BioData Mining

REVIEW

Open Access

Network-based analyses of multiomics data in biomedicine

Rachit Kumar[1,2], Joseph D. Romano[3,4] and Marylyn D. Ritchie[3,4,5*]

*Correspondence:
Marylyn D. Ritchie

marylyn@pennmedicine.upenn.edu

Full list of author information is available at the end of the article

Abstract

Network representations of data are designed to encode relationships between concepts as sets of edges between nodes. Human biology is inherently complex and is represented by data that often exists in a hierarchical nature. One canonical example is the relationship that exists within and between various -omics datasets, including genomics, transcriptomics, and proteomics, among others. Encoding such data in a network-based or graph-based representation allows the explicit incorporation of such relationships into various biomedical big data tasks, including (but not limited to) disease subtyping, interaction prediction, biomarker identification, and patient classification. This review will present various existing approaches in using network representations and analysis of data in multiomics in the framework of deep learning and machine learning approaches, subdivided into supervised and unsupervised approaches, to identify benefits and drawbacks of various approaches as well as the possible next steps for the field.

Keywords Review, Multiomics, Networks, Graphs, Deep learning, Machine learning, Supervised learning, Unsupervised learning

Background

Multiomics - also known as "multiple omics" - includes data types such as genomics, transcriptomics, epigenomics, metabolomics, and proteomics, among many others [1, 2, 3]. The analysis and integration of multiomics data has become an increasingly relevant topic in biomedicine as advances in methods for collecting omics data has led to such data becoming increasingly accessible and has even allowed for the collection of multiple -omics data in the same experimental process from the same biological samples [4].

Such analyses have enabled interrogation of the underlying mechanisms behind the pathogenesis of many diseases, most notably various types of cancers through data collected by The Cancer Genome Atlas (TCGA) Research Network, used by many of the papers discussed in this review in part due to the wide availability and coverage of many -omics datasets within TCGA [5, 6]. Furthermore, many datasets with varying levels of multiomics data types have become available in more recent years to study a number of disease domains including frontotemporal dementia [7], Alzheimer's disease [8, 9], kidney disease [10, 11, 12], and liver disease [13, 14], to name a few, highlighting the

need for additional multiomics integration methods. However, multiomics analyses face many unique challenges; multiomics datasets are often very heterogeneous, sparse, and extraordinarily high-dimensional. At the same time, sample sizes for many of these datasets are still relatively small despite advancements in data collection, in part due to technical limitations and complexities that still arise with these techniques in a wet lab setting [15, 16, 17].

Network, or graph, representations of data provide a structured way to explicitly model and represent these relationships by representing concepts as nodes and representing relationships between concepts as edges [18, 19]. In some methods, the concepts are drawn directly from the features found in the dataset in question and relationships obtained from pre-existing knowledge graphs [20, 21]. In others, the relationships are inferred in an attempt to find or identify novel relationships [22, 23]. These networks can then be used alongside graph machine learning and traditional network analyses to explicitly incorporate known relationships into analyses [24, 25, 26]. In multiomics, network approaches and analyses thereof are becoming increasingly common as a way to try to address many of the issues inherent to multiomics analysis. The nodes often represent features of various omics datasets, such as variants found in genomics and genes found in transcriptomics [22]. The edges are often constructed from underlying biological knowledge stored in databases such as the Kyoto Encyclopedia of Genes of Genomes (KEGG) and others [27]. Alternatively, omics data can be embedded into the network as features induced on nodes or edges [24]. Within the subset of network analyses of multiomics data, many papers have made use of traditional machine learning approaches. However, there has been an emergence of methods that make use of deep learning on graphs for analysis of the same data. In this topical review, we will discuss both traditional and deep learning methods on graphs as well as their respective benefits and drawbacks as applied to multiomics data integration and analysis.

Furthermore, when considering methods to analyze multiomics data on a broad scale, such tasks are usually divided into two types of common paradigms: supervised learning and unsupervised learning. In the case of supervised learning, known labels are used as a target; for example, a classification task that seeks to identify if a network indicates that a patient has a disease or not based on biological data [28]. In unsupervised learning, the goal is to identify patterns that exist within unlabeled data that best characterize the data [28]. In this review, we will highlight a two-axis framework to discuss supervised and unsupervised methods for multiomics data integration using networks via traditional machine learning and deep learning (see Fig. 1).

While many prior reviews have explored either (1) network analyses in multiomics data [29, 30], (2) deep learning and machine learning methods in multiomics analyses [31], or (3) have compared supervised and unsupervised approaches broadly in multiomics [32, 33], none have explicitly discussed the benefits and drawbacks of deep learning and machine learning methods on networks. Considering these methods in this context of networks is particularly important because, as we discuss further below, using network representations to incorporate prior knowledge into analyses can mitigate some of the drawbacks of both deep learning and traditional machine learning.

Furthermore, no previous reviews have further subdivided these approaches into supervised and unsupervised learning to create a two-axis categorization of supervised deep learning, unsupervised deep learning, supervised machine learning, and

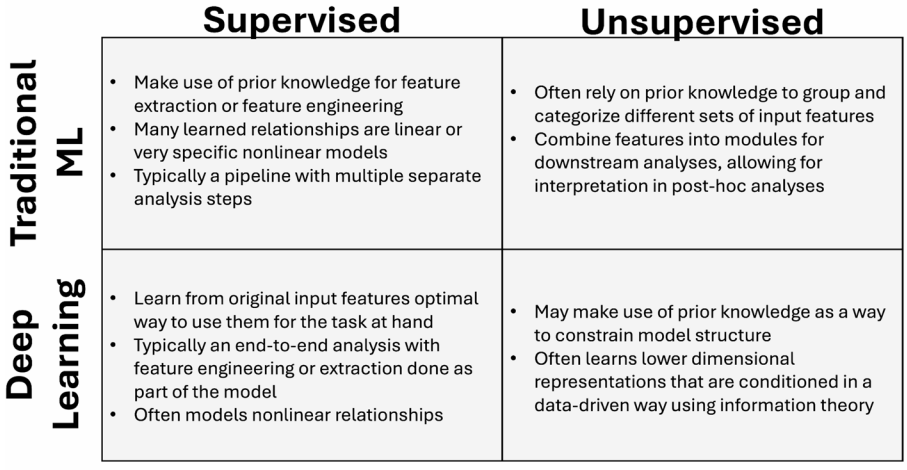

Fig. 1 A brief visualization showing some of the characteristics of different algorithms within the four quadrants of the supervised-unsupervised and traditional ML-deep learning dual dichotomies. ML = "machine learning". Each bullet point of this visual are further discussed in the relevant sections of the paper

unsupervised machine learning. As network representations and analysis of multiomics data alongside deep learning on graphs becomes increasingly popular and powerful, it is important to elucidate important criteria that often play a role in the decision-making process of data scientists and biomedical informaticians when deciding how to analyze their data, including interpretability, resource requirements, and technical limitations. It is also crucial to identify existing gaps in the field so that future methods can aim to address these gaps.

This review thus presents a novel framework for considering and comparing these methods as well as identifying possible next steps that will improve the landscape of future multiomics analyses that make use of network representations, both from a practical perspective (of methods application) and from a theoretical perspective (of methods development).

Existing network analysis methods

In this topical review, we first subdivide network analysis methods into classes of supervised and unsupervised approaches, as this is often the first decision point in any analysis paradigm that is determined based on the definition of the problem [28, 34]. For example, those trying to identify whether a patient has a disease or not by using data from cases and controls will likely look primarily for supervised approaches, while those trying to subtype a disease into different subclassifications without having access to known labels for these subtypes will likely look for unsupervised approaches. In this second example, the goal is to identify whether there are subtypes of a disease, and perhaps how many subtypes exist; thus, there are no known labels to learn.

Within each class of approach, we will further subdivide into traditional (non-deep-learning) and deep learning approaches. As deep learning becomes increasingly popular, many are faced with the question of whether to consider deep learning approaches in their analyses and what the limitations and drawbacks thereof may be, especially as compared to more traditional approaches. While deep learning approaches are sometimes the more novel, innovative, and exciting approach, the research question and underlying model may not warrant that level of sophistication. Sometimes, a deep learning model

is overkill for the question at hand and a more traditional machine learning approach is favorable. In this review, our goal is to aid a researcher in making that assessment for their multiomics data/question for their study.

We note that we do not include semi-supervised learning or reinforcement learning in this review. This is due to a relative lack of existing published literature in either network-based semi-supervised learning in multiomics or network-based reinforcement learning in multiomics at the time of writing this review.

Supervised approaches

Supervised learning generally refers to the class of approaches where a target or outcome is known for a subset of the data used to optimize and evaluate the method in question [28]. In multiomics, methods have been developed for a variety of supervised tasks, including disease category classification and survival prediction, among many others [35]. Some selected papers are discussed below.

In the context of traditional machine learning and deep learning, approaches to supervised learning can be quite varied, especially when considering network-based approaches. In many cases for supervised traditional approaches, they make use of prior knowledge to create networks that allow for feature aggregation or selection (commonly termed "feature extraction" or "feature engineering") that they then use as inputs to standard machine learning algorithms, creating a sort of pipeline that allows for training on a task in a data-driven way. However, the feature aggregation or selection is not always dependent on the task at hand, and the "trainable" part of such approaches is often the downstream classification or regression algorithm that learns from those features. Furthermore, many of these learned relationships tend to be linear in nature or encode very specific nonlinear relationships. In contrast, deep learning methods for supervised learning on networks for multiomics data often act in an end-to-end manner where they learn from the original input features a set of "engineered/extracted features" that are optimized for the task at hand and simultaneously learn how to use those features for the downstream task at hand, allowing them to essentially learn the optimal way to "use" (through combining, selection, or aggregation) the input features for the specific task at hand.

Supervised traditional machine learning approaches: a set of examples

iOmicsPASS [36]: Koh et al. use multiomics data to calculate scores for all interactions found in a pathway database known as ConsensusPathDB through co-expression that aggregates relationships from a variety of other biological knowledge sources. They then use these scores from multiomics data (in their paper, genomics, transcriptomics, and proteomics) as features for classifying tumor subtypes in TCGA and Clinical Proteomic Tumor Analysis Consortium (CPTAC) data by using a modified nearest shrunken centroid (NSC) classification algorithm that reweights centroids to account for class imbalances. They then use feature weights from the NSC algorithm to identify specific interactions that contributed to the classification algorithm.

Integrative Network Fusion [37]: Chierici et al. apply principles of similarity network fusion (SNF) as well as variable juxtaposition (juXT) to find a set of optimal features from multiomics data (gene expression, proteomics, copy number variation, and methylomics in varying combinations) that they then use as inputs to an arbitrary classifier

(that can be switched out) to predict a variety of metrics, including estrogen receptor status and cancer subtyping in the TCGA Breast Invasive Carcinoma (BRCA) dataset as well as overall survival on other TCGA datasets. They first filter features separately using SNF on the multiomics data to find similar features as well as using juXT to rank features that they then intersect to use as a set of features in downstream classifiers.

iDINGO [38]: Class et al. present an interactive tool that allows for integration of multiomics data using a chain graph model where individual omics datasets are assumed to impact other omics data in a hierarchical, one-way manner. They can then make use of these integrated omics networks to perform differential pathway enrichment analyses across different groups (such as individuals with a disease versus those without a disease), showing their method in a breast cancer dataset.

NetMIM [39]: Zhu et al. propose a method that integrates multiomics data using a network-based Bayesian model that seeks to better handle missingness where individuals may only have information from select omics types. They incorporate gene-pathway mappings alongside feature dependencies using Markov random fields to perform initial feature selection; they then integrate features across biological mechanisms using a Bayesian model before identifying subunits that are then used to aggregate features and predict clinical outcomes. They handle missing data by imputing the data for each individual from available omics data using a conditional distribution at each iteration of the model parameter optimization.

Supervised deep learning approaches: a set of examples

GraphSurv [40]: Wang et al. implement a graph convolutional network (GraphSurv) with the goal of predicting cancer prognosis survival on TCGA datasets. They developed a gene-gene network using data and knowledge from KEGG, using information about which gene products interacted in which pathways as the information to build their edges between genes. They then embed multiomics data from the modalities of RNA-sequencing, copy number variation, and DNA methylation as node features on each gene. They then take this network and apply a graph convolutional network with a regression layer to predict patient survival from each network, achieving better performance than non-graph-based methods.

LAGProg [41]: Zhang et al. implement a local augmented graph convolutional network (LAGProg), also to predict cancer prognosis and survival from TCGA datasets. Very much like in GraphSurv, they develop a gene-gene network with multiomics data as node features. However, they then refine these features by using an autoencoder that incorporates information from neighboring genes to compute a set of augmented features for each gene. They then use a graph convolutional network with a regression layer to process the gene-gene network and node features to predict patient survival from each network, noting that their method is similar to GraphSurv but that their augmentation improves performance by allowing for more robust sets of multiomics features to be considered from central genes.

MoGCN [42]: Li et al. use a graph convolutional network on a patient-patient similarity network to classify cancer subtypes from BRCA samples in TCGA from theoretically any combination of patient-level data, though they test on genomics, transcriptomics, and proteomics data. They do this by first using the multiomics data individually alongside similarity network fusion to create a patient-patient similarity network. They then

pass the multiomics data through an autoencoder to generate a set of lower-dimensional features for every patient that they then use as node features in their patient-patient similarity network. They then apply a graph convolutional network on this patient-patient similarity network to perform node classification - that is, they predict the cancer subtype for each patient (node) within their network.

MOGAT [43]: Tanvir et al. use a graph attention network on a patient-patient similarity network to classify cancer subtypes from BRCA samples in TCGA, similarly to as in MoGCN with the exception of using different multiomics data (copy-number variations, co-expression data, RNA-seq data, methylation data, and single-nucleotide variations, comprising genomics, transcriptomics, and epigenomics) along clinical data. They create a patient-patient similarity network using their multiomics data and then embed a set of pre-selected features from each sample's multiomics data into the network as node features on each patient. After this, they apply a graph attention network to this network to create embeddings that they pass to a classification layer and predict cancer subtypes for each patient within their network.

MOGONet [44]: Wang et al. use a graph convolutional network in tandem with a view-correlation discovery network (VCDN) to perform a wide variety of biomedical classification tasks, including Alzheimer's disease prediction and cancer subtyping with an arbitrary number of omics datasets (in this paper, they use mRNA expression, methylation, and miRNA expression). They do so by constructing patient-patient similarity networks on individual omics datasets, processing those networks using a graph convolutional network to produce initial predictions for the classification and calculating correlations between each omics dataset, and then using the VCDN to integrate the correlation of the predictions from each omics dataset together into one final prediction for each sample.

MOGDx [45]: Ryan et al. propose a tool that makes use of a graph convolutional network that they call Multi-Omic Graph Diagnosis (MOGDx) that operates on patient similarity networks in order to perform subtype classification on multiple datasets. They perform initial feature selection to compute similarity using multiple omics modalities and then fuse across modalities using similarity network fusion, providing the network structure for the graph neural network where nodes are individuals and edges connect the 15 nearest neighbors based on similarity. As node features, they use the original omics modalities without feature selection, processed through a fully-connected encoder. By training in this way, they are able to classify individual patients after multiple graph convolutions, and they additionally show that this model provides interpretability into which modalities contribute to a classification through omics ablations.

Unsupervised approaches

Unsupervised learning generally refers to the class of approaches where there is no specific target or label for the data used; instead, the aim is to create a representation of the data by inferring patterns that incorporate a useful set of information or provide insights about the data [28]. In multiomics data analysis, many of these methods focus on single-cell analyses to disaggregate different cell types by using lower dimensional embeddings, while others operate at the patient level to create embeddings that can be used to compare patients [46]. Recently, network integration methods have been used to pair unpaired data across omics datasets or to better model relationships between

omics datasets when generating embeddings. In many cases, these embeddings are tested by evaluating how well they can be used on downstream tasks such as classification or regression, but the labels themselves are not necessary for the construction of the embeddings themselves - an important distinction between supervised and unsupervised learning.

In the context of traditional machine learning and deep learning approaches, these methods often have similar underlying goals of producing an embedding of the original data in a lower-dimensional space; however, the approaches that are taken to produce these embeddings can be very different. In traditional machine learning, methods often rely on prior knowledge to group and categorize different sets of input features together into what many methods call "modules", using aggregation techniques that are relatively standard to combine features assigned to a module (such as mean aggregation), which enables easier interpretation of the contribution of individual modules to any downstream task or cluster. In contrast, deep learning approaches often learn lower dimensional representations that are conditioned in a data-driven way using information theory - for example, trying to reconstruct the original input data in the case of autoencoders - which enforces a constraint that the representations contain some meaningful information; however, it can be difficult to identify the meaning of any given axis within the lower-dimensional embedding as it is not always tied to a specific feature (or "module" as above). Many deep learning approaches that use networks in this fashion also make use of prior knowledge as a way to provide a "seed" to constrain or provide some prior structure to their model in terms of how the features may map to aspects of the lower-dimensional space, which can improve interpretability slightly.

Unsupervised traditional machine learning approaches: a set of examples

netOmics [47]: Bodein et al. construct multiomics networks consisting of relationships between concepts from gene expression, proteomics, and metabolomics to perform analyses and generate network representations of data that can then be used for further downstream analyses in a package they call netOmics. They construct their network by moderately preprocessing their omics data independently and then draw on interaction databases like BioGRID. They show that this network represents useful information by performing enrichment analyses and topology analyses on the resulting network that show relationships that are strongly associated with known relationships on three datasets, including a HeLa cell dataset, a maize genetics dataset, and a diabetes dataset.

PARADIGM [48]: Vaske et al. use genome copy number variation, gene expression, and DNA methylation in their method called PARADIGM where they attempt to infer patient-specific pathway activity to understand personalized mechanisms of disease in breast and brain cancer. They make use of known biological information to construct subnetworks representing individual pathways and then perform knowledge-aware aggregation of data into pathways and construct integrated pathway activity scores that they then use as a lower-dimensional representation of a patient. They show that these scores embed useful information by comparing their usefulness in downstream tasks such as classification, though these embeddings can be generated independently of any labels (as an unsupervised algorithm).

COSMOS [49]: Dugourd et al. make use of prior knowledge encoded in signaling, metabolic, and regulatory networks to integrate transcriptomics, phosphoproteomics, and

metabolomics data in their Causal Oriented Search of Multi-Omics Space (COSMOS) approach. They use these relationships alongside putative activity of various proteins to infer and explore network-level dysregulation and attempt to identify causal mediators of clear cell renal cell carcinoma as a case study. They make use of specific known relationships (such as activation and inhibition) to create knowledge-aware network scores and perform enrichment analyses, finding that they are able to recapitulate many known drug targets in renal cell carcinoma.

SUBATOMIC [50]: Loers and Vermeirssen write about their Subgraph Based Multiomics Clustering Framework (SUBATOMIC). They start by taking in a pre-created set of networks of multiomics data with overlapping nodes. They then create subgraphs of 3 and 2 nodes with different properties based on the node relationships within those subgraphs, taking into account the different node types and edge types corresponding to different multiomics relationships (between and within omics sets). They then use hypergraph processing to represent the subgraphs as new hyperedges within a hypergraph that they then use to represent different "modules" on which expression data is incorporated to represent different conditions. They use the differences between these modules in different disease states and patients to understand patient-specific differences and identify which modules (and groups of modules) are important in a given disease.

NetICS [51]: Dimitrakopoulos et al. present Network-based Integration of Multi-omics Data (NetICS), a method that makes use of multiomics data represented in a functional interaction network to generate priority lists of genes based on lower-dimensional embeddings and relationships representing the gene as it exists within the network. They construct a functional interaction network based on known relationships and use it to represent directed edges between transcriptome and proteome data. They then make use of network diffusion methods to create representations at the gene level that then function as sample-specific gene embeddings. They take the sample-specific embeddings for each gene (for a given subpopulation) and aggregate them to construct a set of scores for all of the genes in their dataset, then order them by their respective weights to prioritize which genes are important within a subpopulation.

Lemon-Tree [52]: Bonnet et al. make use of module network inference in their method called Lemon-Tree to construct sets of co-expression modules that incorporate information from various omics datasets (in this paper, copy number variation and gene expression). They then construct relationships between modules and regulatory programs by building an ensemble of the results from many runs of module network inference as well as annotating the modules using known biological knowledge from Gene Ontology [53], giving them a set of features on each module and on each regulatory program that they can use alongside patient-specific values for each omics dataset as a representation for each patient as well as for enrichment analyses across patients within different subgroups.

Unsupervised deep learning approaches: a set of examples

GCN-SC [54]: Gao et al. integrate single-cell multiomics data (gene expression, chromatin accessibility, and protein expression) by finding pairs of related cells within individual omics layers and between omics layers using mutual nearest neighbors to construct a network that represented nodes as cells and edges between identified pairs across

multiomics datasets. They then use a graph convolutional network and non-negative matrix factorization to create a set of lower-dimensional embeddings of the data that they could then use for cell type clustering.

GLUE [55]: Cao and Gao develop a framework that makes use of a graph convolutional network in the form of graph variational autoencoders in a framework called graph-linked unified embedding (GLUE) to pair and integrate multiomics data (specifically gene expression, chromatin accessibility, and methylation) for the purposes of creating single-cell embeddings. They first use the features from their multiomics data and pull relationships from knowledge graphs, then use a graph variational autoencoder to reduce the graph to a set of latent feature variables. They then apply this reduction to their cell-specific multiomics features to create a lower-dimensional set of latent variables for each patient that they then use to characterize and distinguish cells in a more comprehensive way, showing consistent clustering across omics types and even better discrimination of certain cell types.

MAE [56]: Ma and Zhang use autoencoders across multiomics types with network constraints in a framework that they call Multi-view Factorization Autoencoder (MAE) to integrate multiple multiomics datasets together. They make use of multiple autoencoders, one for each omics data type, to generate embeddings ("views") for each data type and then have a module that combines multiple views together. They impose network constraints using existing biological interaction networks by using such knowledge to determine the similarity of features through shared pathways and interactions, then they use those similarities as inductive biases through incorporated regularization terms to ensure that related features are represented similarly in their embedding as in the original network.

SpatialGlue [57]: Long et al. use methods similar to a graph autoencoder to integrate spatial multiomics data with the goal of identifying spatial domains in a method that they call SpatialGlue. They make use of graph convolutional layers as an encoder, then aggregate data spatially across modalities as well as within modalities, then use further graph convolutional layers as a decoder. The method relies on spatial resolution of multiomics data to construct a graph of nearby points; however, they also enrich this graph with edges corresponding to points that are close to each other based on a principal components analysis (PCA) embedding space. They train this model on its ability to reconstruct the original data as well as its ability to project different omics types into similar representation spaces, showing that this model performs well on various downstream tasks and generalizes across multiple datasets, including a human lymph node dataset and a mouse brain dataset, among others.

SSGATE [58]: Lv et al. propose a model that they call SSGATE, a "multi-omics integration method based on dual-path graph attention autoencoder". SSGATE makes use of either single-cell or spatial multiomics data by constructing nearest-neighbor graphs based on single-cell expression profiles or geometric distance, respectively. It then learns a lower-dimensional embedding of each individual using graph attention mechanisms using standard autoencoder loss functions, allowing the embeddings to be used for downstream tasks. They use this model for cell clustering and developmental trajectory analysis on human bone marrow, mouse spleen, and mouse thymus datasets, integrating transcriptomics and proteomics data.

Comparisons along axes of interest

Supervised versus unsupervised approaches

In terms of considering supervised versus unsupervised approaches, this is usually a domain- and paradigm-specific problem relating to whether one has labels for their data that they would like to predict, succinctly shown in Figs. 1 and 2. It is important to consider and discuss these methods as this is a critical inflection point for most projects making use of multiomics data; however, comparing supervised to unsupervised approaches directly can be difficult due to the inherent different goals of these two approaches. That being said, many commonalities can be observed between these methods: in many cases, both classes of approaches have several examples that make heavy use of prior knowledge from existing knowledge graphs or annotations, which may sometimes be flawed or incomplete; however, they are generally designed to be robust to changing knowledge (either by being able to operate on different versions of the knowledge graph or by being able to reconstruct the relationships from an external data source).

Additionally, more supervised approaches operate at the patient or individual level while more unsupervised approaches operate on the single-cell level. Both of these are viable paradigms for multiomics integration but represent vastly different levels of biological organization; however, at a fundamental level, many of the approaches in supervised learning can be used on a single-cell level and vice-versa, and this may represent an interesting area of future research and development. For example, the frameworks used in MOGAT, MOGONet, or MOGDx could all theoretically be used at the single-cell level as opposed to the patient level. However, there are some distinctions between patient-level and single-cell-level data relating to differences in how data may

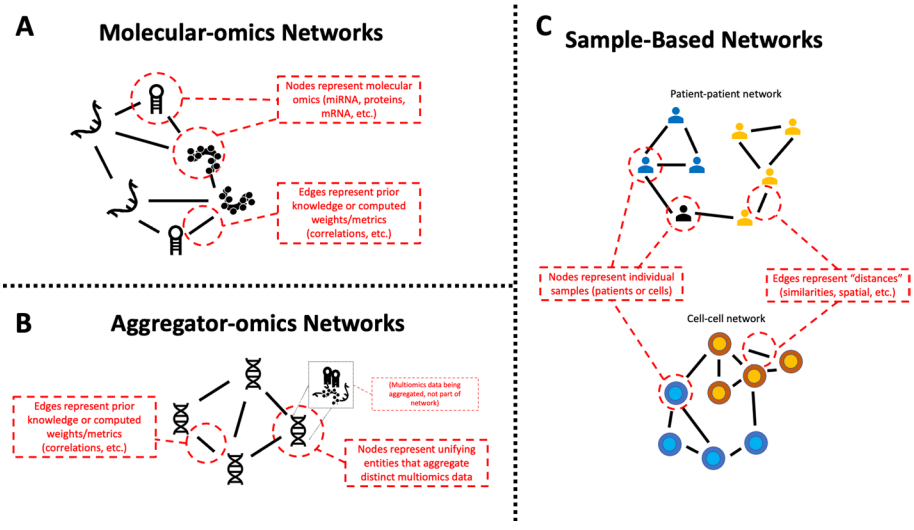

Fig. 2 A visual of several network representations of multiomics data that have been used in multiomics integration. We classify the ways networks have been represented broadly into three categories: (**A**) Molecular-omics networks, where nodes represent individual omics types and edges usually represent prior knowledge or computed weights; (**B**) Aggregator-omics networks, where nodes are a single entity (such as a gene) that aggregates multiple omics types into one and edges usually represent prior knowledge or computed weights; (**C**) Sample-based networks, where nodes are individual samples where multiple omics types have been collected (such as patients or cells) and edges generally represent some metric of similarity or distance between samples, such as feature similarity or spatial distance

be distributed or collected (such as increased sparsity in single-cell data) that may make direct translation between these settings more difficult.

Furthermore, lower-dimensional embeddings generated by unsupervised learning approaches are inherently generalizable, which can allow them to be translated to a variety of downstream tasks. This enables one to train the process of generating enriched embeddings on much larger datasets that may not be task-specific or labeled and then "transfer" that knowledge, in a sense, to specific, supervised tasks such as patient or cell classification. This means that supervised and unsupervised tasks do not necessarily need to be inherently exclusive, and in fact, some of the methods discussed such as MoGCN make use of unsupervised embedding generation prior to making a prediction.

Deep learning versus traditional machine learning

When considering whether to use deep learning or traditional machine learning, there are many critical considerations to make, as described in Fig. 1; Table 1. Among the methods discussed (particularly for the supervised approaches), deep learning tends to outperform traditional machine learning; however, this comes at the expense of interpretability. Within both supervised and unsupervised approaches, the traditional machine learning methods tend to be more interpretable with an easier path to understanding the underlying contribution of individual omics features and even entire omics layers to ensure that their relation to the outputs are intuitive and potentially clinically actionable [59, 60, 61]. Furthermore, the relative simplicity of these approaches mean that overfitting (while still possible) is far less likely due to the reduced parameter space and complexity [62, 63, 64]. However, it is important to note that if the relationship(s) between inputs and outputs is known, then the standard considerations are less reliable and one should instead determine the simplest model that models those relationships effectively rather than looking to the dichotomy of deep learning or traditional learning.

In contrast, the deep learning methods tend to rely on the training process and defined objective functions to ensure reliable relationships between omics features and the outputs (whether supervised or unsupervised). By their nature, they attempt to incorporate and model potential nonlinear relationships between their feature set at each step, which makes understanding the exact relationship of a feature to the output or to intermediate learned features difficult [59, 60, 61]. Furthermore, they tend to be much more highly parameterized, leading to the potential for overfitting with small sample sizes or for the methods to be poorly generalizable to external datasets [62, 63, 64].

It is important to note that recent developments in the field of deep learning can help to mitigate many of the canonical drawbacks of using deep learning over machine learning, especially in the context of network analyses. For example, the field of explainability and interpretability of graph convolutional networks and other graph neural networks has been making steady progress, with many methods being developed to highlight important regions of a graph or even specific edges within a graph that contribute to a prediction [65, 66, 67]. This has been underutilized in multiomics integration (in the methods discussed in this review) and represents a potential area of tackling both of the issues related to deep learning in this space: (1) being able to identify specific contributions; and (2) being able to determine whether overfitting is occurring by comparing important components across individual samples or patients.

Table 1 A discussion of the primary decision point for deciding between supervised and unsupervised learning (top subtable) and the various considerations that a researcher might make when deciding between deep learning approaches and traditional machine learning (ML) methods (bottom subtable). Note that this is merely a sampling of possible considerations and that not every consideration is listed here; for example, having access to limited computational resources May require a researcher to more heavily consider traditional ML techniques

Consideration	Answer for Supervised Learning	Answer for Unsupervised Learning	Comments/Caveats
Do you have meaningful labels for your data that you seek to predict?	Yes, you do have labels that you seek to predict	No, you do not have labels or you can only create inductive labels such as reconstruction	Note that this consideration is not as binary as it may seem. It is possible that unsupervised learning can still play a role even if you do have labels for your data; for example, if you have a large quantity of data but only a subset is labeled, you can use unsupervised learning to generate embeddings using the much larger dataset (which may be richer due to the larger quantity of data) and then train a prediction head on top of the unsupervised embeddings for the labeling.
Consideration	**Answer for Traditional Machine Learning**	**Answer for Deep Learning**	**Comments/Caveats**
How important is it that your model is interpretable or explainable?	Very important	Not as important	Many deep learning approaches, particularly those that use graph neural networks or attention-based methods, are often able to layer in interpretability through either architectural design or post-hoc model explainability. Depending on which deep learning approach you are considering, you may be able to maintain this consideration even when using deep learning.
Do you have a good number of samples relative to features?	No	Yes	Deep learning methods tend to be more prone to overfitting with small sample sizes due to their general capacity to model nonlinear relationships.
How important is it that your model can capture nonlinear complexities?	Can compromise	Critical	Deep learning models are more effective at capturing such nonlinear complexities without a priori knowledge of what nonlinear relationships may be important, which may allow them to achieve better performance in the presence of good enough sample sizes.
How much domain expertise do you have? (Can you perform feature engineering?)	Substantial expertise (yes)	Limited expertise (no)	Deep learning models are constructed in an end-to-end manner, often learning representations of the input data optimized for the training paradigm in question. This means that it can perform its own feature engineering (and extraction) that can help mitigate a lack of domain knowledge in constructing valid features from more complex or high-dimensional data.

Network representations of multiomics data: observations

Table 2 provides a brief overview of all of the methods discussed in this paper, categorizing them by supervised and unsupervised learning as well as deep learning versus traditional machine learning, and Table 3 provides additional information about the nodes and edges in each paper discussed which gives an insight into how the networks that represent multiomics data are defined and constructed.

From the papers that we have presented in this review alone, it is clear that there exists a huge amount of diversity in the types of network structures and representations that have been developed for analyzing multiomics data. We approach our observations from two directions: one that focuses on how nodes are defined in these networks and one that focuses on how edges are defined in these networks.

Table 2 Methods and their categorizations

Method	Supervised / Unsupervised	Deep Learning / Traditional Learning	Reference
iOmicsPASS	Supervised	Traditional Learning	[36]
Integrative Network Fusion	Supervised	Traditional Learning	[37]
iDINGO	Supervised	Traditional Learning	[38]
NetMIM	Supervised	Traditional Learning	[39]
GraphSurv	Supervised	Deep Learning	[40]
LAGProg	Supervised	Deep Learning	[41]
MoGCN	Supervised	Deep Learning	[42]
MOGAT	Supervised	Deep Learning	[43]
MOGONet	Supervised	Deep Learning	[44]
MOGDx	Supervised	Deep Learning	[45]
netOmics	Unsupervised	Traditional Learning	[47]
PARADIGM	Unsupervised	Traditional Learning	[48]
COSMOS	Unsupervised	Traditional Learning	[49]
SUBATOMIC	Unsupervised	Traditional Learning	[50]
NetICS	Unsupervised	Traditional Learning	[51]
Lemon-Tree	Unsupervised	Traditional Learning	[52]
GCN-SC	Unsupervised	Deep Learning	[54]
GLUE	Unsupervised	Deep Learning	[55]
MAE	Unsupervised	Deep Learning	[56]
SpatialGlue	Unsupervised	Deep Learning	[57]
SSGATE	Unsupervised	Deep Learning	[58]

When it comes to nodes, many methods define nodes as molecular entities. For example, some methods create what we call "molecular-omics networks" where each node is a single entity from each multiomics dataset, such as an mRNA, protein, methylation site, or miRNA, among many others. Other methods create "aggregator-omics networks", where each entity represents a harmonized mapping of multiple multiomics types, such as a gene aggregating information from mRNA, miRNA, or methylation probes into a single entity rather than as distinct entities. Many methods also define nodes as individual samples in "sample-based networks", where a sample can depend on the dataset. These include graphs where each node is an individual patient or graphs where each node is an individual cell.

In the context of edges, some methods define their edges based on prior knowledge such as information from KEGG about what genes are in the same pathways or how different multiomics data types interact with each other. Other methods define their edges in an entirely data-driven way by computing correlations or relationships between multiomics features, such as calculating the coexpression of gene expression, calculating correlations between features, using Euclidean distances for spatial data, or calculating metrics based on similarities such as how similar two cells or patients are to each other. Notably, this is often intertwined with the definition of the nodes in the graph for which edges are being defined, with graphs that have individual sample nodes often using similarity-based metrics to define edges.

The diversity of network definitions described here are prevalent across all possible categorizations of models, from traditional machine learning to deep learning approaches as well as supervised to unsupervised approaches. We also attempt to provide a visual overview of everything described in this section in Fig. 2.

Table 3 Methods and their network representations

Method	Algorithm / Analysis Approach	Nodes	Edges	Omics
iOmicsPASS	Modified nearest shrunken centroid classification that favors subnetworks	Molecules or features within each individual omics set	Defined based on correlation of abundance between molecules	mRNA, proteomics from TCGA breast cancer
Integrative Network Fusion	Similarity network fusion + variable juxtaposition followed by random forest or linear SVM	Individual samples/patients	Similarity between patients, computed using individual omics features	Arbitrary, tested gene expression, protein expression and copy number variants from TCGA breast cancer (multiple combinations tested)
iDINGO	Differential network analysis across patient groups	Molecules such as miRNA, mRNA, proteins	Defined based on partial correlations between molecules	Arbitrary, tested miRNA, mRNA, proteins from TCGA breast cancer
NetMIM	Bayesian network model incorporating Dirac spike-and-slab feature selection	Genes or methylation probes (+ other factors such as clinical covariates)	Methylation effects on gene expression (or other factors on gene expression)	Gene expression, methylation from TCGA kidney renal clear cell carcinoma
GraphSurv	Graph convolutional network with Cox proportional hazards prediction output	Genes (omics integrated as node features)	Interactions between genes based on KEGG pathways	mRNA, copy number variation, and methylation from TCGA (33 cancer types)
LAGProg	Conditional variational autoencoder for augmenting features + graph convolutional network with Cox proportional hazards prediction output	Genes (omics integrated as node features through augmentation and original values)	Interactions between genes based on KEGG pathways	mRNA, CNV, and DNA methylation data from TCGA (15 cancer types)
MoGCN	Autoencoder to generate patient-level expression features and then graph convolutional network on a similarity network fusion generated graph	Patients (node features are autoencoder generated features)	Patient similarity edges (fused similarity based on original omics similarities)	Copy number variation, RNA sequencing, and proteomics data from TCGA breast cancer
MOGAT	Graph attention network applied to eight different patient similarity networks to generate embeddings for each network, concatenate, and predict	Patients (node features from each data type)	Patient similarity (8 different sets based on different omics types)	Copy-number variation, coexpression, gene expression, lncRNA, methylation, miRNA, single nucleotide variants (and clinical data) from TCGA breast cancer + METABRIC (some subsets also tested)
MOGONet	Graph convolutional network on each omics network + view correlation discovery network	Patients (node features from each data type)	Patient similarity (for each omics type)	mRNA, methylation, miRNA from ROSMAP (Alzheimer's) + TCGA low-grade glioma, breast cancer, and kidney cancer
MOGDx	Two-layer encoder for dimensionality reduction + graph convolutional network on patient similarity network generated with similarity network fusion	Patients (node features generated by encoder)	Patient similarity (generated by similarity network fusion)	TCGA low-grade glioma, breast cancer, and kidney cancer

Table 3 (continued)

Method	Algorithm / Analysis Approach	Nodes	Edges	Omics
netOmics	Random-walk propagation analysis of multiple sub-omics networks connected together	Various (genes, molecules, proteins)	Calculated interactions (gene interactions), prior knowledge interactions (protein interactions), or relationships between layers (genes and proteins) based on KEGG	mRNA, translation produces, and proteins from HeLa cell cycling dataset + genes, TF-encoding genes, and metabolites from maize aphid dataset
PARADIGM	Parameter estimation using EM algorithm on probabilistic graph model constructed through gene networks (internal and between genes)	Genes (within subnetworks representing pathways) with entities for each of the omics types for each gene	Pathway annotations of how gene omics affect other omics (copy variation -> mRNA, for example) as well as how active products affect other pathways	Copy number variation, gene expression, and DNA methylation from TCGA glioblastoma
COSMOS	Network optimization to minimize mismatches between nodes for causal networks and the size of the selected subgraph	Individual omics data entities (proteins, transcription factors, metabolites)	Prior knowledge based on protein-protein interactions and metabolite-protein interactions	Self-generated transcriptomics, proteomics, and metabolomics from kidney cancer
SUBATOMIC	Hypergraph clustering of submodules determined by subgraph clustering	Molecules or features within each individual omics set	Relationships between omics molecules based on regulatory modules and submodules	Genes and miRNA from hypoxia dataset (with arbitrary extensibility)
NetICS	Network propagation through hierarchical diffusion based on aberrations and differential expression	Genes, miRNA, with different types based on categorization	Interactions representing different biological operations (phosphorylation, expression, etc.)	Somatic mutations with copy number variations, mRNA differential expression, miRNA differential expression from TCGA uterine, liver, bladder, breast, and lung cancers
Lemon-Tree	Model-based clustering and development of gene modules using gene regulatory networks	Genes (from expression data)	Coexpression predictions based on expression data	Gene expression and copy number variation from TCGA glioblastoma
GCN-SC	Spectral graph convolutional network on non-negative matrix factorization reduced features	Individual single cells	Cell pairs matrix determined by mutual nearest neighbors across omics data types	scRNA-seq, CITE-seq, and scATAC-seq (different combinations thereof) from six datasets in humans and mice
GLUE	Graph variational autoencoder paired with standard variational autoencoder	Molecules or features within each individual omics set	Prior knowledge about regulatory interactions between omics features, encoded in a signed and weighted manner	SNARE-seq, SHARE-seq and 10X Multiome (paired), and Nephron and MOp (unpaired)
MAE	Multiple autoencoders with graph-imposed constraints on feature interactions	Molecules or features within each individual omics set	Within each omics, the interactions between omics features (prior knowledge)	Gene expression, miRNA, proteomics, and methylation from TCGA bladder cancer and brain cancer

Table 3 (continued)

Method	Algorithm / Analysis Approach	Nodes	Edges	Omics
SpatialGlue	Graph convolutional network with aggregation within and between modalities	"Spots" from spatial data (based on Euclidean information)	Nearest neighbors between spots in Euclidean space and nearest neighbors between spots based on feature similarities	Spatial transcriptome, epigenome, and proteome from in-house lymph node data (and simulated data)
SSGATE	Graph attention autoencoder for individual modalities with alignment	Individual single cells	Either based on feature similarity (for non-spatial data) or based on spatial distance (for spatial data)	CITE-seq from GEO: BMNC or GEO: SLN111_D1 or Stereo-CITE-seq from GEO: SCS_MT

Usage of algorithms and analysis approaches: observations

Table 2 provides a brief overview of all of the methods discussed in this paper, categorizing them by supervised and unsupervised learning as well as deep learning versus traditional machine learning, and Table 3 provides additional information about the algorithms that are used in each of the methods discussed. While this review is not meant to be a comprehensive overview of every one of the four subcategories assessed, we do note some interesting patterns and trends in what algorithms or analysis approaches are taken within each of the subcategories.

Broadly, traditional machine learning approaches appear to have a more diverse array of approaches that operate on multiomics data as structured as a graph either for preprocessing or for the actual analysis of the graphs. Across the ten papers discussed here that are described as such (across both supervised and unsupervised methods), there are at least six meaningfully distinct approaches that operate on the networks, including but not limited to: a modified nearest shrunken centroid algorithm, differential network analysis, Bayesian probabilistic networks, random-walk network propagation, hypergraph clustering, and parameter estimation, and many of the other methods present their own modifications to multiomics data processing or graph construction that allow for reasonable arguments that they are equally distinct.

In contrast, deep learning approaches appear to have broadly converged onto the usage of graph convolutional networks for supervised approaches and graph-based autoencoders for unsupervised approaches. Five of the six supervised deep learning approaches make use of graph convolutional networks in some capacity (GraphSurv, LAGProg, MoGCN, MOGONet, and MOGDx); the remaining one uses graph attention networks (MOGAT). Three of the five unsupervised deep learning approaches explicitly make use of autoencoders with some imposition of graph structure (GLUE, MAE, SSGATE); the remaining two (GCN-SE and SpatialGlue) do not explicitly describe their methods as autoencoders, but they both make use of reconstruction losses and architecture structures that make such a label reasonable.

There also exist several similarities from a high-level perspective, particularly in how the data is preprocessed for analysis. Many methods within the traditional and deep learning buckets make use of feature selection or dimensionality reduction approaches. For example, some deep learning approaches make use of autoencoders – distinct from the autoencoders that operate on the graph-structured data, these typically operate on individual omics data with no graph structure – to perform learnable dimensionality

reduction in an unsupervised manner; these reduced features are then sometimes used as node features for the graph-based approach. For traditional approaches, dimensionality reduction is done in a variety of ways, including at the graph level, such as selecting subnetworks or submodules for downstream analysis, using feature selection, or using differential expression to select features that are distinct across groups of patients. However, these approaches are not universal, and many methods make use of as many features or nodes (as they are defined) as they can with some degree of interpretability built in.

The centralization of deep learning methods for multiomics integration around graph convolutional networks, especially in the context of the relative wealth of diversity in their traditional machine learning counterparts, highlights a possible area for improvement when it comes to cutting-edge methods for multiomics integration. For example, graph attention networks which have been used by some papers (MOGAT and to some extent SSGATE) can allow for improved parameterization of how the network structure is incorporated by adding learnable parameters to the aggregation of neighbor information.

Observations on the landscape of the broader field

Table 2 provides a brief overview of all of the methods discussed in this paper, categorizing them by supervised and unsupervised learning as well as deep learning versus traditional machine learning. These methods demonstrate the variety of machine learning network approaches that can be applied to problems in multiomics association analysis. Generally, the methods discussed face the same issues that all multiomics analysis methods must address. The most prominent in this review were the curse of dimensionality (having far more features than samples in many cases) [68, 69] and the issue of addressing noise and sparsity in some datasets [15, 70]. Network-based approaches in both the traditional machine learning and deep learning regimes provide rigorous solutions to this problem, as discussed above by the myriad number of methods that inherently perform dimensionality reduction or naturally relate concepts to each other.

For example, in methods that deal with patient or individual-level data, issues relating to high dimensionality and noise due to loss of information can be quite important due to the inherent aggregation of omics data in such datasets (for example, bulk RNA sequencing often leads to a transcriptomic profile that is, in essence, a weighted sum of the transcriptomic profiles of the cells that are in the bulk data that is sequenced). Using networks can address this concern by allowing one to integrate the data in a constrained way in a variety of ways, such as through the integration of prior knowledge such as knowledge about what pathways or gene modules are impacted by specific genes to create an integrated pathway/gene module representation that helps mitigate some of the loss of information and dimensionality through informed aggregation or selection in an informed way.

In the case of methods that operate on single-cell-level data, high dimensionality remains an issue, but high amounts of sparsity due to the data being collected at cellular-level resolution can provide an additional challenge in generating useful representations or learning information about individual features in a meaningful way. Network-based approaches can resolve this issue in a similar way - by providing an informed way to relate sparse elements to one another, they can provide a way to induce information or

a useful informed/constrained representation based on only a subset of features due to sparsity.

Furthermore, it is important to recognize that there exist issues of reproducibility and data leakage that must be considered in all machine learning analyses. While this is not unique to multiomics analysis or network analyses, it is important to avoid such pitfalls when designing one's own analysis; we direct readers to the relevant paper by Kapoor and Narayanan to learn more [71].

It is also worth repeating that we did not include semi-supervised learning or reinforcement learning in this review, despite their relative prominence in machine learning as a whole, as these are fields that have not seen much use in the context of network analyses for multiomics data. However, we do want to point the reader to work by Wang et al. in MOSEGCN, which is a graph convolutional network-based approach that uses semi-supervised learning, if they are interested in reading more about how such an approach can be developed [72]. Semi-supervised learning offers a promising opportunity for combining labeled and unlabeled data in the multiomics paradigm where labeled data may be sparse but unlabeled data may be abundant; as such, the lack of usage of semi-supervised learning approaches represents a gap that can be further explored in network-based multiomics data analysis.

Discussion and future directions for the field

Based on our perusal of the field, we observed many limitations and gaps in existing methods that could serve as future directions or areas of possible improvement in the future.

One major gap is in the apparent, emerging consolidation of methods around particular algorithms or approaches, especially in the domain of deep learning approaches for multiomics integration, where graph convolutional networks and layers appear to dominate. While many methods appear able to improve on performance with graph convolutional networks alone, there are many advances in the field of graph deep learning that could lead to even more improvement. For example, graph transformers and graph attention networks are alternative architectures to graph convolutional networks that have shown their own benefits with novel ways of aggregating information that may improve performance [73, 74]. Additionally, heterogeneous graph neural networks or more elaborate methods for multi-view networks that operate on heterogeneous graphs where there are more than one type of node or edge (much like the molecular-omics networks or networks where edges can carry one of multiple definitions) have also been seeing significant improvements, but despite those improvements, such methods appear to be underutilized in the field of multiomics integration [75, 76, 77, 78, 79].

Another key gap is in interpretability and in reliance on prior knowledge as a ground truth. These often lead to algorithms having limited clinical relevance due to the increased rigor and expectations that exist for clinically-adjacent methods that may impact patient care, and they can also limit the research applications by limiting the capacity for hypothesis generation and testing through limited understanding or rapidly becoming outdated as biological knowledge progresses. Future considerations for network representation and analyses of multiomics data may include incorporating advances in graph deep learning interpretability and consciously designing methods to be robust to changing knowledge to mitigate many of these issues. This will allow future

methods to push the boundaries of accuracy and value without having to compromise on important components of any machine learning model or data representation, especially those that have the potential to be clinically relevant.

Furthermore, there exists a possibility for integrating both deep learning and traditional machine learning methods, particularly for applications where end-to-end deep learning training may be overkill or prone to overfitting. For example, one could train unsupervised graph autoencoders with much larger, unlabeled datasets to generate lower-dimensional representations of data, and then use those unsupervised representations with traditional network-based algorithms such as diffusion or random-walk propagation for classification or other downstream tasks. This would allow one to take advantage of the generally higher representative power of graph neural networks while mitigating the downside of their lower effectiveness for small-sample-size applications.

Conclusion

Network analysis in multiomics is an established, but rapidly growing field that has many benefits over non-network analyses in the realms of interpretability, sparsity, and dimensionality. Recent advances in deep learning, particularly the development of graph neural networks, as well as in traditional machine learning methods that operate on networks has led to an increased number of papers making use of these approaches in their analyses. A recent coalescence can be seen between supervised and unsupervised approaches, with many papers making use of similar underlying approaches to integrate network information, representing an exciting opportunity for improvements within both domains by incorporating advances in the other.

We provide in this review an overview of many of these recent approaches while making focused observations about the tradeoffs between traditional machine learning and deep learning approaches as well as the distinction between supervised and unsupervised approaches to help researchers in the decision-making process when it comes to their decisions about how to analyze their own data. We use this framework and these observations to propose a variety of possible changes that may lead to improvements in the field of multiomics integration as a whole, particularly for network-based multiomics integration.

Abbreviations

TCGA	The Cancer Genome Atlas
KEGG	Kyoto Encyclopedia of Genes of Genomes
CPTAC	Clinical Proteomic Tumor Analysis Consortium
NSC	Nearest shrunken centroid
SNF	Similarity network fusion
juXT	Variable juxtaposition
BRCA	Breast Invasive Carcinoma (a TCGA dataset)
VCDN	View-correlation discovery network
COSMOS	Causal Oriented Search of Multi-Omics Space
SUBATOMIC	Subgraph Based Multiomics Clustering Framework
NetICS	Network-based Integration of Multi-omics Data
GLUE	Graph-linked unified embedding
MAE	Multi-view Factorization Autoencoder

Acknowledgements
Not applicable.

Author contributions
RK and MDR conceptualized the idea for this review and the framework thereof. RK and MDR drafted and revised the paper. JDR contributed substantially to the revisions of this paper and to the final draft. All authors read and approved the final manuscript.

Funding
RK was partially supported by the Training Program in Computational Genomics grant from the National Human Genome Research Institute to the University of Pennsylvania (T32HG000046). JDR was supported by the National Library of Medicine of the National Institutes of Health (R00LM013646). MDR was supported by grants U01AG066833, R01AI077505, UL1TR001878, and P30AG073105.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2]Medical Scientist Training Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3]Division of Informatics, Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[5]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Received: 26 August 2024 / Accepted: 10 May 2025
Published online: 27 May 2025

References

1. Sun YV, Hu YJ. Integrative analysis of Multi-omics data for discovery and functional studies of complex human diseases. Adv Genet. 2016;93:147–90. https://doi.org/10.1016/bs.adgen.2015.11.004.
2. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83. https://doi.org/10.1186/s13059-017-1215-1.
3. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinforma Biol Insights. 2020;14:1177932219899051. https://doi.org/10.1177/1177932219899051.
4. Roy S, Radivojevic T, Forrer M, et al. Multiomics data collection, visualization, and utilization for guiding metabolic engineering. Front Bioeng Biotechnol. 2021;9:612893. https://doi.org/10.3389/fbioe.2021.612893.
5. Silva TC, Colaprico A, Olsen C, et al. TCGA workflow: analyze cancer genomics and epigenomics data using bioconductor packages. F1000Research. 2016;5:1542. https://doi.org/10.12688/f1000research.8923.2.
6. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol. 2015;19(1A):A68–77. https://doi.org/10.5114/wo.2014.47136.
7. Menden K, Francescatto M, Nyima T, et al. A multi-omics dataset for the analysis of frontotemporal dementia genetic subtypes. Sci Data. 2023;10(1):849. https://doi.org/10.1038/s41597-023-02598-x.
8. Reddy JS, Heath L, Linden AV, et al. Bridging the Gap: Multi-omics profiling of brain tissue in Alzheimer's disease and older controls in multi-ethnic populations. Alzheimers Dement J Alzheimers Assoc. 2024;20(10):7174–92. https://doi.org/10.1002/alz.14208.
9. Greenwood AK, Montgomery KS, Kauer N, et al. The AD knowledge portal: A repository for Multi-Omic data on Alzheimer's disease and aging. Curr Protoc Hum Genet. 2020;108(1):e105. https://doi.org/10.1002/cphg.105.
10. Muto Y, Yoshimura Y, Wu H, et al. Multiomics profiling of mouse polycystic kidney disease progression at a single-cell resolution. Proc Natl Acad Sci. 2024;121(43):e2410830121. https://doi.org/10.1073/pnas.2410830121.
11. Li H, Li D, Ledru N, et al. Transcriptomic, epigenomic, and Spatial metabolomic cell profiling redefines regional human kidney anatomy. Cell Metab. 2024;36(5):1105–e112510. https://doi.org/10.1016/j.cmet.2024.02.015.
12. Fernandes M, Husi H. Establishment of a integrative multi-omics expression database CKDdb in the context of chronic kidney disease (CKD). Sci Rep. 2017;7(1):40367. https://doi.org/10.1038/srep40367.
13. Li Z, Zhang H, Li Q, et al. GepLiver: an integrative liver expression atlas spanning developmental stages and liver disease phases. Sci Data. 2023;10(1):376. https://doi.org/10.1038/s41597-023-02257-1.
14. Sveinbjornsson G, Ulfarsson MO, Thorolfsdottir RB, et al. Multiomics study of nonalcoholic fatty liver disease. Nat Genet. 2022;54(11):1652-63. https://doi.org/10.1038/s41588-022-01199-5.
15. Flores JE, Claborne DM, Weller ZD, Webb-Robertson BJM, Waters KM, Bramer LM. Missing data in multi-omics integration: recent advances through artificial intelligence. Front Artif Intell. 2023;6:1098308. https://doi.org/10.3389/frai.2023.1098308.
16. Krassowski M, Das V, Sahu SK, Misra BB. State of the field in Multi-Omics research: from computational needs to data mining and sharing. Front Genet. 2020;11:610798. https://doi.org/10.3389/fgene.2020.610798.
17. Odenkirk MT, Reif DM, Baker ES. Multiomic big data analysis challenges: increasing confidence in the interpretation of artificial intelligence assessments. Anal Chem. 2021;93(22):7763–73. https://doi.org/10.1021/acs.analchem.0c04850.
18. Strogatz SH. Exploring complex networks. Nature. 2001;410(6825):268–76. https://doi.org/10.1038/35065725.
19. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68. https://doi.org/10.1038/nrg2918.

20. Hänsel K, Dudgeon SN, Cheung KH, Durant TJS, Schulz WL. From data to wisdom: biomedical knowledge graphs for Real-World data insights. J Med Syst. 2023;47(1):65. https://doi.org/10.1007/s10916-023-01951-2.
21. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J. 2020;18:1414–28. https://doi.org/10.1016/j.csbj.2020.05.017.
22. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from Multi-Omics data. Front Genet. 2019;10. https://doi.org/10.3389/fgene.2019.00535.
23. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16(2):85–97. https://doi.org/10.1038/nrg3868.
24. Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. Nat Biomed Eng. 2022;6(12):1353–69. https://doi.org/10.1038/s41551-022-00942-x.
25. Bechtel W. Hierarchy and levels: analysing networks to study mechanisms in molecular biology. Philos Trans R Soc Lond B Biol Sci. 2020;375(1796):20190320. https://doi.org/10.1098/rstb.2019.0320.
26. Merico D, Gfeller D, Bader GD. How to visually interpret biological data using networks. Nat Biotechnol. 2009;27(10):921–4. https://doi.org/10.1038/nbt.1567.
27. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61. https://doi.org/10.1093/nar/gkw1092.
28. Jovel J, Greiner R. An introduction to machine learning approaches for biomedical research. Front Med. 2021;8. https://doi.org/10.3389/fmed.2021.771607.
29. Agamah FE, Bayjanov JR, Niehues A et al. Computational approaches for network-based integrative multi-omics analysis. *Front Mol Biosci*. 2022;9. Accessed April 19, 2023. https://www.frontiersin.org/articles/https://doi.org/10.3389/fmolb.2022.967205
30. Zhou G, Li S, Xia J. Network-Based approaches for Multi-omics integration. Methods Mol Biol Clifton NJ. 2020;2104:469–87. https://doi.org/10.1007/978-1-0716-0239-3_23.
31. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform. 2022;23(1):bbab454. https://doi.org/10.1093/bib/bbab454.
32. Arjmand B, Hamidpour SK, Tayanloo-Beik A, et al. Machine learning: A new prospect in Multi-Omics data analysis of Cancer. Front Genet. 2022;13:824451. https://doi.org/10.3389/fgene.2022.824451.
33. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnol Adv. 2021;49:107739. https://doi.org/10.1016/j.biotechadv.2021.107739.
34. Moore JH, Boland MR, Camara PG, et al. Preparing next-generation scientists for biomedical big data: artificial intelligence approaches. Pers Med. 2019;16(3):247-57. https://doi.org/10.2217/pme-2018-0145.
35. Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience. 2022;25(2):103798. https://doi.org/10.1016/j.isci.2022.103798.
36. Koh HWL, Fermin D, Vogel C, Choi KP, Ewing RM, Choi H. iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. Npj Syst Biol Appl. 2019;5(1):1-10. https://doi.org/10.1038/s41540-019-0099-y.
37. Chierici M, Bussola N, Marcolini A et al. Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling. *Front Oncol*. 2020;10. Accessed April 19, 2023. https://www.frontiersin.org/articles/https://doi.org/10.3389/fonc.2020.01065
38. Class CA, Ha MJ, Baladandayuthapani V, Do KA. iDINGO—integrative differential network analysis in genomics with Shiny application. Bioinformatics. 2018;34(7):1243-5. https://doi.org/10.1093/bioinformatics/btx750.
39. Zhu B, Zhang Z, Leung SY, Fan X. NetMIM: network-based multi-omics integration with block missingness for biomarker selection and disease outcome prediction. Brief Bioinform. 2024;25(5):bbae454. https://doi.org/10.1093/bib/bbae454.
40. Wang Y, Zhang Z, Chai H, Yang Y. Multi-omics Cancer Prognosis Analysis Based on Graph Convolution Network. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.; 2021:1564-1568. https://doi.org/10.1109/BIBM52615.2021.9669797
41. Zhang Y, Xiong S, Wang Z, et al. Local augmented graph neural network for multi-omics cancer prognosis prediction and analysis. Methods. 2023;213:1-9. https://doi.org/10.1016/j.ymeth.2023.02.011.
42. Li X, Ma J, Leng L, et al. MoGCN: A Multi-Omics integration method based on graph convolutional network for Cancer subtype analysis. Front Genet. 2022;13:806842. https://doi.org/10.3389/fgene.2022.806842.
43. Tanvir RB, Islam MM, Sobhan M, Luo D, Mondal AM. MOGAT: an improved Multi-Omics integration framework using graph attention networks. Published online April 2, 2023:2023.04.01.535195. https://doi.org/10.1101/2023.04.01.535195
44. Wang T, Shao W, Huang Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12(1):3445. https://doi.org/10.1038/s41467-021-23774-w.
45. Ryan B, Marioni RE, Simpson TI. Multi-Omic graph diagnosis (MOGDx): a data integration tool to perform classification tasks for heterogeneous diseases. Bioinformatics. 2024;40(9):btae523. https://doi.org/10.1093/bioinformatics/btae523.
46. Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. Nat Rev Mol Cell Biol. 2023;24(10):695-713. https://doi.org/10.1038/s41580-023-00615-w.
47. Bodein A, Scott-Boyer MP, Perin O, Lê Cao KA, Droit A. Interpretation of network-based integration from multi-omics longitudinal data. Nucleic Acids Res. 2021;50(5):e27. https://doi.org/10.1093/nar/gkab1200.
48. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinforma Oxf Engl. 2010;26(12):i237-245. https://doi.org/10.1093/bioinformatics/btq182.
49. Dugourd A, Kuppe C, Sciacovelli M, et al. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol Syst Biol. 2021;17(1):e9730. https://doi.org/10.15252/msb.20209730.
50. Loers JU, Vermeirssen V. SUBATOMIC: a subgraph based mulTi-OMIcs clustering framework to analyze integrated multi-edge networks. BMC Bioinformatics. 2022;23(1):363. https://doi.org/10.1186/s12859-022-04908-3.
51. Dimitrakopoulos C, Hindupur SK, Häfliger L, et al. Network-based integration of multi-omics data for prioritizing cancer genes. Bioinforma Oxf Engl. 2018;34(14):2441-8. https://doi.org/10.1093/bioinformatics/bty148.
52. Bonnet E, Calzone L, Michoel T. Integrative multi-omics module network inference with Lemon-Tree. PLoS Comput Biol. 2015;11(2):e1003983. https://doi.org/10.1371/journal.pcbi.1003983.
53. Gene Ontology Consortium, Aleksander SA, Balhoff J, et al. The gene ontology knowledgebase in 2023. Genetics. 2023;224(1):iyad031. https://doi.org/10.1093/genetics/iyad031.

54. Gao H, Zhang B, Liu L, Li S, Gao X, Yu B. A universal framework for single-cell multi-omics data integration with graph convolutional networks. Brief Bioinform Published Online March. 2023;17:bbad081. https://doi.org/10.1093/bib/bbad081.
55. Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol. 2022;40(10):1458-66. https://doi.org/10.1038/s41587-022-01284-4.
56. Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using Multi-view factorization autoencoder (MAE). BMC Genomics. 2019;20(11):944. https://doi.org/10.1186/s12864-019-6285-x.
57. Long Y, Ang KS, Sethi R, et al. Deciphering Spatial domains from Spatial multi-omics with spatialglue. Nat Methods. 2024;21(9):1658-67. https://doi.org/10.1038/s41592-024-02316-4.
58. Lv T, Zhang Y, Liu J, Kang Q, Liu L. Multi-omics integration for both single-cell and spatially resolved data based on dual-path graph attention auto-encoder. Brief Bioinform. 2024;25(5):bbae450. https://doi.org/10.1093/bib/bbae450.
59. Teng Q, Liu Z, Song Y, Han K, Lu Y. A survey on the interpretability of deep learning in medical diagnosis. Multimed Syst. 2022;28(6):2335-55. https://doi.org/10.1007/s00530-022-00960-4.
60. Linardatos P, Papastefanopoulos V, Kotsiantis S, Explainable AI. A review of machine learning interpretability methods. Entropy. 2020;23(1):18. https://doi.org/10.3390/e23010018.
61. Sarker IH. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. Sn Comput Sci. 2021;2(6):420. https://doi.org/10.1007/s42979-021-00815-1.
62. Demšar J, Zupan B. Hands-on training about overfitting. PLoS Comput Biol. 2021;17(3):e1008671. https://doi.org/10.1371/journal.pcbi.1008671.
63. López OAM, López AM, Crossa DJ, Overfitting. Model tuning, and evaluation of prediction performance. Multivariate statistical machine learning methods for genomic prediction [Internet]. Springer; 2022. https://doi.org/10.1007/978-3-030-89010-0_4.
64. Ying X. An overview of overfitting and its solutions. J Phys Conf Ser. 2019;1168(2):022022. https://doi.org/10.1088/1742-6596/1168/2/022022.
65. Kokhlikyan N, Miglani V, Martin M et al. Captum: A unified and generic model interpretability library for pytorch. Published online 2020.
66. Luo D, Cheng W, Xu D et al. Parameterized Explainer for Graph Neural Network. Published online November 9, 2020. Accessed March 28, 2023. http://arxiv.org/abs/2011.04573
67. Ying Z, Bourgeois D, You J, Zitnik M, Leskovec J, GNNExplainer. Generating Explanations for Graph Neural Networks. In: *Advances in Neural Information Processing Systems*. Vol 32. Curran Associates, Inc.; 2019. Accessed March 28, 2023. https://proceedings.neurips.cc/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html
68. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735-46. https://doi.org/10.1016/j.csbj.2021.06.030.
69. Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. Machine learning and integrative analysis of biomedical big data. Genes. 2019;10(2):87. https://doi.org/10.3390/genes10020087.
70. Wissel D, Rowson D, Boeva V. Systematic comparison of multi-omics survival models reveals a widespread lack of noise resistance. Cell Rep Methods. 2023;3(4). https://doi.org/10.1016/j.crmeth.2023.100461.
71. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. https://doi.org/10.1016/j.patter.2023.100804.
72. Wang J, Liao N, Du X, Chen Q, Wei B. A semi-supervised approach for the integration of multi-omics data based on transformer multi-head self-attention mechanism and graph convolutional networks. BMC Genomics. 2024;25(1):1-12. https://doi.org/10.1186/s12864-024-09985-7.
73. Shehzad A, Xia F, Abid S, et al. Graph Transformers: A survey. Published Online July. 2024;13. https://doi.org/10.48550/arXiv.2407.09777.
74. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Atten Networks Published Online Febr. 2018;4. https://doi.org/10.48550/arXiv.1710.10903.
75. Ata SK, Fang Y, Wu M, Shi J, Kwoh CK, Li X. Multi-View collaborative network embedding. ACM Trans Knowl Discov Data. 2021;15(3):39:1-39. https://doi.org/10.1145/3441450.
76. Bing R, Yuan G, Zhu M, Meng F, Ma H, Qiao S. Heterogeneous graph neural networks analysis: a survey of techniques, evaluations and applications. Artif Intell Rev. 2023;56(8):8003-42. https://doi.org/10.1007/s10462-022-10375-2.
77. Li Q, Chen W, Fang Z, Ying C, Wang C. A multi-view contrastive learning for heterogeneous network embedding. Sci Rep. 2023;13(1):6732. https://doi.org/10.1038/s41598-023-33324-7.
78. Wang X, Bo D, Shi C, Fan S, Ye Y, Yu PS. A survey on heterogeneous graph embedding: methods, techniques, applications and sources. Published Online November. 2020;30. https://doi.org/10.48550/arXiv.2011.14867.
79. Shang Y, Ye X, Sakurai T. Multi-view Network Embedding with Structure and Semantic Contrastive Learning. In: *2023 IEEE International Conference on Multimedia and Expo (ICME).*; 2023:870-875. https://doi.org/10.1109/ICME55011.2023.00154

Publisher's note