Jarada T, Rokne J, Alhajj R. SNF-NN: Computational Method To Predict Drug-Disease Interactions Using Similarity Network Fusion and Neural Networks. doi:10.21203/rs.3.rs-56433/v1.
[102] Chierici M et al. Integrative network fusion: a multi-omics approach in molecular profiling. Front Oncol 2020;10:1065.
[103] Valdeolivas A et al. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics 2019;35:497–505.
[104] Shang H, Liu Z-P. Network-based prioritization of cancer genes by integrative ranks from multi-omics data. Comput Biol Med 2020;119:103692.
[105] Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. (1999).
[106] Liu W et al. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. Bioinformatics 2013;29:2169–77.
[107] Kim SY, Jeong H-H, Kim J, Moon J-H, Sohn K-A. Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. Biol Direct 2019;14.
[108] Kim SY, Choe EK, Shivakumar M, Kim D, Sohn K-A. Multi-layered network-based pathway activity inference using directed random walks: application to predicting clinical outcomes in urologic cancer. doi: 10.1101/2020.07.22.163949.
[109] Nelson W et al. To embed or not: network embedding as a paradigm in computational biology. Front Genet 2019;10.
[110] Yue X et al. Graph embedding on biomedical networks: methods, applications and evaluations. Bioinformatics 2019. https://doi.org/10.1093/bioinformatics/btz718.
[111] Luo Y et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 2017;8.
[112] Cho H, Berger B, Peng J. Diffusion component analysis: unraveling functional topology in biological networks. Res. Comput. Mol. Biol. 2015;9029:62–4.
[113] Xuan P et al. Gradient boosting decision tree-based method for predicting interactions between target genes and drugs. Front Genet 2019;10.
[114] Gligorijević V, Barot M, Bonneau R. deepNF: deep network fusion for protein function prediction. Bioinformatics 2018;34:3873–81.
[115] Zeng X et al. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics 2019;35:5191–8.
[116] Chen Y, de Rijke M. A Collective Variational Autoencoder for Top-N Recommendation with Side Information. in Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems 3–9 (Association for Computing Machinery, 2018).
[117] Manica M, Cadow J, Mathis R, Martínez MR. PIMKL: Pathway-induced multiple kernel learning. npj Syst Biol Appl 2019;5.
[118] Tepeli YI, Ünal AB, Akdemir FM, Tastan O. PAMOGK: a pathway graph kernel-based multiomics approach for patient clustering. Bioinformatics 2021;36:5237–46.
[119] Wu Z et al. A Comprehensive survey on graph neural networks. IEEE Trans Neural Networks Learn Syst 2021;32:4–24.
[120] Wang Z, Zhou M, Arnold C. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing. Bioinformatics 2020;36:i525–33.
[121] Singha, M. et al. GraphGR: A graph neural network to predict the effect of pharmacotherapy on the cancer cell growth. doi: 10.1101/2020.05.20.107458.

[122] Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 2018;34:i457–66.
[123] Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. doi:10.1101/2020.07.08.192930.
[124] Muzio G, O'Bray L, Borgwardt K. Biological network analysis with deep learning. Briefings Bioinf 2021;22:1515–30.
[125] Zhang L et al. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front Genet 2018;9:477.
[126] Xu J et al. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. BMC Bioinf 2019;20.
[127] Lin Y, Zhang W, Cao H, Li G, Du W. Classifying breast cancer subtypes using deep neural networks based on multi-omics data. Genes 2020;11:888.
[128] Islam MM et al. An integrative deep learning framework for classifying molecular subtypes of breast cancer. Comput Struct Biotechnol J 2020;18:2185–99.
[129] Yang H, Chen R, Li D, Wang Z. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. Bioinformatics 2021. https://doi.org/10.1093/bioinformatics/btab109.
[130] Gaynanova I, Li G. Structural learning and integrative decomposition of multi-view data. Biometrics 2019;75:1121–32.
[131] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–91.
[132] Zhang S et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucl Acids Res 2012;40:9379–91.
[133] Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 2015;btv544. https://doi.org/10.1093/bioinformatics/btv544.
[134] Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLoS ONE 2017;12: e0176278.
[135] Cantini L et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun 2021;12.
[136] Luo Y, Tao D, Ramamohanarao K, Xu C, Wen Y. Tensor canonical correlation analysis for multi-view dimension reduction. In: 2016 IEEE 32nd International Conference on Data Engineering (ICDE). https://doi.org/10.1109/icde.2016.7498374.
[137] Tenenhaus M, Tenenhaus A, Groenen PJF. Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. Psychometrika 2017;82:737–77.
[138] Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinf 2014;15:162.
[139] Chalise P, Ni Y, Fridley BL. Network-based integrative clustering of multiple types of genomic data using non-negative matrix factorization. Comput Biol Med 2020;118:103625.
[140] Fujita N, Mizuarai S, Murakami K, Nakai K. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. Sci Rep 2018;8.
[141] Argelaguet R et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol 2018;14.
[142] Argelaguet R et al. MOFA: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol 2020;21.
[143] Clark C, Dayon L, Masoodi M, Bowman GL, Popp J. An integrative multi-omics approach reveals new central nervous system pathway alterations in Alzheimer's disease. Alzheimers. Res. Ther. 2021;13:71.
[144] Palou-Márquez G, Subirana I, Nonell L, Fernández-Sanlés A, Elosua R. DNA methylation and gene expression integration in cardiovascular disease. Clin. Epigenetics 2021;13:75.
[145] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 2009;25:2906–12.
[146] Mo Q et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl. Acad. Sci. U.S.A. 2013;110:4245–50.
[147] Mo Q et al. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics 2018;19:71–86.
[148] Ma M et al. Integrative analysis of genomic, epigenomic and transcriptomic data identified molecular subtypes of esophageal carcinoma. Aging 2021;13:6999–7019.
[149] Zhao Y, Gao Y, Xu X, Zhou J, Wang H. Multi-omics analysis of genomics, epigenomics and transcriptomics for molecular subtypes and core genes for lung adenocarcinoma. BMC Cancer 2021;21:257.
[150] Mo Q, Li R, Adeegbe DO, Peng G, Chan KS. Integrative multi-omics analysis of muscle-invasive bladder cancer identifies prognostic biomarkers for frontline chemotherapy and immunotherapy. Commun Biol 2020;3:784.
[151] Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Statistics 2013;7.
[152] Feng Q, Jiang M, Hannig J, Marron JS. Angle-based joint and individual variation explained. J Multivariate Anal 2018;166:241–65.
[153] Kaplan A, Lock EF. Prediction with dimension reduction of multiple molecular data sources for patient survival. Cancer Inf 2017;16.
[154] Ponzi E, Thoresen M, Nøst, TH, Møllersen K. Integrative analyses of multi-omics data improves model predictions: an application to lung cancer. bioRxiv 2020.10.02.299834 (2020) doi: 10.1101/2020.10.02.299834.
[155] Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. IEEE/ACM Trans Comput Biol Bioinf 2019;16:841–50.
[156] Wang T. et al. MORONET: Multi-omics Integration via Graph Convolutional Networks for Biomedical Data Classification. doi: 10.1101/2020.07.02.184705.
[157] Wang W et al. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics 2013;29:149–59.
[158] Zhu R, Zhao Q, Zhao H, Ma S. Integrating multidimensional omics data for cancer outcome. Biostatistics 2016;17:605–18.
[159] Chai H et al. Analysis of cancer gene expression data with an assisted robust marker identification approach. Genet Epidemiol 2017;41:779–89.
[160] Wu C, Zhang Q, Jiang Y, Ma S. Robust network-based analysis of the associations between (epi)genetic measurements. J. Multivar. Anal. 2018;168:119–30.
[161] Kim D-C et al. Integration of DNA Methylation, Copy Number Variation, and Gene Expression for Gene Regulatory Network Inference and Application to Psychiatric Disorders. in 2014 IEEE International Conference on Bioinformatics and Bioengineering 238–242 (2014).
[162] Cai X, Bazerque JA, Giannakis GB. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS Comput Biol 2013;9:e1003068.
[163] Fortelny N, Bock C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. doi:10.1101/794503.
[164] Balluff B et al. Integrative clustering in mass spectrometry imaging for enhanced patient stratification. Proteomics Clin Appl 2019;13:e1800137.
[165] Ma S, Ren J, Fenyö D. Breast cancer prognostics using multi-omics data. AMIA Jt Summits Transl Sci Proc 2016;2016:52–9.
[166] McDonald ME et al. Molecular characterization of non-responders to chemotherapy in serous ovarian cancer. Int J Mol Sci 2019;20.
[167] Zhu Y et al. Ensemble transfer learning for the prediction of anti-cancer drug response. Sci Rep 2020;10:18040.
[168] Herrmann M, Probst P, Hornung R, Jurinovic V, Boulesteix A-L. Large-scale benchmark study of survival prediction methods using multi-omics data. Brief. Bioinform. 2020. https://doi.org/10.1093/bib/bbaa167.
[169] Pierre-Jean M, Deleuze J-F, Le Floch E, Mauger F. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. Brief. Bioinform. 2020;21:2011–30.
[170] Wei Z, Zhang Y, Weng W, Chen J, Cai H. Survey and comparative assessments of computational multi-omics integrative methods with multiple regulatory networks identifying distinct tumor compositions across pan-cancer data sets. Briefings Bioinf 2020. https://doi.org/10.1093/bib/bbaa102.
[171] McCabe SD, Lin D-Y, Love MI. Consistency and overfitting of multi-omics methods on experimental data. Brief. Bioinform. 2020;21:1277–84.
[172] Zhu W, Xie L, Han J, Guo X. The application of deep learning in cancer prognosis prediction. Cancers 2020;12.