Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 135-152

Unraveling Biological Complexity: AI and Statistical Approaches to Multi-Omics Data Integration

D. C. Mishra[1], Shesh Nath Rai[2], Mamatha Y. S.[1], K. K. Chaturvedi[1], Sudhir Srivastava[1], Neeraj Budhlakoti[1] and Girish Kumar Jha[1]

[1]*Division of Agricultural Bioinformatics*
ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012
[2]*Cancer Data Science Center*
University of Cincinnati, College of Medicine, Cincinnati, OH, USA.

Received: 16 June 2025; Revised: 14 August 2025; Accepted: 17 August 2025

Abstract

In the era of precision medicine, understanding the intricate biological mechanisms underlying diseases requires a comprehensive analysis of multi-omics data, including genomics, transcriptomics, proteomics and metabolomics. The sheer volume and complexity of these datasets present significant challenges in deciphering the interactions and regulatory networks that govern cellular functions. This paper will explore how cutting-edge artificial intelligence (AI) and statistical methodologies, including deep learning approaches like Variational Autoencoder (VAE) and Graph Neural Networks (GNNs), are transforming the integration of multi-omics data, enabling new insights into biological complexity. We will discuss advanced statistical models, such as Bayesian Networks, Canonical Correlation Analysis (CCA) and Multi-Omics Factor Analysis (MOFA), that facilitate the integration of diverse data types, revealing deeper layers of biological information that are often obscured in traditional analyses. From identifying biomarkers for early disease detection to uncovering therapeutic targets, the combination of AI, deep learning and statistical approaches holds great promise in advancing our understanding of health and disease.

Key words: Multiomics; Data integration; MOFA; Deep learning; Network based approach.

AMS Subject Classifications: 62K05, 05B05.

1. Introduction

The central dogma of molecular biology, which describes the flow of genetic information from DNA to RNA to protein, has long served as a cornerstone of biological understanding. However, a comprehensive understanding of biological systems requires the integration of data from multiple 'omics' layers. Genomics, transcriptomics, proteomics and

Corresponding Author: D.C. MISHRA
Email: dwij.mishra@gmail.com

metabolomics each offer a unique perspective, capturing different aspects of cellular function and regulation. The advent of high-throughput technologies, such as next-generation sequencing and mass spectrometry, has led to an explosion of omics data, creating both opportunities and challenges for systems biology, see Misra (2018).

While each omics layer provides valuable information, studying them in isolation offers an incomplete and potentially misleading picture. For instance, changes in mRNA transcript levels do not always directly correlate with corresponding protein abundances due to post-transcriptional regulation, protein turnover and other factors. Multi-omics integration seeks to address these limitations by combining data from multiple sources to provide a more holistic and accurate representation of biological systems, see Subramanian *et al.* (2020).

In this paper, we explore a range of statistical and AI-based methods for multi-omics data integration, with a focus on Canonical correlation analysis, Network modeling, Bayesian inference and Deep learning strategies like Variational autoencoders. We review existing tools such as mixOmics, RGCCA, and PINSPlus, which leverage these methods for practical applications in agricultural and biomedical research.

2. Statistical approaches to multi-omics data integration

Statistical methods play a crucial role in managing the high-dimensional, heterogeneous nature of multi-omics data. Several widely used methods for integrating multi-omics data are given below, see Naserkheil *et al.* (2022).

2.1. Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis is a statistical method designed to identify and quantify the linear relationships between two multidimensional datasets. In the context of multi-omics data integration, CCA helps in discovering correlated patterns across different omics layers—such as transcriptomics and proteomics—thus uncovering shared biological signals, see Wróbel *et al.* (2024).

Let $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$ be two centered datasets representing two omics layers, where n is the number of samples, and p and q are the number of variables in each omics type. CCA seeks linear combinations of the variables in each dataset such that the correlation between these combinations is maximized. We aim to find vectors $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ such that the correlation between the canonical variates $X\mathbf{a}$ and $Y\mathbf{b}$ is maximized:

$$\max_{\mathbf{a},\mathbf{b}} \rho = \frac{\mathbf{a^T C_{XY} b}}{\sqrt{\mathbf{a^T C_{XX} a}}\sqrt{\mathbf{b^T C_{YY} b}}} \tag{1}$$

where:

- $\mathbf{C}_{XX} = \frac{1}{n-1}X^TX$ is the covariance matrix of $\mathbf{X}$.
- $\mathbf{C}_{YY} = \frac{1}{n-1}Y^TY$ is the covariance matrix of $\mathbf{Y}$.
- $\mathbf{C}_{XY} = \frac{1}{n-1}X^TY$ is the cross covariance matrix of $\mathbf{XY}$.

This leads to the generalized eigenvalue problem:

$$\mathbf{C}_{XY}\,\mathbf{C}_{YY}^{-1}\,\mathbf{C}_{YX}\,\mathbf{a}=\lambda\,\mathbf{C}_{XX}\,\mathbf{a}$$

$$\mathbf{C}_{YX}\mathbf{C}_{XX}^{-1}\mathbf{C}_{XY}\mathbf{b}=\lambda\mathbf{C}_{YY}\mathbf{b}$$

The first pair $(\mathbf{a}_1, \mathbf{b}_1)$ gives the directions of maximal correlation. Subsequent canonical directions are obtained by enforcing orthogonality constraints with previous variates.

In high-dimensional multi-omics data (where p or q is much larger than n), classical CCA may become ill-posed. In such cases, regularized or sparse variants are used.

2.1.1. Regularized CCA

Regularized CCA adds penalties to the denominator to stabilize the solution, see Parkhomenko *et al.* (2009).

$$\max_{\mathbf{a},\mathbf{b}} \rho=\frac{\mathbf{a}^T\mathbf{C}_{XY}\mathbf{b}}{\sqrt{\mathbf{a}^T(\mathbf{C}_{XX}+\kappa_x\mathbf{I})\mathbf{a}}\,\sqrt{\mathbf{b}^T(\mathbf{C}_{YY}+\kappa_y\mathbf{I})\mathbf{b}}} \tag{2}$$

where κ_x and κ_y are regularization parameters.

2.1.2. Sparse CCA (sCCA)

Sparse CCA (sCCA) imposes sparsity constraints on $\mathbf{a}$ and $\mathbf{b}$, leading to feature selection and interpretability:

$$\begin{aligned} \max_{\mathbf{a},\mathbf{b}} \quad & \mathbf{a}^T\mathbf{C}_{XY}\mathbf{b} \\ \text{subject to} \quad & \|\mathbf{a}\|_2 \le 1, \quad \|\mathbf{b}\|_2 \le 1 \\ & \|\mathbf{a}\|_1 \le c_1, \quad \|\mathbf{b}\|_1 \le c_2 \end{aligned} \tag{3}$$

These constraints $\|\cdot\|_1$ enforce sparsity, making sCCA particularly useful in the context of omics data where many variables are irrelevant or noisy, see Witten and Tibshirani (2009).

2.1.3. Advantages and limitations

CCA is a powerful tool for identifying relationships between multi-omics datasets. It can handle high-dimensional data and identify complex dependencies. However, CCA is sensitive to outliers and assumes a linear relationship between the variables. In cases where the relationship is non-linear, other methods, such as kernel CCA, may be more appropriate.

2.1.4. Tools implementing CCA for multi-omics data integration

a. mixOmics R package with multivariate methods (including CCA) for exploring and integrating omics datasets, see Rohart *et al.* (2017).

b. RGCCA R package offering generalized CCA for integrating multiple datasets.

c. BLOCCS R package for Block Sparse CCA, estimating multiple canonical directions for enhanced interpretability.

2.2. Similarity-based approaches

Similarity-based approaches represent a powerful class of methods in multi-omics data integration. These methods focus on quantifying the similarity or distance between samples within each omics layer and then combining these relationships to gain a unified understanding of biological patterns, such as disease subtypes, cellular states, or treatment responses.

Unlike direct feature-level integration, which merges raw data matrices, similarity-based methods operate by first computing sample-sample similarity matrices independently for each omics type (*e.g.*, transcriptomics, proteomics, metabolomics). These matrices reflect the relationship between samples based on their respective omics profiles.

Let us consider K different omics datasets $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}\}$, each with n samples and their respective similarity matrices $\{\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(K)}\}$, where each $\mathbf{S}^{(k)} \in \mathbb{R}^{n \times n}$.

The key idea is to integrate these K similarity matrices into a single consensus matrix $\mathbf{S}_{\text{integrated}}$, which captures the shared structure across all data types.

2.2.1. Similarity Network Fusion (SNF)

One of the most popular similarity-based methods is Similarity Network Fusion, which iteratively updates each similarity matrix using neighborhood information from other omics layers, see Wang *et al.* (2014).

The SNF algorithm involves the following steps:

1. **Compute sample similarity matrices $\mathbf{S}^{(k)}$** for each omics data type using a distance metric (*e.g.*, Euclidean distance or Gaussian kernel similarity)
2. **Normalize** the matrices to maintain comparability.
3. **Iteratively update** each matrix by combining it with others through a message-passing mechanism:

$$\mathbf{W}_{t+1}^{(k)} = \alpha \mathbf{P}^{(k)} \cdot \left(\frac{1}{K-1} \sum_{l \neq k} \mathbf{W}_t^{(l)} \right) \mathbf{P}^{(k)T} + (1-\alpha) \mathbf{W}_t^{(k)} \tag{4}$$

   where $\mathbf{P}^{(k)}$ is the transition probability matrix of $\mathbf{S}^{(k)}$, and α is a regularization parameter (typically 0.5).

4. **Fuse the final networks** after convergence:

$$\mathbf{S}_{\text{integrated}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{W}_T^{(k)} \tag{5}$$

The resulting integrated similarity matrix is then used for downstream tasks such as spectral clustering, dimensionality reduction, or classification.

2.2.2. Other tools and methods

a. PINSPlus An extension of perturbation clustering that performs multiple clustering runs on each omics dataset and integrates the results using co-clustering frequencies.

b. NEMO (Neighborhood-based multi-omics clustering) Designed for partial datasets with missing omics layers, it builds local sample neighborhoods and combines them across modalities.

c. iClusterPlus Although fundamentally a latent variable model, it also aligns sample similarities and can be categorized under similarity-based frameworks.

2.2.3. Advantages and limitations

Similarity based integration methods offer several advantages and challenges. Among the advantages, they are robust to missing data, as similarity matrices can still be computed even when some features are absent. They also allow flexible integration, effectively handling heterogeneous omics types without requiring normalization across different data scales. Additionally, these methods enhance interpretability by providing integrated similarity networks that visually and intuitively represent relationships among samples. However, there are notable challenges as well. The choice of similarity metric is critical, as different distance measures can produce significantly different outcomes. Computational complexity is another concern, especially with large datasets, as calculating pairwise similarities can be both memory and time-intensive. Lastly, parameter tuning is essential for algorithms like Similarity Network Fusion, which rely on parameters such as the number of neighbors and kernel width, requiring careful adjustment to ensure reliable results.

2.3. Bayesian models

Bayesian models offer a powerful and principled framework for multi-omics data integration by treating uncertainty explicitly and allowing incorporation of prior biological knowledge. These models are particularly useful in handling heterogeneous, high-dimensional and often noisy datasets typical in multi-omics studies, such as genomics, transcriptomics, epigenomics and proteomics, see Kirk *et al.* (2012).

2.3.1. Bayesian clustering models

These models assign samples to latent clusters using probability distributions, rather than hard assignments. A popular non parametric Bayesian clustering method is the Dirichlet Process Mixture Model (DPMM).

In multi-omics integration, each omics dataset contributes to the clustering through its own likelihood component. For instance, assuming omics data $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}$ share

a common clustering structure **Z**:

$$P\left(Z, \theta^{(1)}, \ldots, \theta^{(K)} \mid \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}\right) \propto P(Z) \prod_{k=1}^{K} P\left(\mathbf{X}^{(k)} \mid Z, \theta^{(k)}\right) P\left(\theta^{(k)}\right) \tag{6}$$

where: θ^k is cluster specific parameter.

Tools and methods include:

a. MDI (Multiple Dataset Integration) A joint Bayesian model that performs clustering on multiple omics layers and identifies consensus clusters.

b. BCC (Bayesian Consensus Clustering) Estimates shared cluster structure while allowing for data-specific variations.

c. LRAcluster (Low-Rank Approximation Clustering) Incorporates low-rank approximations to simplify the Bayesian model for high-dimensional omics data.

2.3.2. Bayesian networks

Bayesian networks are graphical models that represent conditional dependencies among random variables. In multi-omics integration, they are used to model causal relationships between genes, proteins, and metabolites.

A Bayesian network is a directed acyclic graph (DAG), where nodes represent variables (*e.g.*, gene expression, protein levels), and edges encode conditional dependencies. The joint distribution is factorized as:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \text{Parents}(X_i)\right) \tag{7}$$

This formulation enables modeling of regulatory pathways or signaling cascades across omics layers. Examples:

a. PARADIGM (Pathway Recognition Algorithm using Data Integration on Genomic Models) Integrates copy number and gene expression data to infer pathway activity, see Vaske *et al.* (2021).

b. CONEXIC (COpy Number and EXpression In Cancer) Uses Bayesian networks to identify driver genes by integrating copy number alterations and expression profiles, see Akavia *et al.* (2010).

2.3.3. Advantages and limitations

Bayesian models offer several compelling advantages and face notable challenges. On the positive side, they excel at uncertainty modeling by providing full posterior distributions, which yield credible intervals and enhance confidence in predictions. They also allow the incorporation of prior knowledge, such as known biological pathways or disease associations,

directly into the model. Thanks to modern techniques like variational inference and Markov Chain Monte Carlo (MCMC) sampling, Bayesian methods have become scalable to large datasets. Additionally, they handle missing data naturally as part of the inference process, eliminating the need for imputation. However, these benefits come with challenges. Bayesian inference can be computationally expensive, particularly when dealing with multiple omics layers or a high number of variables. The complexity of designing and validating hierarchical models or directed acyclic graphs (DAGs) demands significant expertise and domain knowledge. Moreover, the results can be sensitive to the choice of priors—poorly chosen or inadequate priors may bias outcomes or impede model convergence.

2.4. Multivariate methods

Multivariate methods are essential tools in multi-omics data integration, offering the capability to jointly analyze multiple variables from different omics layers. Unlike univariate methods that treat each variable independently, multivariate approaches capture correlations, co-variations, and shared structures across datasets, making them ideal for discovering hidden biological relationships and reducing dimensionality in high-throughput omics data. These methods are particularly valuable when integrating datasets from genomics, transcriptomics, proteomics, metabolomics, and other omics types, where the number of variables far exceeds the number of observations, and variables often interact in complex, non-linear ways.

2.4.1. Principal Component Analysis (PCA)

PCA is one of the most widely used unsupervised multivariate techniques for dimensionality reduction. It identifies orthogonal directions (principal components) that capture the maximum variance in the data. When applied to multi-omics datasets either jointly or separately, PCA can reveal dominant variation patterns, batch effects, and clustering structures, see Jolliffe and Cadima (2016).

Given a centered data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, PCA solves the eigenvalue problem:

$$\mathbf{X}^T\mathbf{X}\mathbf{v} = \lambda\mathbf{v} \tag{8}$$

where $\mathbf{v}$ is the eigenvector corresponding to the principal component, and λ is its associated eigenvalue.

2.4.2. Partial Least Squares (PLS)

PLS is a supervised multivariate method that models relationships between predictor and response datasets, see Tenenhaus (1998). In multi-omics, PLS is useful for integrating two or more omics layers (*e.g.*, gene expression and metabolite levels) and relating them to phenotypic outcomes, see Lê C. *et al.* (2008).

PLS finds weight vectors $\mathbf{w}_X$ and $\mathbf{w}_Y$ such that the covariance between the projections $\mathbf{X}\mathbf{w}_X$ and $\mathbf{Y}\mathbf{w}_Y$ is maximized:

$$\max_{\mathbf{w}_X, \mathbf{w}_Y} \operatorname{Cov}(\mathbf{X}\mathbf{w}_X, \mathbf{Y}\mathbf{w}_Y) \tag{9}$$

Variants like sparse PLS introduce regularization to enable feature selection.

2.4.3. Multi-Omics Factor Analysis (MOFA)

MOFA is a latent variable model specifically developed for the integration of multi-omics data. It decomposes each omics dataset into shared and data-specific factors, which correspond to biological or technical sources of variation, see Argelaguet *et al.* (2018).

Given K omics matrices $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}\}$, MOFA models each as:

$$\mathbf{X}^{(k)} = \mathbf{Z}\mathbf{W}^{(k)} + \mathbf{E}^{(k)} \tag{10}$$

where:

- $\mathbf{Z} \in \mathbb{R}^{n \times d}$ is a matrix of latent factors shared across datasets,
- $\mathbf{W}^{(k)} \in \mathbb{R}^{d \times p_k}$ are weights for dataset k,
- $\mathbf{E}^{(k)}$ is residual noise.

MOFA is probabilistic and handles missing data naturally. It enables unsupervised clustering, dimensionality reduction, and exploration of latent drivers in biological systems, see Vahabi and Michailidis (2022).

2.4.4. Sparse Multi-Block PLS (sMBPLS)

sMBPLS extends PLS to more than two data blocks and incorporates sparsity to identify the most informative features across all omics layers, see Li *et al.* (2012). It is especially suited for studies where multiple omics are related to a common response (*e.g.*, disease status or treatment outcome).

This method builds a global latent structure and optimizes for interpretability, making it useful in complex systems biology studies.

2.4.5. Gene-wise weights and feature selection

In some multivariate frameworks, gene-wise weights are assigned to different omics variables to evaluate their contribution to observed variance or phenotype association. These weights help rank and select biologically relevant features from high-dimensional data.

One example is the CNAmet model, which integrates copy number, methylation, and expression data using correlation structures and statistical weighting.

2.4.6. Advantages and limitations

Multivariate methods offer a range of advantages and face several challenges in the analysis of complex datasets. They enable joint analysis by accounting for co-variation and correlations among variables, which enhances the understanding of interdependencies in the data. These methods also facilitate dimensionality reduction, making high-dimensional omics data more tractable and interpretable. Additionally, they are powerful tools for discovering

latent factors that may represent hidden biological drivers of variation. Their flexibility allows them to be applied in both supervised and unsupervised learning contexts. However, multivariate methods can be computationally intensive, especially when applied to large-scale omics datasets, necessitating efficient algorithmic implementations. They are also prone to overfitting, particularly in scenarios with small sample sizes, which requires the use of regularization techniques. Furthermore, while these methods can uncover latent components, interpreting these components in terms of clear biological processes can be challenging.

3. AI and machine learning approaches

3.1. Variational Autoencoders (VAEs) in multi-omics data integration

Variational Autoencoders are a class of generative models that have gained popularity in multi-omics data integration due to their ability to model complex, non-linear relationships and uncover latent representations of high-dimensional biological data, see Kingma and Welling (2013). VAEs are especially well-suited for handling the noise, sparsity, and heterogeneity commonly found in multi-omics datasets, see Simidjievski *et al.* (2019).

3.1.1. Theoretical foundations of VAEs

VAEs belong to the family of probabilistic generative models and extend classical autoencoders by introducing a probabilistic framework. Instead of encoding an input x into a deterministic latent vector, VAEs encode it into a distribution over latent variable z. The goal is to learn the parameters of the generative model $p_\theta(x \mid z)$, and the inference model $q_\phi(z \mid x)$, typically with neural networks.

The VAE objective is to maximize the evidence lower bound (ELBO):

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - D_{\mathrm{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\right) \tag{11}$$

where:

- $\mathbb{E}[\log p_\theta(x \mid z)]$ is the reconstruction loss,
- D_{KL} is the Kullback–Leibler divergence between the approximate posterior and the prior $p(z)$, typically $\mathcal{N}(0, I)$.

This formulation ensures that the latent space z is both continuous and regularized, which enables smooth sampling and interpolation—useful for capturing underlying biological variation.

3.1.2. Application in multi-omics integration

In multi-omics studies, VAEs can be used to learn shared or modality-specific latent representations that capture the biological signal common across omics layers while accounting for layer-specific variation.

3.1.2.1. Integration strategies

a. Early integration (Full Fusion) Concatenate all omics datasets as input to a single VAE model.

b. Intermediate integration Each omics layer has a separate encoder, but a shared latent space is learned.

c. Late integration Separate VAEs are trained for each omics dataset, and their latent embeddings are later combined for downstream tasks (*e.g.*, clustering, classification).

These approaches support modularity, scalability, and flexibility in integrating omics with different feature spaces and distributions.

3.1.3. Tools

a. scVI A VAE model for single-cell RNA-seq data, modeling gene expression while correcting batch effects.

b. Multi-omics VAE Custom-built frameworks where omics-specific encoders feed into a joint decoder, enabling integrative modeling of transcriptomics, proteomics, and epigenomics, see Xin *et al.* (2024)

3.1.4. Advantages and limitations

Variational Autoencoders offer several benefits in biological research, particularly in the analysis of complex omics data. They enable dimensionality reduction by compressing high-dimensional data into low-dimensional latent factors that capture key biological variation. Their probabilistic framework enhances robustness to noise and batch effects, making them well-suited for real-world biological datasets. VAEs also handle missing data naturally by modeling the underlying data distribution, allowing for effective imputation. The latent space learned by VAEs often reveals meaningful clusters that correspond to phenotypes or disease subtypes, aiding in visualization and interpretation. Furthermore, VAEs support biomarker discovery by identifying important features that contribute to latent factors, which can be biologically interpreted. However, VAEs also come with challenges. The interpretability of latent dimensions can be limited, as they may not directly map to known biological processes. Training complexity is another issue, requiring careful tuning of the model architecture and learning parameters. Additionally, data scaling is crucial, as different omics types must be normalized to prevent bias in the latent space. Lastly, over-regularization due to the KL divergence term can overly constrain the latent space, potentially leading to underfitting and loss of important biological signals.

3.2. Graph-based learning in multi-omics data integration

Graph-based learning has emerged as a powerful strategy for integrating multi-omics data, particularly because biological systems are naturally structured as networks—whether they be gene regulatory networks, protein–protein interaction (PPI) networks, metabolic pathways, or cell–cell communication maps. Graph-based methods model the relationships

between entities (*e.g.*, genes, proteins, samples) as edges in a graph, enabling the analysis of topological structure, dependency, and contextual interactions across multiple omics layers.

In traditional machine learning, samples are often treated as independent and identically distributed. However, in multi-omics analysis, samples or features often exhibit nonlinear dependencies and interconnected behaviors that are better captured by graphs. For example: 1. Genes may co-express or be co-regulated, 2. Proteins interact physically or functionally, 3. Samples (patients) may be similar based on integrated omics profiles. Graph-based learning encodes this structure using nodes (*e.g.*, genes, proteins, samples) and edges (*e.g.*, co-expression, similarity, interaction), and applies machine learning techniques tailored for graphs, see Bengio *et al.* (2013).

3.2.1. Types of graph-based approaches

a. Similarity networks In this approach, each omics dataset is used to construct a similarity matrix between samples, which is then converted into a graph. These graphs are fused to form a unified network using methods such as Similarity Network Fusion. The final network can be analyzed using spectral clustering or community detection to identify subgroups (*e.g.*, disease subtypes).

b. Graph Neural Networks (GNNs) GNNs are deep learning models designed to operate on graph-structured data. They aggregate information from neighboring nodes and learn node embeddings that capture structural and feature information, see Kipf and Welling (2017). For multi-omics, nodes may represent genes with features from multiple omics. Edges may encode gene–gene relationships or pathway links. The GNN learns to predict phenotypes or latent node properties using neighborhood context, see Velickovic *et al.* (2017).

A common formulation in a GNN layer is:

$$\mathbf{h_v^{(l+1)}} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \frac{1}{c_{vu}} \mathbf{W^{(l}h_u^{(l)}}\right) \tag{12}$$

where:

- $\mathbf{h_v^{(l)}}$ is the representation of node v at layer l,
- $\mathcal{N}(v)$ is the set of neighbors of node v,
- c_{vu} is a normalization constant,
- $\mathbf{W^{(l)}}$ is the learnable weight matrix, and
- σ is a non-linear activation function.

c. Network propagation and diffusion These algorithms propagate information (*e.g.*, expression signals, mutation scores) over a network to prioritize relevant nodes, see Köhler *et al.* (2008). Examples include:

- **Random Walk with Restart (RWR)** A random walker starts at a node and probabilistically explores the network, returning to the start with probability **r**. This helps rank nodes based on their proximity to known disease genes.

$$p_{t+1} = (1 - r)Wp_t + rp_0 \tag{13}$$

where: p_t is the probability vector at time t, W is the transition matrix, p_0 is the initial distribution.

- **NetICS, TieDIE** Used for integrating mutation data with expression or pathway data using directed propagation, see Paull *et al.* (2013).

d. Probabilistic graphic models These include Bayesian Networks and Markov Random Fields (MRFs) that model conditional dependencies among variables (genes, proteins, *etc.*). For example, PARADIGM infers pathway activities by combining multiple omics layers within a Bayesian graphical model framework.

3.2.2. Advantages and limitations

Graph-based learning has emerged as a powerful approach in multi-omics analysis due to its ability to model complex, structured biological relationships. It has been applied in various domains such as cancer subtype classification, where methods like Graph Neural Networks and Similarity Network Fusion cluster patients based on integrated omics profiles; biomarker discovery, where network diffusion identifies genes or proteins functionally related to known disease markers; pathway activity inference, with tools like PARADIGM integrating gene expression and copy number data to predict pathway status; and feature selection, where GNN attention mechanisms highlight informative nodes for downstream analysis. The advantages of graph-based methods include their natural representation of biological systems using existing knowledge like gene networks, flexibility in handling non-Euclidean and structured data, context-aware learning through neighborhood-informed node embeddings, and scalability enabled by recent computational advances. However, challenges remain, such as the need for careful data preprocessing to construct reliable graphs, limited interpretability of deep graph models, complexity in integrating heterogeneous omics layers without introducing bias or losing specificity, and the high computational demands of training large-scale graph models.

4. Conclusion

Multi-omics data integration is at the forefront of systems biology, enabling a holistic view of cellular function by combining genomic, transcriptomic, proteomic, metabolomic, and other omics data types. Each method explored—statistical, machine learning, and network-based—offers unique strengths in addressing the challenges of high-dimensionality, heterogeneity, and noise inherent in biological data. Statistical approaches, particularly Canonical Correlation Analysis and its variants (sparse and regularized CCA), provide interpretable linear models for discovering cross-domain correlations between omics layers. These models are well-suited for moderate-dimensional data and are often used as a first step in integrative analysis. Similarity-based methods, such as Similarity Network Fusion, excel in clustering

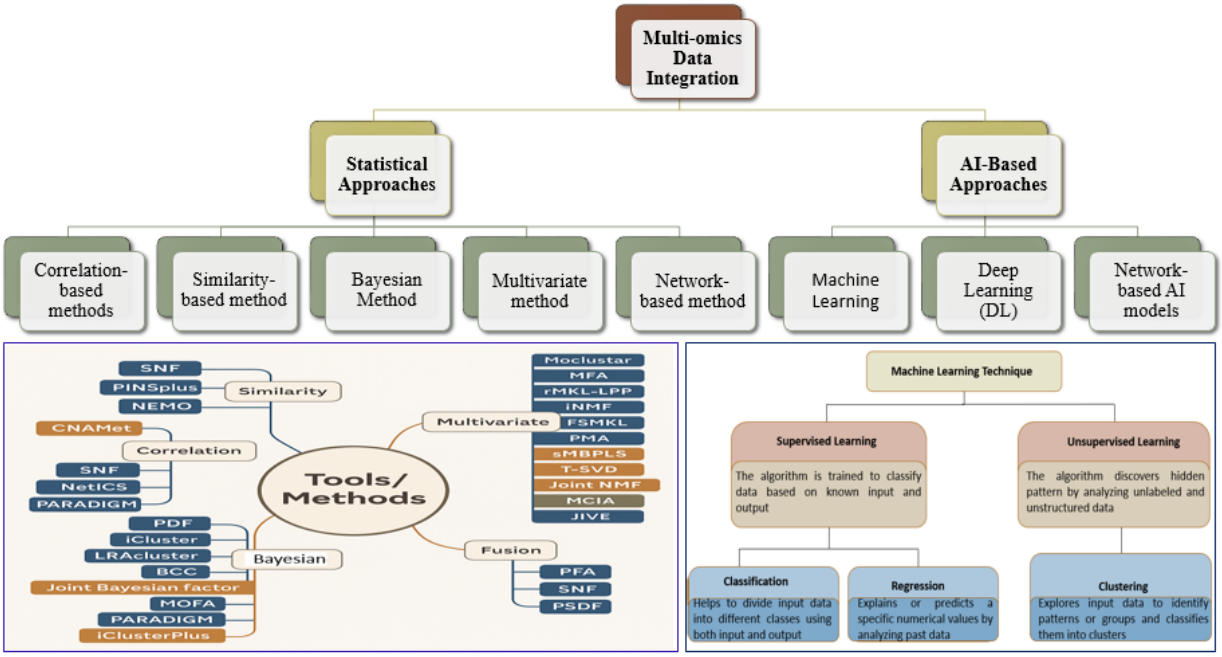

Figure 1: Workflow diagram of AI and statistical methods of multi-omics data integration

Table 1: Multi-omics public datasets and compatible methods

Dataset / Resource	Multi-omics layers	Compatible methods
TCGA (*via* GDC portal)	mRNA, miRNA, methylation, CNV, proteomics	PCA, PLS, SNF, BCC, PARADIGM, MOFA, *etc.*
ICGC	Genomics, transcriptomics, epigenomics	Same as TCGA, broader diversity
CMOB benchmark (TCGA-based)	Processed multi-cancer data	All listed ML/stat methods
MixOmics example sets	mRNA, proteome, metabolome	PCA, PLS, sMBPLS, CCA
BioGRID interactions + TCGA	Network /expression/proteomics	GNN, RWR

and patient stratification by leveraging sample-level relationships across different datasets. These methods are robust to missing features and offer flexible data-type integration through graph-based fusion strategies. Bayesian models introduce a probabilistic framework that explicitly handles uncertainty and allows for the incorporation of prior biological knowledge. They are particularly effective in unsupervised clustering, causal inference, and modeling hidden structures in multi-omics data, though often computationally demanding. Multivariate methods, including PCA, PLS, MOFA, and sMBPLS, help in reducing dimensionality and uncovering latent variables that drive shared or specific biological variation across omics layers. These techniques are scalable and interpretable, making them widely adopted in both research and clinical settings.

Variational Autoencoders represent a more recent advancement, leveraging deep learning to capture complex, non-linear patterns and generate latent representations. Their flexibility in integration strategies (early, intermediate, late) and natural handling of missing data make them highly promising for large, noisy, and heterogeneous datasets. Graph-based learning, including Graph Neural Networks and network propagation methods, allows integration of biological interaction networks with omics data. These methods encode structural dependencies, enhance biological interpretability, and enable feature prioritization based on contextual relevance within the network. However, no single method is universally superior; instead, the choice depends on the specific research question, data type, sample size, and computational resources. As computational methods advance and multi-omics datasets expand, integrative approaches will continue to unlock new insights into complex diseases, biological pathways, and precision medicine.

Acknowledgements

We are indeed grateful to the Editors for their guidance and counsel. We are very grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

Table 2: Multi-omics data integration methods

Method	Function	Advantages	Limitations
Canonical Correlation Analysis (CCA)	Finds linear combinations of features in two datasets that are maximally correlated.	Simple and interpretable; suitable for moderate-dimensional data.	Assumes linearity; unstable when number of variables exceeds samples; sensitive to noise.
Sparse/Regularized CCA	Extends CCA with sparsity (L1) or regularization to improve feature selection or stability.	Feature selection; better suited for high-dimensional omics data.	Parameter tuning required; interpretability can decrease with complexity.
Similarity Network Fusion (SNF)	Constructs sample-sample similarity networks from each omics and fuses them iteratively.	Handles heterogeneous data; robust to missing features; good for clustering.	Sensitive to similarity metric choice; requires careful normalization and parameter tuning.
Bayesian Clustering (MDI, BCC)	Uses probabilistic models to assign samples to latent clusters across datasets.	Models uncertainty; incorporates prior knowledge; captures hidden structure.	Computationally intensive; may require strong assumptions or priors.
Bayesian Networks (*e.g.*, PARADIGM)	Models conditional dependencies among omics variables *via* DAGs.	Captures causal relationships; integrates multiple data types with biological priors.	Complex to construct; inference can be slow and sensitive to data quality.
Principal Component Analysis (PCA)	Reduces dimensionality by capturing directions of maximum variance.	Simple, fast, and unsupervised; good for visualization and variance exploration.	Assumes linearity; may overlook class-specific patterns; not tailored to response variables.
Partial Least Squares (PLS)	Projects data onto latent variables that correlate with outcomes.	Supervised; identifies correlated features across data types.	May overfit with small sample sizes; assumes linear relationships.
Multi-Omics Factor Analysis (MOFA)	Learns shared and specific latent factors across omics layers.	Probabilistic; handles missing data; interpretable latent structure.	Assumes Gaussian distributions; requires tuning of latent dimensionality.
sMBPLS (Sparse Multi-Block PLS)	Integrates multiple omics datasets with sparsity constraints.	Simultaneous integration and feature selection; interpretable loadings.	Computationally demanding; sensitive to sparsity level selection.
Variational Autoencoders	Learns probabilistic latent representations; used for denoising, imputation, clustering.	Captures nonlinear patterns; handles missing data; flexible integration strategies.	Requires deep learning expertise; difficult to interpret latent variables biologically.
Graph Neural Networks (GNNs)	Learns on graph-structured data to capture node-level and graph-level representations.	Exploits interaction networks; context-aware; scalable with recent advances.	Graph construction can be noisy; hard to interpret; requires large labeled datasets.
Network Propagation (*e.g.*, RWR)	Spreads signals across biological networks to prioritize genes or features.	Integrates prior knowledge; useful for ranking and feature prioritization	Performance depends on quality of network; propagation may dilute weak but important signals.

Table 3: Comparison of multi-omics integration methods

Method	Description	Software / Package	Platform	Link / Notes
CCA (Canonical Correlation Analysis)	Identifies linear relationships between two data matrices	mixOmics::rcc / PMA::CCA	R	mixOmics, PMA
SNF (Similarity Network Fusion)	Constructs sample similarity networks and fuses them	SNFtool / SNFpy	R / Python	SNFtool (R)
BCC (Bayesian Consensus Clustering)	Unsupervised clustering across multiple data types	BayesCC	R	GitHub - BayesCC
PARADIGM	Integrates multi-omics using pathway information	Java tool, also in UCSC Cancer Genomics Browser	Java / Web	PARADIGM GitHub, UCSC site
PCA (Principal Component Analysis)	Linear dimensionality reduction	Base R/prcomp, sklearn .decomposition. PCA	R / Python	scikit-learn PCA
PLS (Partial Least Squares)	Projects predictor and response variables to a new space	mixOmics::pls / sklearn.cross $_{d}$*ecomposition.* *PLSRegression*	R / Python	mixOmics, scikit-learn PLS
MOFA (Multi-Omics Factor Analysis)	Probabilistic latent variable model for multiple omics	MOFA2	R / Python	MOFA2 GitHub, Documentation
sMBPLS (Sparse Multi-block PLS)	PLS extension for multi-block data, sparse variant	mixOmics::block .spls	R	mixOmics - block.spls
VAE (Variational Autoencoder)	Deep learning model to learn latent representations	TensorFlow, PyTorch, scVI	Python	scVI, PyTorch VAE example
GNN (Graph Neural Networks)	Deep models on graph-structured omics data	PyTorch Geometric, DGL, Spektral	Python	PyTorch Geometric, DGL, Spektral
RWR (Random Walk with Restart)	Graph-based propagation for gene prioritization	Custom or igraph, NetWalker	R / Python / Java	NetWalker, igraph

References

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A., and Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, **14**, e8124.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1798–1828.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, **374**, 531–547.

Kingma, D. P. and Welling, M. (2013). *Auto-Encoding Variational Bayes*, **1312**, Banff, Canada.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**, 3290–3297.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, **82**, 949–958.

Lê C., Kim, A., Rossouw, D., Robert, G., and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, **7**, p.35.

Li, W., Zhang, S., Liu, C.-C., and Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, **28**, 2458–2466.

Misra, B. B. (2018). New tools and resources in metabolomics: 2016–2017. *Electrophoresis*, **39**, 909–923.

Naserkheil, M., Ghafouri, F., Zakizadeh, S., Pirany, N., Manzari, Z., Ghorbani, S., Banabazi, M. H., Bakhtiarizadeh, M. R., Huq, M. A., and Park, M. N. (2022). Multi-omics integration and network analysis reveal potential hub genes and genetic mechanisms regulating bovine mastitis. *Current Issues in Molecular Biology*, **44**, 309–328.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, **8**, 1–34.

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics*, **29**, 2757–2764.

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, **13**, e1005752.

Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andres Terre, H., Shams, Z., Jamnik, M., and Liò, P. (2019). Variational autoencoders for cancer data integration: design principles and computational practice. *Frontiers in Genetics*, **10**, 1205.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, **14**, 1177932219899051.

Tenenhaus, M. (1998). *La régression PLS: théorie et pratique*. Editions technip.

Vahabi, N. and Michailidis, G. (2022). Unsupervised multi-omics data integration methods: a comprehensive review. *Frontiers in Genetics*, **13**, 854752.

Vaske, C. J., Benz, S. C., Stuart, J. M., and Haussler, D. (2021). Pathway recognition algorithm using data integration on genomic models (paradigm). **10**. US Patent 991,448.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *Stat*, **1050**, 10–48550.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, **11**, 333–337.

Witten, D. and Tibshirani, R. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, **8**, 1–27.

Wróbel, S., Turek, C., Stepień, E., and Piwowar, M. (2024). Data integration through canonical correlation analysis and its application to omics research. *Journal of Biomedical Informatics*, **151**, 104575.

Xin, L., Huang, C., Li, H., Huang, S., Feng, Y., Kong, Z., Liu, Z., Li, S., Yu, C., and Shen, F. (2024). Artificial intelligence for central dogma-centric multi-omics: Challenges and breakthroughs. *arXiv preprint arXiv:2412.12668*, **.**