Multi-Omics Integration strategies

Summary single-omics

Genomics

- Study DNA
- Techniques employed : NGS, WES, WGS
- Goal : To identify genetic variants that modify the DNA sequence

Transcriptomics

- Study RNA
- Techniques employed : NGS, RNA-seq
- Goal : To identify aberrant gene expression, alternative splicing events and non-coding transcripts

Proteomics

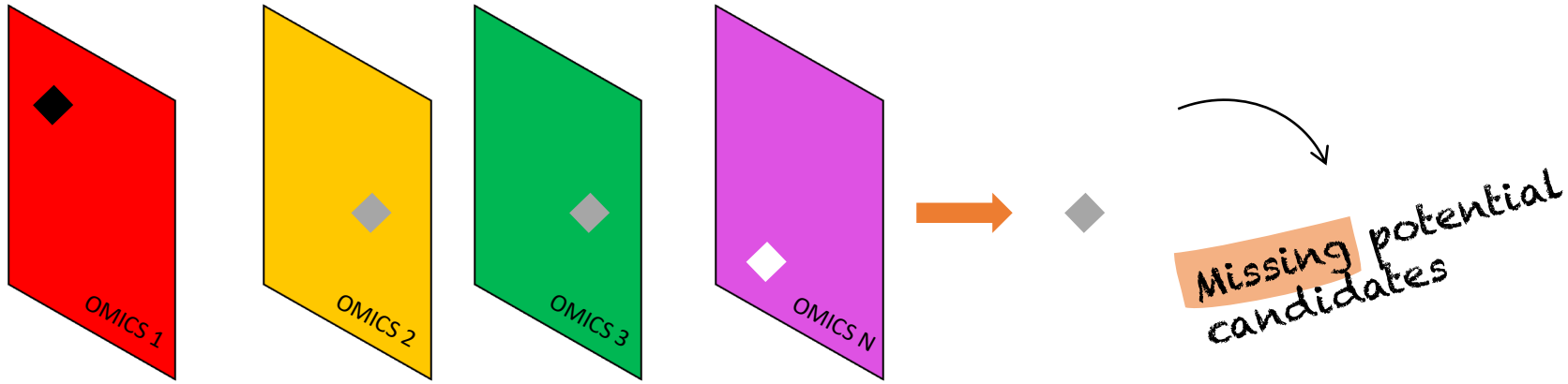
- Study proteins
- Technique employed : Mass spectrometry
- Goal : To charcarterize the abundance of proteins, and the quantification of post-translational modifications

Metabolomics

- Study metabolites
- Technique employed : Mass spectrometry
- Goal : To to identify biomarkers and therapeutic targets

Why multi-omics integration?

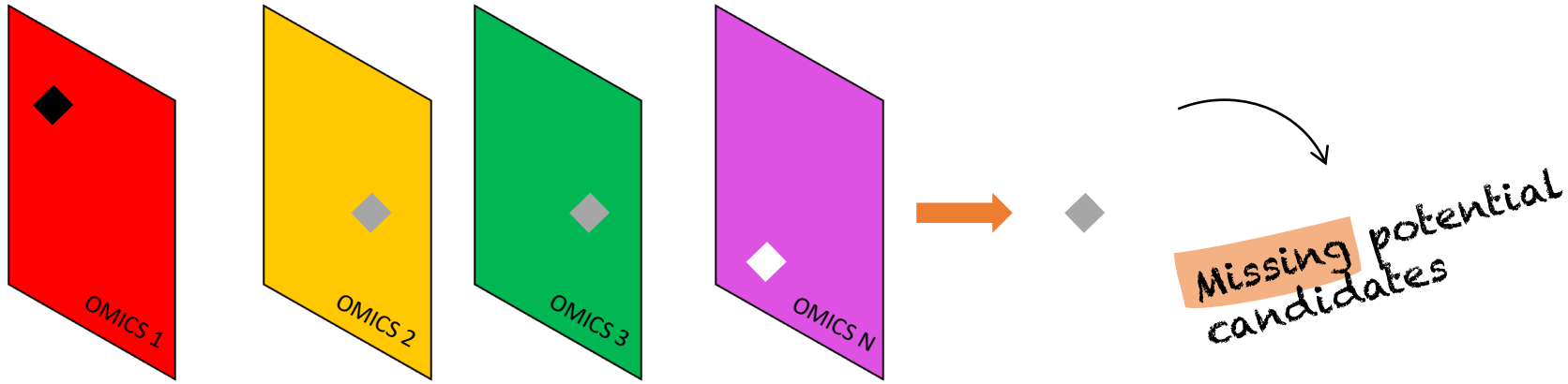
Multi-omics approach **without** integration



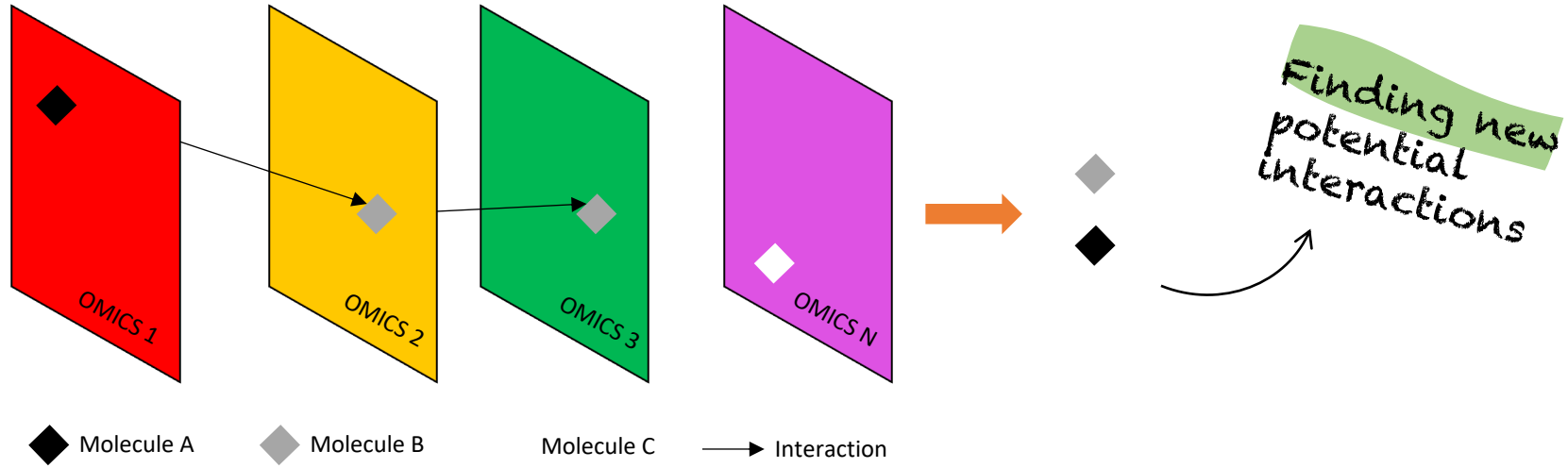
Molecule A Molecule B Molecule C Interaction

Why multi-omics integration?

Multi-omics approach **without** integration

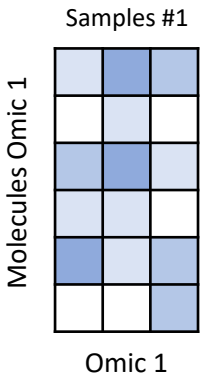


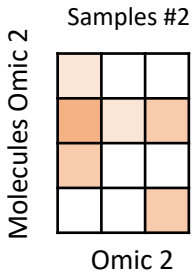
Multi-omics approach **with** integration



How to do multi-omics integration?

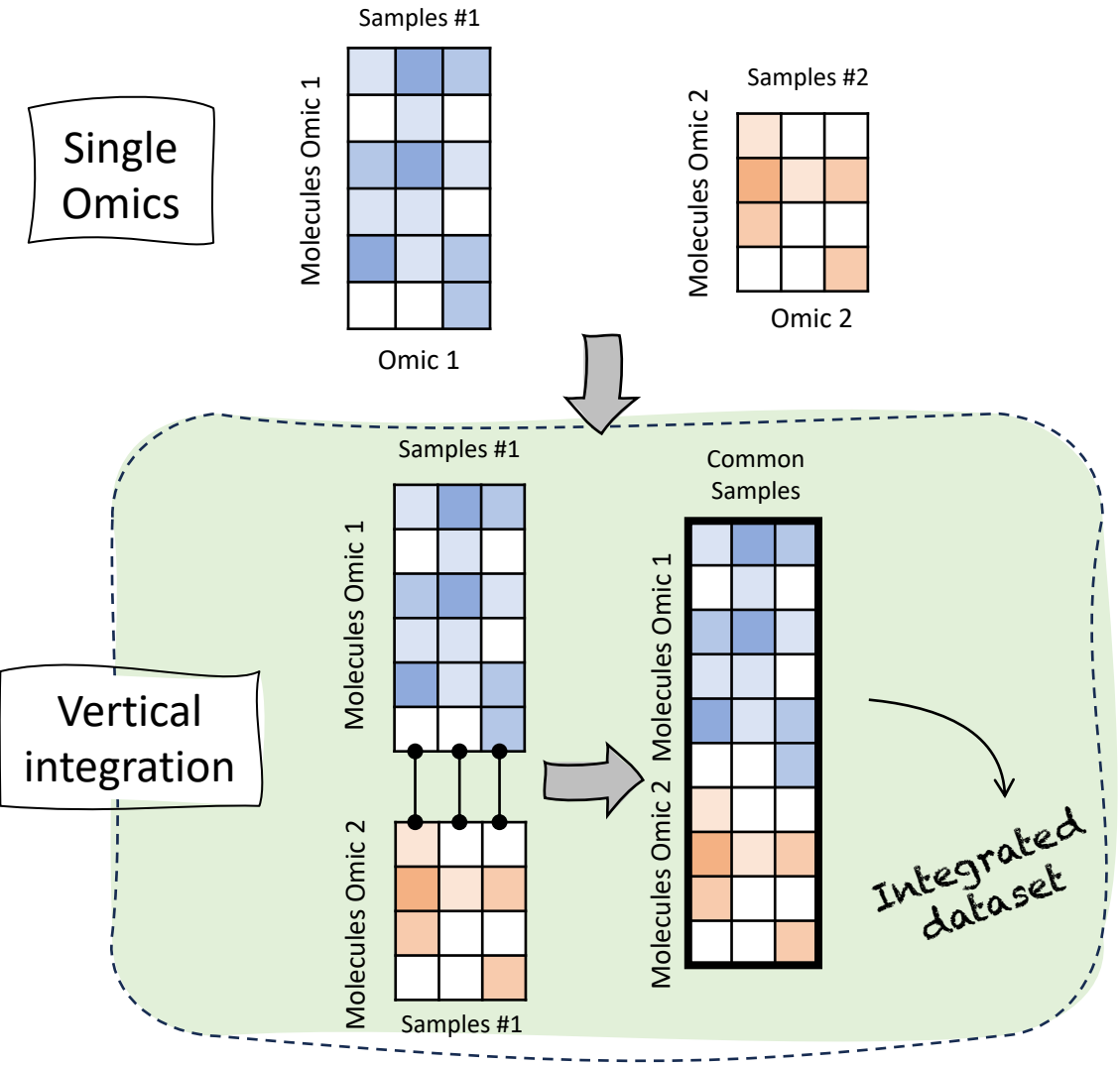
Single Omics





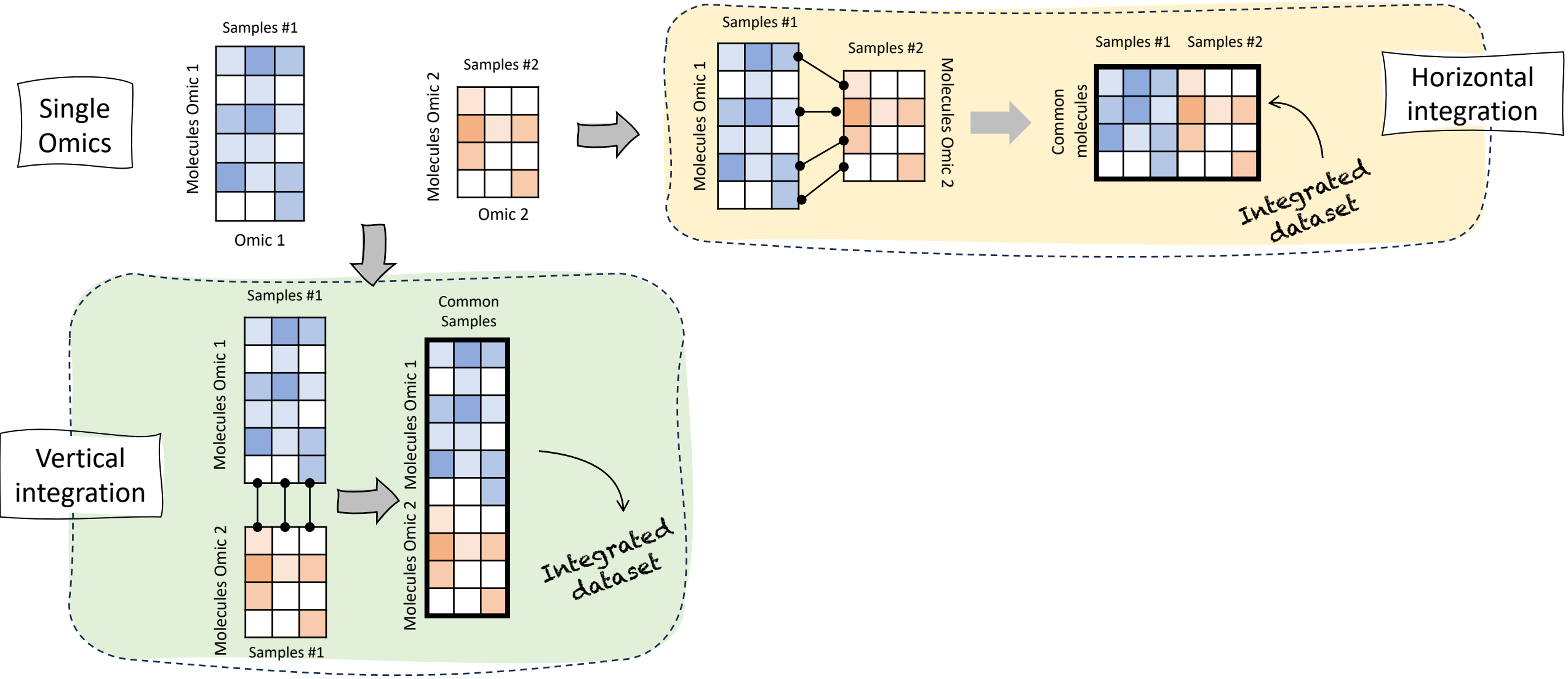
Labory J, Bottini S. The multi-omics revolution in the era of deep learning: allies or enemies? 2023, invited book chapter, in press

How to do multi-omics integration?



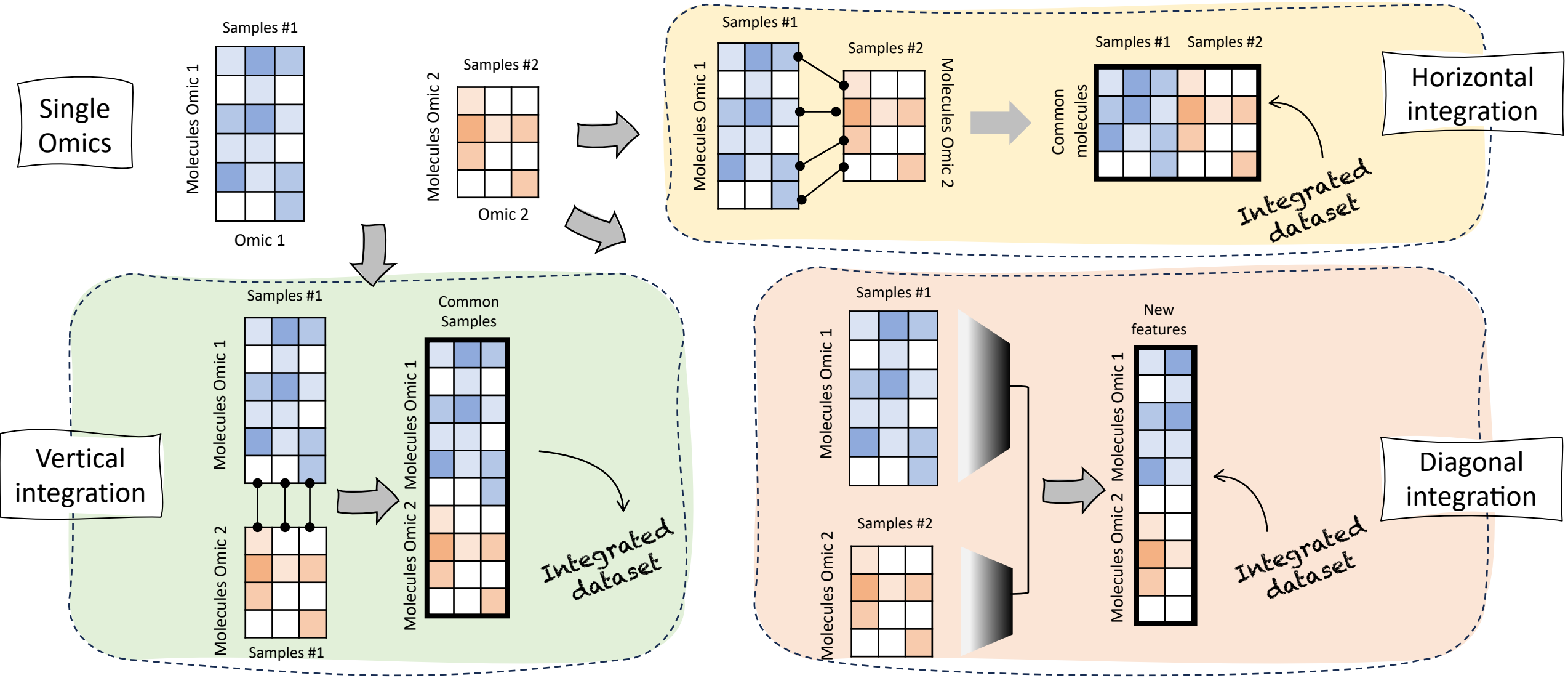
Labory J, Bottini S. The multi-omics revolution in the era of deep learning: allies or enemies? 2023, invited book chapter, in press

How to do multi-omics integration?



Labory J, Bottini S. The multi-omics revolution in the era of deep learning: allies or enemies? 2023, invited book chapter, in press

How to do multi-omics integration?

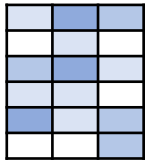


Labory J, Bottini S. The multi-omics revolution in the era of deep learning: allies or enemies? 2023, invited book chapter, in press

Challenges in multi-omics integration analysis

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

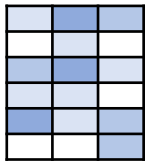
Challenges in multi-omics integration analysis



Heterogeneity, sparsity and uneven datasets.

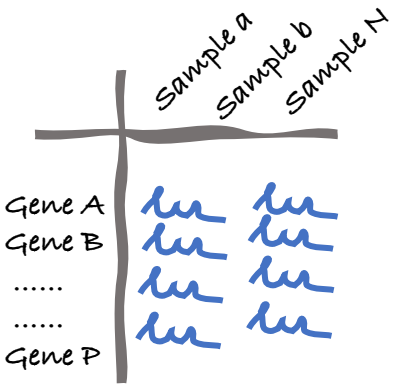
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Challenges in multi-omics integration analysis



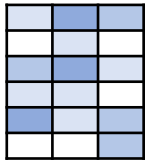
Heterogeneity, sparsity and uneven datasets.

More features than data (p >> n).



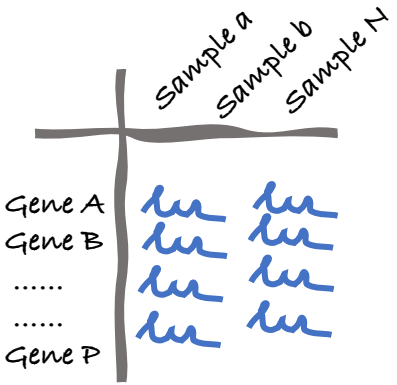
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Challenges in multi-omics integration analysis



Heterogeneity, sparsity and uneven datasets.

More features than data (p >> n).

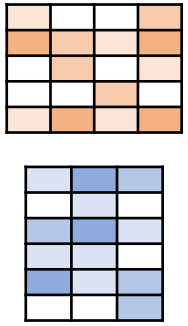


NaN

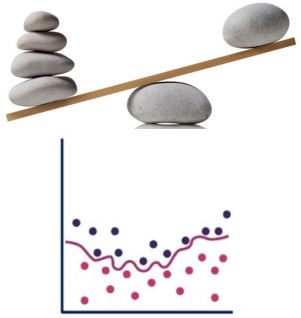
Missing data.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

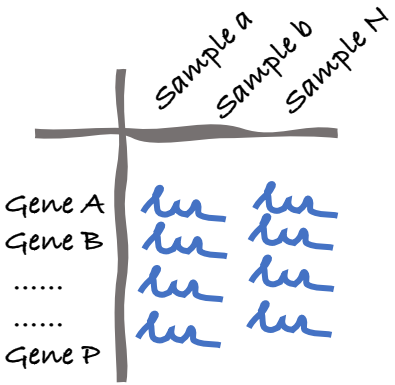
Challenges in multi-omics integration analysis



Heterogeneity, sparsity and uneven datasets.



Class imbalance and overfitting.



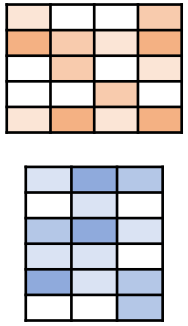
More features than data (p >> n).

NaN

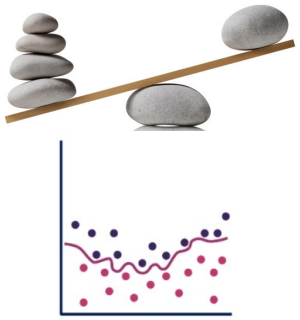
Missing data.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Challenges in multi-omics integration analysis

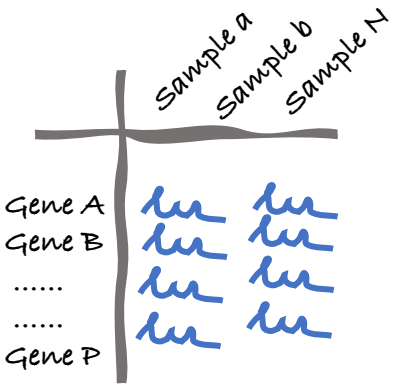


Heterogeneity, sparsity and uneven datasets.



Class imbalance and overfitting.

More features than data (p >> n).



How to choose a good model to perform the integration?

NaN

Missing data.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Main objectives of multi-omics data integration studies

- Patients grouping/stratification

Subtype identification. The discovery of disease subtypes can improve the definition of personalized and thus more effective treatments including biological drugs, hormonal therapy and immunotherapy. Also, subtype identification can identify heterogeneous groups within cancer cohorts with differences in disease progression or response to treatment. Usually, histopathology features, patient clinical profiles and symptoms are used by physicians to stratify patients accordingly. Currently, research studies are investigating new methodologies to assess disease subtype classifications by finding signatures at the molecular level. At first, these signatures were investigated at the single-omic level for example by finding common genes with perturbed expression. Recently multi-omics studies signatures are used to determine disease subtypes.

Drug response prediction. Patients affected by the same disease can respond to drug treatment differently. One of the central questions nowadays to achieve successful personalised medicine approaches is to be able to predict whether a drug will work on a group of patients with a similar molecular profile. Multi-omics integration analysis can be useful to study drug effects on specific cell lines or patient cells tailoring the way toward personalised medicine.

Main objectives of multi-omics data integration studies

- Molecular signatures and salient disease-pathways identification

Detect disease-associated molecular patterns. Similar to the previous objective and more generally, the identification of molecular markers associated with clinical markers or measurable characteristics is valuable in clinical practice. These molecular markers can be used as disease stage indicators or to identify disease-specific pathways and mechanisms. Multi-omics integration methods can reveal disease-associated molecules leading to the identification of salient patterns signature as relationships or patterns.

Understand regulatory processes. The previous objective can lead to the inference of disease-specific gene regulatory networks (GRNs) by combining measurements from multi-omics studies (E. Liu et al., 2019).

Diagnosis/Prognosis. The diagnosis of multifactorial diseases can be difficult to assess due to their complex genotype and phenotype. Thus diagnosis can be tedious and very long, resulting in patients in diagnostic stalemate. By using integrated multi-omics analysis, complex molecular signatures can be identified by employing approaches that are able to catch complex relationships among different molecular layers. By identifying the relationship between these indicators at the molecular level, a better explanation of the complex phenotype can be achieved and allow the prediction of disease progression and severity and course.

Multi-OMICS integration

Data-driven

Concatenation first

- Easy
- Straightforward

- Need proper normalization
- Do not take into account the different distribution of each omic
- Computationally intensive

Feature selection first

- Reduce computational costs

- Inter-omics relationships are lost
- Weak signals could be lost

Conversion first

- Allow easy representation of data

- May prevents the identification of indirect mechanisms
- Transformations can be challenging

Hypothesis-driven

Interactome first

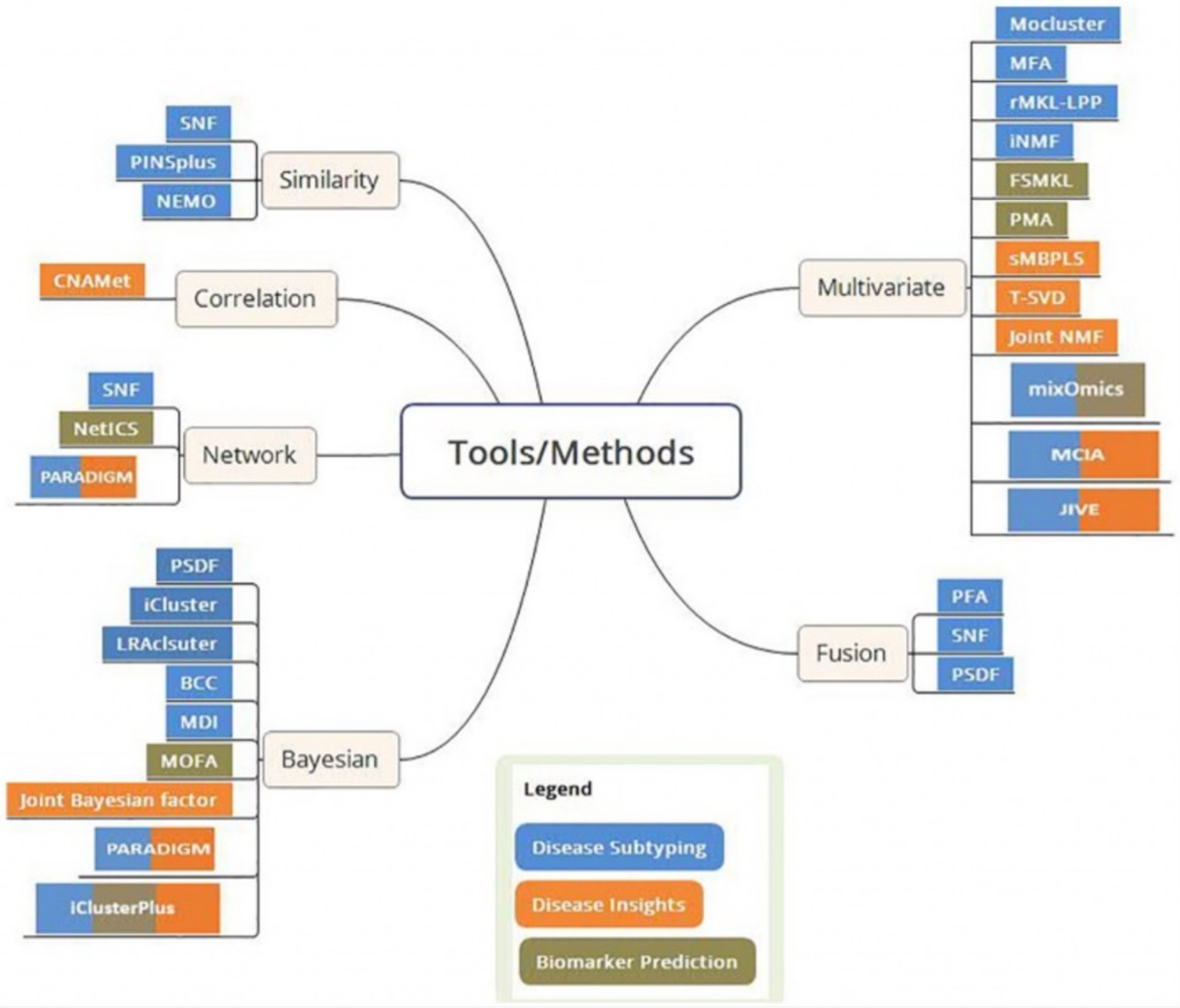
- Allow inter-omics connections

- Strongly depends on existing knowledge

Pathway as network

- Incorporate pathway topology

- Hard to extend
- Computationally intense


Tools/Methods
Similarity
SNF
PINSplus
NEMO
Correlation
CNAMet
Network
SNF
NetICS
PARADIGM
Bayesian
PSDF
iCluster
LRAclsuter
BCC
MDI
MOFA
Joint Bayesian factor
PARADIGM
iClusterPlus
Multivariate
Mocluster
MFA
rMKL-LPP
iNMF
FSMKL
PMA
sMBPLS
T-SVD
Joint NMF
mixOmics
MCIA
JIVE
Fusion
PFA
SNF
PSDF
Legend
Disease Subtyping
Disease Insights
Biomarker Prediction

Multi-omics methods for patient grouping/stratification

Concatenation-first approaches

The concatenation-first approach consists to create a unique dataset by joining all omics containing all features.

These methods take advantage of all the concatenated features to select the most discriminating features for a given phenotype.

This is the easiest strategy for analysing continuous or categorical data and the final comprehensive dataset can be used as the input for numerous classical machine learning algorithms

The drawback of these methodologies consists of training difficulties due to the high dimensionality of features of the joint dataset compared to the relatively small number of patients/samples which decreases the performance and increases the computational time.

Because of the striking difference in the feature dimensionality in multi-omics data, another pitfall of this technique is the model's tendency to learn more from the omics with the larger number of features.

Examples

A joint matrix of multi-omics features (including gene expression, copy number variation and mutation) was used with classical RF and SVM to predict anti-cancer drug response (Stetson et al., 2014).

Other than classical learning algorithms, such as deep neural networks (Tang et al., 2019) have also been widely used, or instance to identify robust survival sub-groups of liver cancer using RNA, miRNA and methylation data (Chaudhary et al., 2018).

Dimensionality reduction approaches

To minimize the number of features and avoid the problems of using a learning algorithm on the concatenated multi-omics dataset, one can select a subset of more relevant features (feature selection) or apply a dimensionality reduction technique in the pre-processing by keeping a small number of discriminating features (feature extraction).

This approach is based on the selection of the most salient features independently on each omic, either by selecting a subset of features based on some knowledge-based criteria or by using a feature extraction algorithm.

Sometimes a combination of the two techniques is used, because even after feature selection, most multi-omics datasets might still have too many features and high ration of features/samples.

Overall, the pitfall of these methods consists of a possible loss of relationship inter-omics due to the feature selection performed before integration.

Feature selection

In computational modelling, it is common practice to use any prior knowledge of the system to include into the model to achieve better performances.

Therefore, biological expertise can offer allow to select the features to include in the model with the highest probability to be relevant. Athreya et al. reduced over 7 million features consisting of SNPs and metabolites, to 65 predictor variables by using several biological-driven criteria (Athreya et al., 2018).

Usually, regarding transcriptome, only the genes showing the highest variability among samples showing different phenotypes are retained whereas genes with consistently low activity levels are usually discarded. Similarly, relevant features from epigenomic data can be selected by only keeping the loci near relevant genes or regions encompassing multiple methylation sites.

Feature extraction

Feature extraction is the process of condensing features into a lower number of new features (decided by the user). Despite the efficacy of these algorithms, the degree of explainability is sometimes lost since it is not always easy to explain what the new features are and how they are connected to the original ones. This is by far the widest application of learning algorithms to multi-omics integration studies. Several methods have been proposed, listed hereafter.

Factorization

Factorization-based techniques take all omics as input matrices and decompose them into two parts: (i) factors that are common to all omics and (ii) weights for each omic.

Common factors can be utilized for patient clustering/stratification, and weights help identify disease-causing molecules and/or biomarkers.

The assumption behind these models is that the biological mechanisms under investigation can be revealed by biological factors shared among multiple omics.

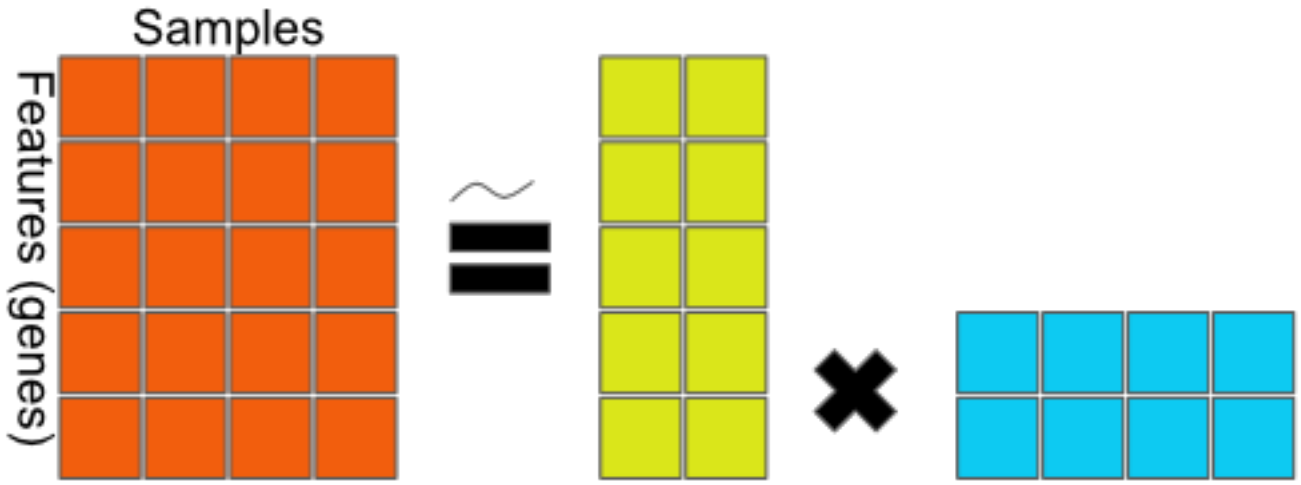
This type of approach allows to identify complex inter-omics structures and can be accomplished in two ways: tensor-based and matrix-based.

Factorization-based approaches have been extensively investigated in the literature for multi-omics integration, however the assumption to consider a global shared space among omics while neglecting partial common structures prevents the identification of indirect mechanisms.

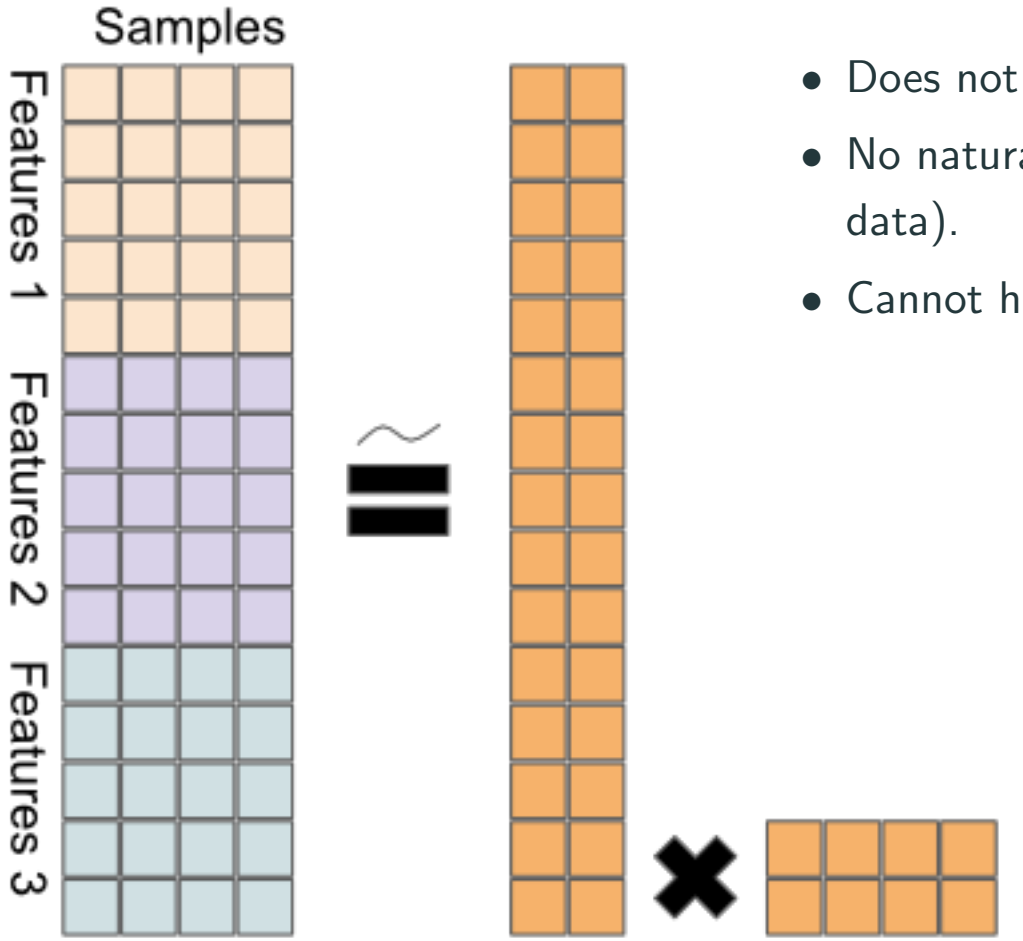
Matrix-based factorization

The most popular matrix factorization method is principal component analysis (PCA), which decomposes the covariance matrix of data to extract hidden biological factors.

A downside of PCA is that data are linearly transformed to generate the new features, whereas in real life relationship are rarely linear.



PCA



PCA is a great exploratory tool for single multivariate data sets, but it has important pitfalls in the analysis of multi-omics data:

- Does not generalise to an arbitrary number of data modalities.
- No natural way to combine different data modalities (binary data with continuous data).
- Cannot handle missing values.

MFA

Formally, we have

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_L \end{bmatrix} = WH,$$

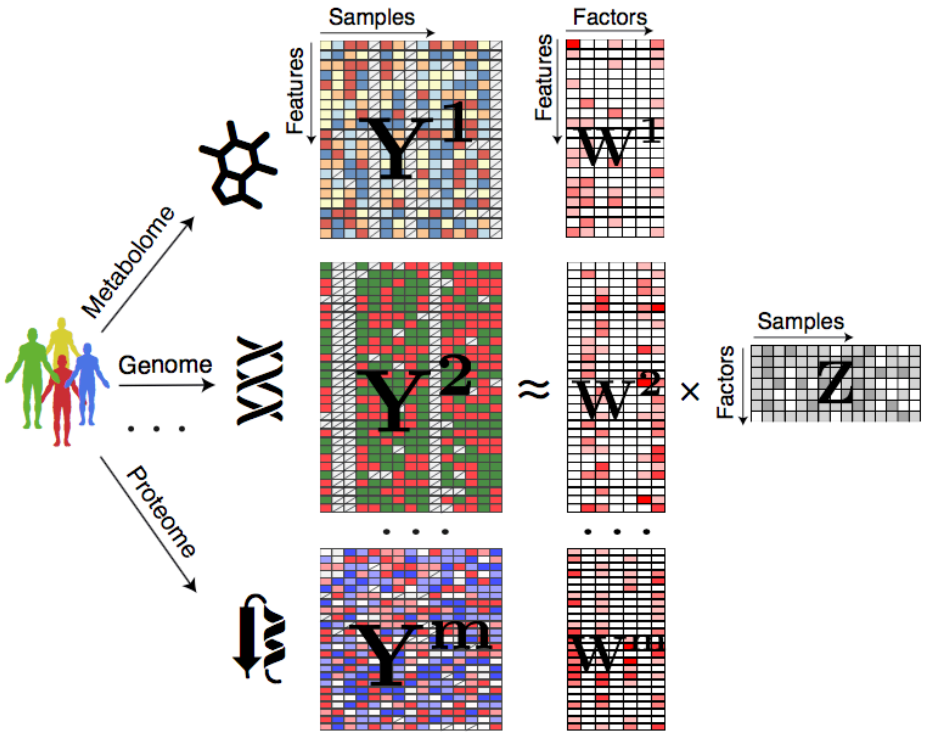
a joint decomposition of the different data matrices (X_i) into the factor matrix W and the latent variable matrix H. This way, we can leverage the ability of PCA to find the highest variance decomposition of the data, when the data consists of different omics types. As a reminder, PCA finds the linear combinations of the features which, when the data is projected onto them, preserve the most variance of any K-dimensional space. But because measurements from different experiments have different scales, they will also have variance (and co-variance) at different scales.

Multiple Factor Analysis addresses this issue and achieves balance among the data types by normalizing each of the data types, before stacking them and passing them on to PCA. Formally, MFA is given by

$$X_n = \begin{bmatrix} X_1/\lambda_1^{(1)} \\ X_2/\lambda_1^{(2)} \\ \vdots \\ X_L/\lambda_1^{(L)} \end{bmatrix} = WH,$$

where $\lambda_1^{(i)}$ is the first eigenvalue of the principal component decomposition of X_i.

MOFA



- The structure of the data is specified in the prior distributions of the Bayesian model
- The critical part of the model is the use sparsity priors, which enable automatic relevance determination of the factors

jNMF

NMF (Non-negative Matrix Factorization) is an algorithm from 2000 that seeks to find a non-negative additive decomposition for a non-negative data matrix

In the multi-omics context, we will, as in the MFA case, wish to find a decomposition for an integrated data matrix of the form

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_L \end{bmatrix},$$

with X_is denoting data from different omics platforms.

As NMF seeks to minimize the reconstruction error $||X - WH||_F$, some care needs to be taken with regards to data normalization. Different omics platforms may produce data with different scales (i.e. real-valued gene expression quantification, binary mutation data, etc.), and so will have different baseline Frobenius norms. To address this, when doing Joint NMF, we first feature-normalize each data matrix, and then normalize by the Frobenius norm of the data matrix. Formally, we run NMF on

$$X = \begin{bmatrix} X_1^N/\alpha_1 \\ X_2^N/\alpha_2 \\ \vdots \\ X_L^N/\alpha_L \end{bmatrix},$$

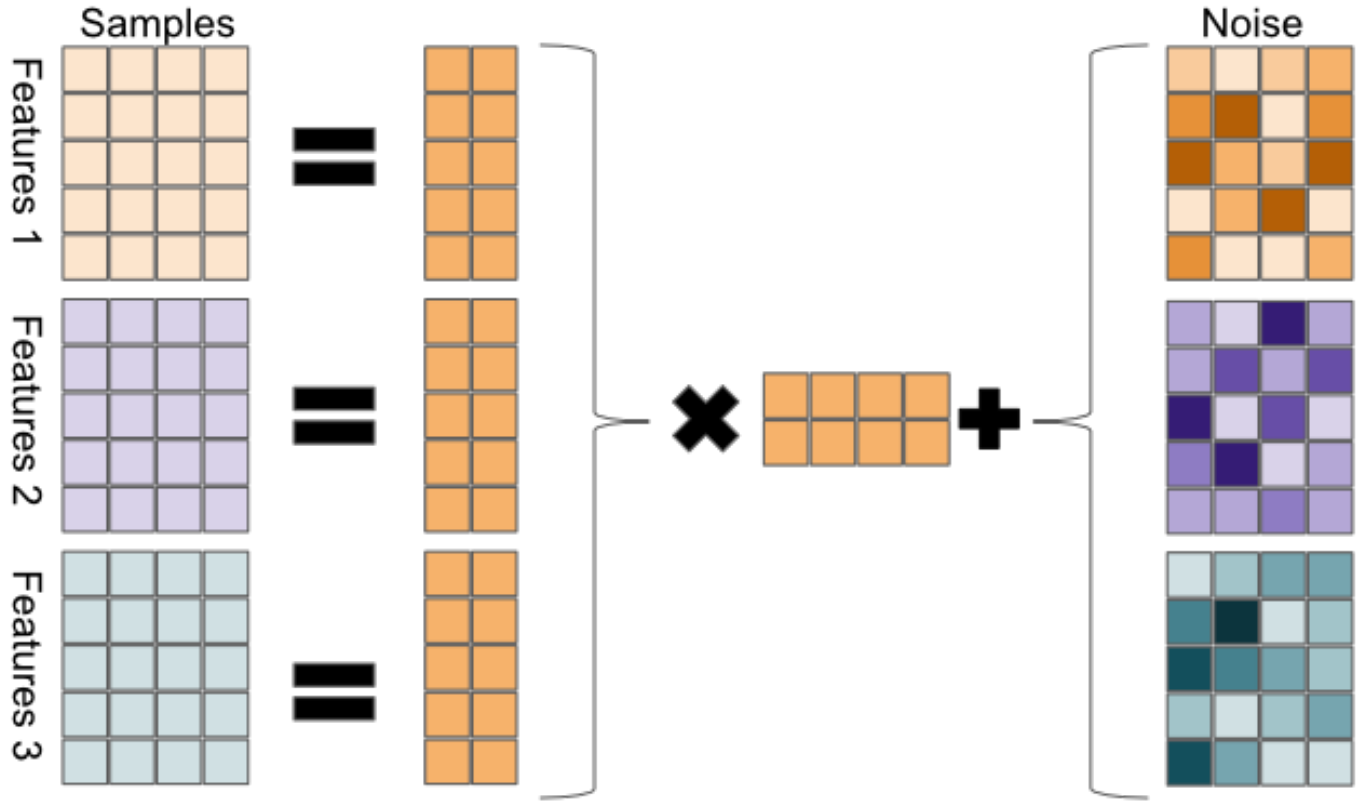
where X_i^N is the feature-normalized data matrix $X_i^N = \frac{x^{ij}}{\sum_j x^{ij}}$, and $\alpha_i = ||X_i^N||_F$.

iCluster

iCluster takes a Bayesian approach to the latent variable model. In Bayesian statistics, we infer distributions over model parameters, rather than finding a single maximum-likelihood parameter estimate. In iCluster, we model the data as

$$X_{(i)} = W_{(i)}Z + \epsilon_i,$$

where $X_{(i)}$ is a data matrix from a single omics platform, $W_{(i)}$ are model parameters, Z is a latent variable matrix, and is shared among the different omics platforms, and ϵ_i is a "noise" random variable, $\epsilon \sim N(0, \Psi)$, with $\Psi = diag(\psi_1, \ldots \psi_M)$ is a diagonal covariance matrix.



iCluster+

iCluster+ is an extension of the iCluster framework, which allows for omics types to arise from distributions other than a Gaussian. While normal distributions are a good assumption for log-transformed, centered gene expression data, it is a poor model for binary mutations data, or for copy number variation data, which can typically take the values $(-2, 1, 0, 1, 2)$ for heterozygous / monozygous deletions or amplifications. iCluster+ allows the different Xs to have different distributions:

- for binary mutations, X is drawn from a multivariate binomial
- for normal, continuous data, X is drawn from a multivariate Gaussian
- for copy number variations, X is drawn from a multinomial
- for count data, X is drawn from a Poisson.

In that way, iCluster+ allows us to explicitly model our assumptions about the distributions of our different omics data types, and leverage the strengths of Bayesian inference.

Both iCluster and iCluster+ make use of sophisticated Bayesian inference algorithms (EM for iCluster, Metropolis-Hastings MCMC for iCluster+), which means they do not scale up trivially. Therefore, it is recommended to filter down the features to a manageable size before inputting data to the algorithm. The exact size of “manageable” data depends on your hardware, but a rule of thumb is that dimensions in the thousands are ok, but in the tens of thousands might be too slow.

CCA

Canonical Correlation Analysis (CCA) is a simple extension of PCA to find linear components that capture correlations between **two** datasets[3].

Given two data matrices $\mathbf{Y}_1 \in \mathbb{R}^{N \times D_1}$ and $\mathbf{Y}_2 \in \mathbb{R}^{N \times D_2}$ CCA finds a set of linear combinations $\mathbf{U} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$ with maximal cross-correlation.

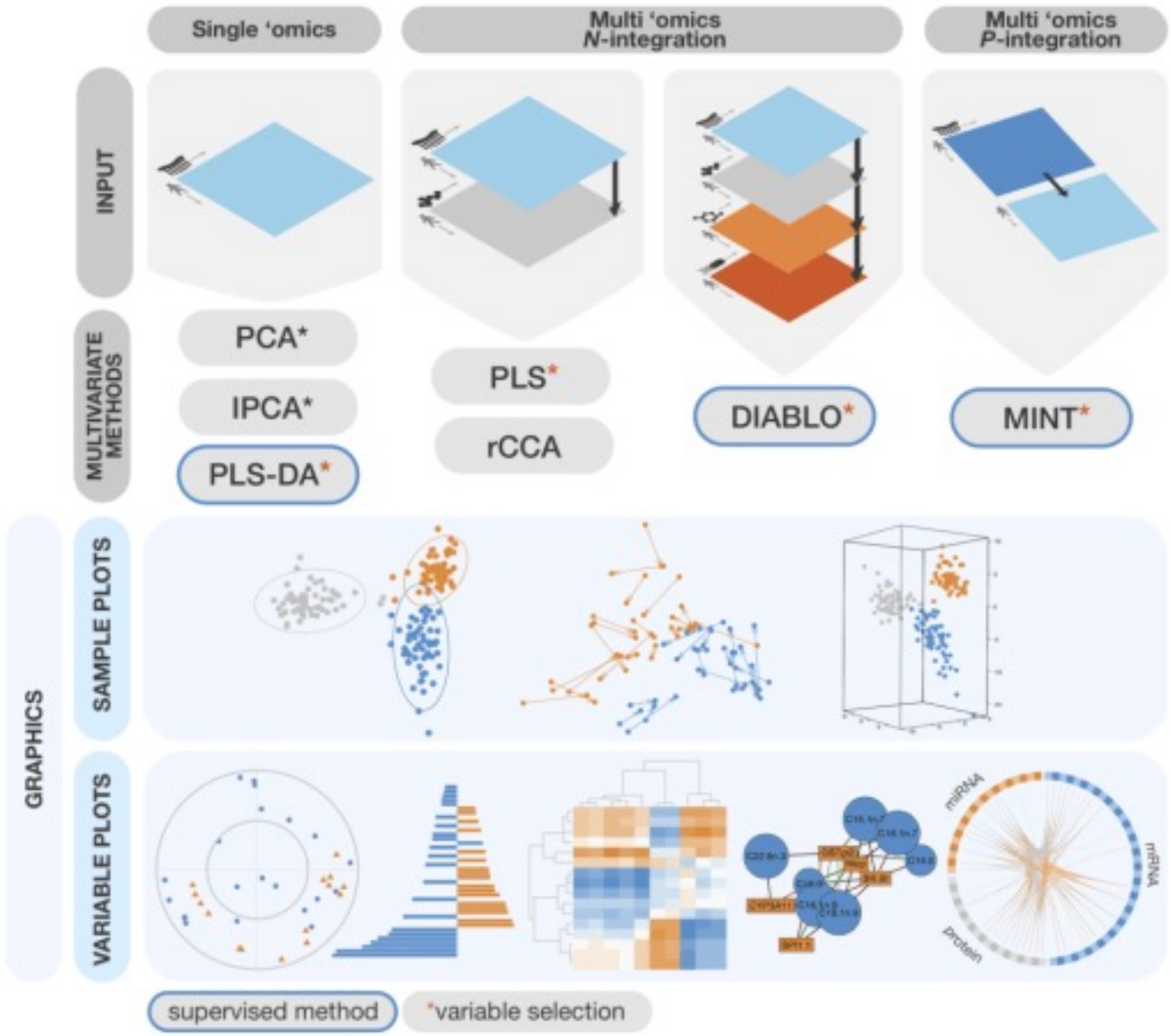
For the first pair of canonical variables, the optimisation problem is:

$$(\hat{\mathbf{u}_1}, \hat{\mathbf{v}_1}) = \underset{\|\mathbf{u}_1\|=1, \|\mathbf{v}_1\|=1}{\arg\max} \; corr(\mathbf{u}_1^T \mathbf{Y}_1, \mathbf{v}_1^T \mathbf{Y}_2)$$

In CCA the canonical components are defined as linear combinations of features that maximise the cross-correlation between the two data sets. This implies that:

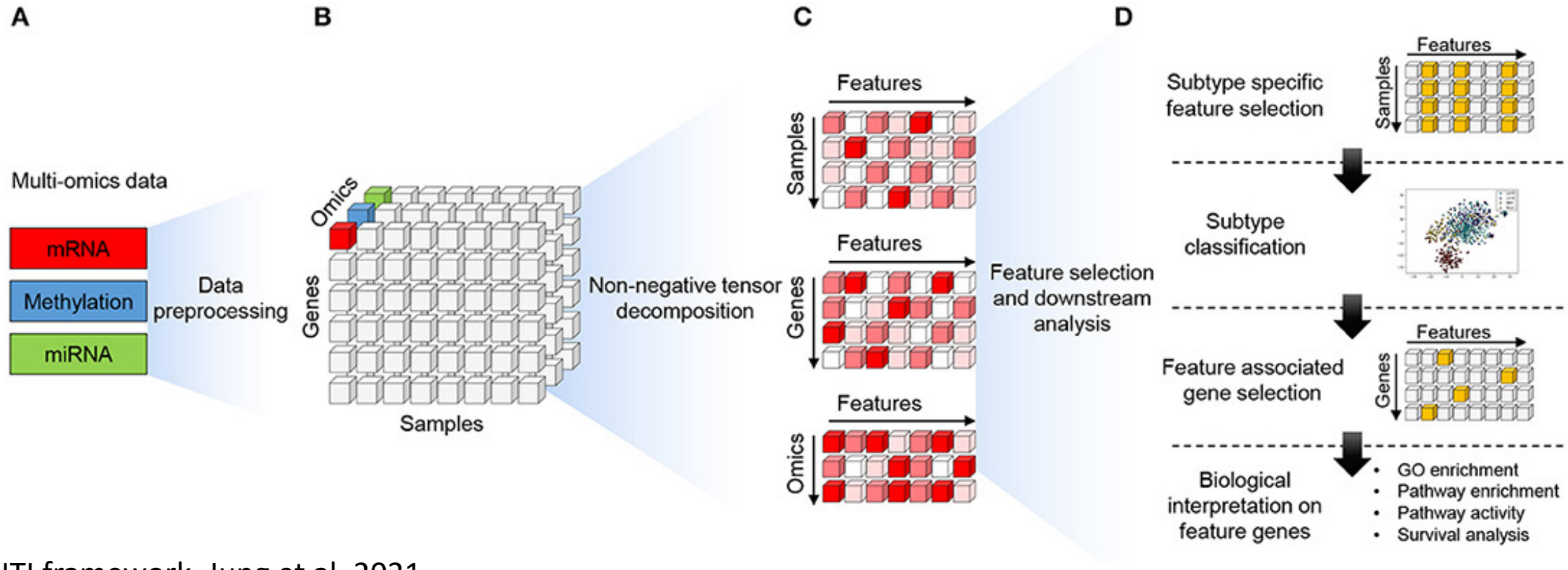
- It only works for the integration of *2* data sets
- It only finds sources of covariation between the two data sets. CCA is not able to find the sources of variation that are present within individual data sets

Mixomics


Single 'omics
Multi 'omics N-integration
Multi 'omics P-integration
INPUT
MULTIVARIATE METHODS
PCA*
IPCA*
PLS-DA*
PLS*
rCCA
DIABLO*
MINT*
GRAPHICS
SAMPLE PLOTS
VARIABLE PLOTS
miRNA
mRNA
protein
supervised method
*variable selection

Tensor-based factorization

Tensor factorization is a generalization of matrix factorization. Typically identifies higher-order relationships among biological variables by extracting factors that play essential roles in describing these relationships



MONTI framework, Jung et al. 2021

Kernel-based

is a form of transformation that uses kernel functions to map original features onto a new space with higher dimensions. Kernels allow such methods to work in high-dimensional space to explore similarities and relationships between samples.

Support Vector Machine and Multiple kernel learning (MKL) are common machine learning algorithms for working with kernels.

Despite the attractive performances, this approach is computationally expensive compared to other transformation-based techniques.

Examples

fMKL-DR is a method that generates kernel matrices from each input omics data, then combines them and finally uses SVM to stratify the data (Giang et al., 2020).

Neural network-based

This is a new and fast-growing area of research in multi-omics integration due to its superior performance in numerous domains of multimodal learning.

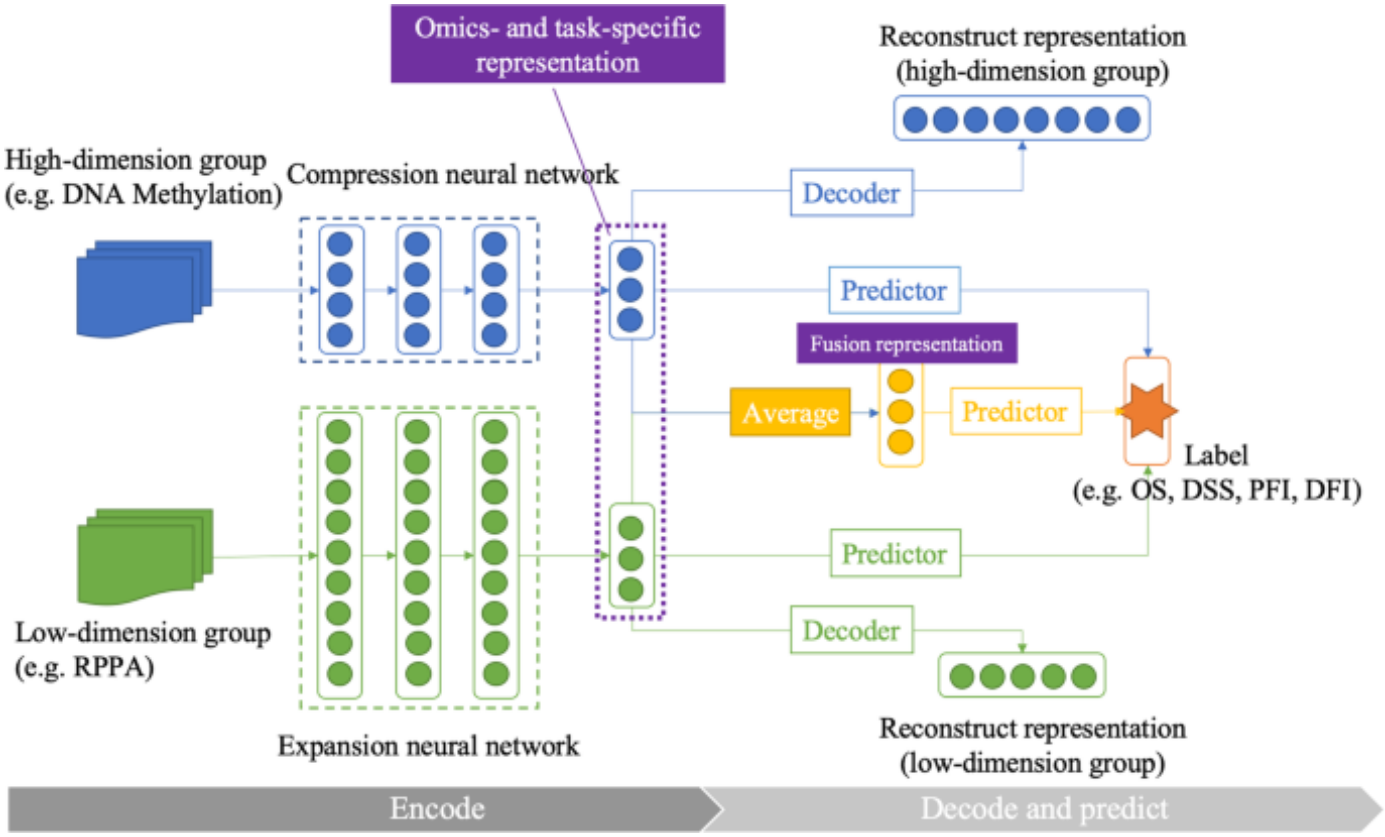
Usually, a network is trained for each omic to learn a joint representation of the inputs. The hidden layers of the built networks are then passed into another neural network.

They are particularly suitable for multi-omics integration because their particular structure allows for learning complex non-linear relationships of features.

Furthermore, they can deal with both structured data, like gene or protein expressions (Chen et al., 2016) and unstructured data, such as medical images (X. Liu et al., 2021; McBee et al., 2018).

Examples

MOSAE (Multi-omics Supervised Autoencoder) (Tan et al., 2020) was developed for pan-cancer analysis and compared with conventional ML methods such as SVM, DT, naïve Bayes, KNN, RF and AdaBoost. MOSAE uses a specific AE for each omics, according to their size of dimensions to generate omics-specific representations. Then, a supervised autoencoder is constructed based on specific autoencoder by using labels to enforce each specific autoencoder to learn both omics-specific and task-specific representations. Finally, MOSAE fuses all different omics-specific representations generated from supervised AEs and the fused representation is used for predictive tasks.

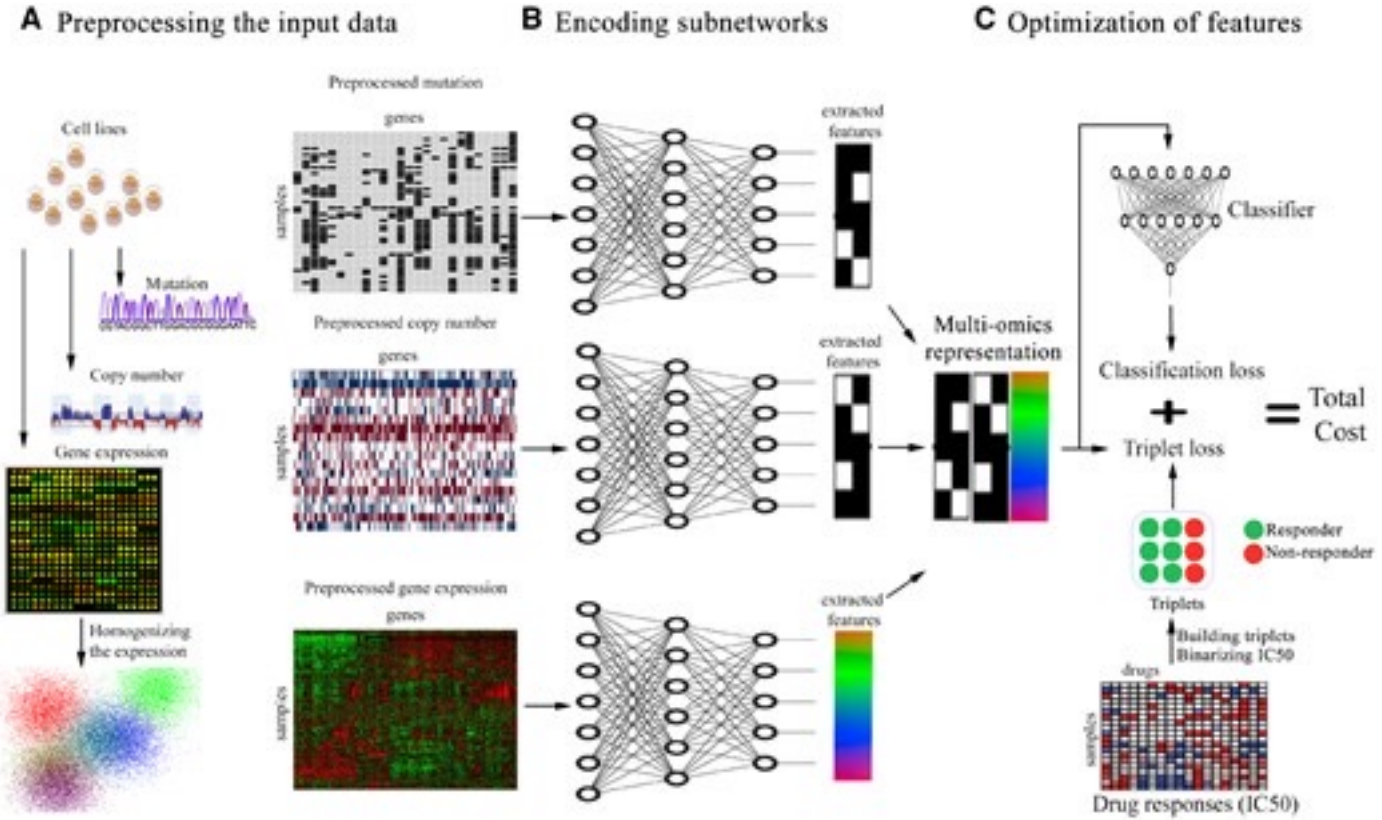


Examples

MAUI (multiomics autoencoder integration) is a non-linear dimension reduction method for multi-omics integration based on the use of variational AE to produce latent features that can be used for either clustering or classification (Ronen et al., 2019).

Examples

MOLI (multi-omics late integration) (Sharifi-Noghabi et al., 2019) is a deep-learning model that uses distinct encoding subnetworks to learn features of each omics data type. Then, the final network uses the concatenated features to classify the response of cancer cells to a given drug.



Multi-omics methods for molecular signatures and salient disease-pathways identification

Regression models

The earliest methods including learning models for multi-omics integration were mainly based on regression techniques. Usually, in these models, the expression of each gene is represented by a regression model that includes measurements from other omic datasets and their parameters. Thus to infer the relationship between a set of N genes and their regulators is decomposed into a set of N regression problems.

iGRN (Zarayeneh et al., 2017) was used to infer a gene regulatory network from human brain data of patients with 3 psychiatric disorders: schizophrenia, bipolar disorder and major depression. The samples have 25k gene expression (GE), 1028 copy number variations (CNVs) and 24k sites for DNA methylation. The method produces a gene-to-gene adjacency matrix and two bi-adjacency matrices for the interactions of CNV and DNA methylation with genes. The expression of each gene is then modelled by a sparse linear model incorporating other genes and also interaction effects of its nearby CNVs and DNA methylations.

Another example is BMNPGRN (Yuan et al., 2018). It uses non-convex penalty-based regression methods as an alternative to LASSO when dealing with sparse problems such as estimating interactions from multi-omics datasets. They studied the mechanisms of breast cancer, using to a multi-omics dataset of 760 cases and 80 control samples and identified potential regulators for key driver genes in the context of breast cancer.

Network based

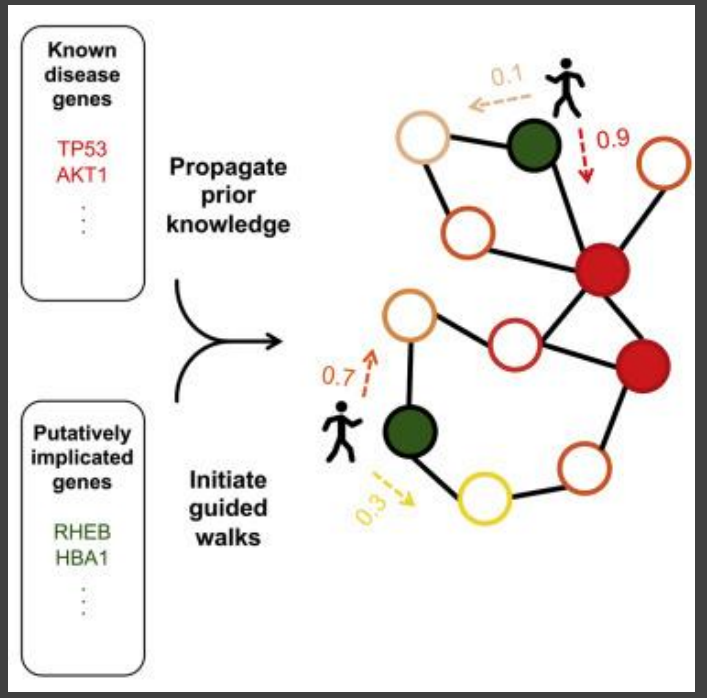
is one of the most used techniques for multi-omics integration in biomedical and healthcare studies. These techniques are capable to capture molecular interactions thus allowing to perform sample classifications and disease subtyping. Two network-based techniques are employed.

The first one models each omic as a network and combines them to make a unified network for carrying out further analysis and is usually combined with clustering methods.

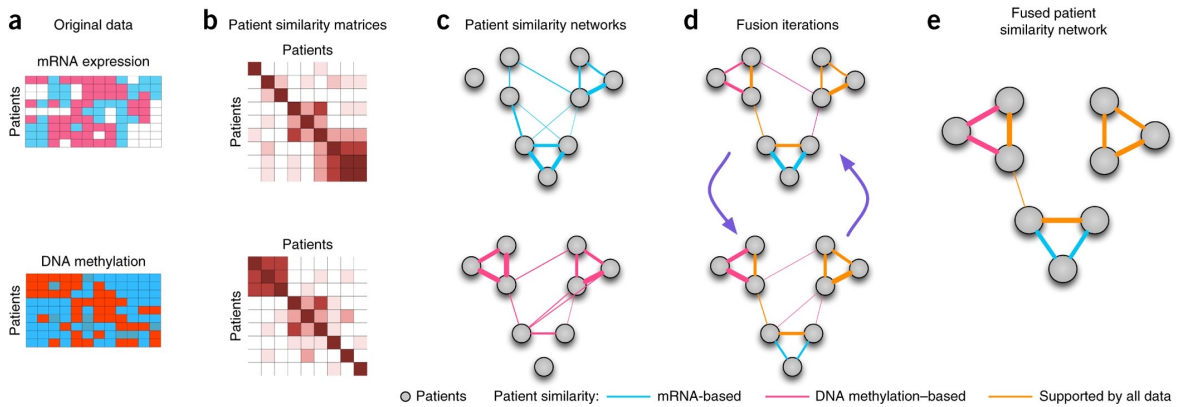
A network is created for each omic, the nodes represent samples, and the edges represent relationships between pairs of samples. Then, networks are converted to similarity matrices in an iterative optimization process or single iteration algorithm. Similarity Network Fusion (SNF) (B. Wang et al., 2014) and Neighbourhood-based Multi-Omics (NEMO) (Rappoport & Shamir, 2019) clustering are two examples belonging to this category.

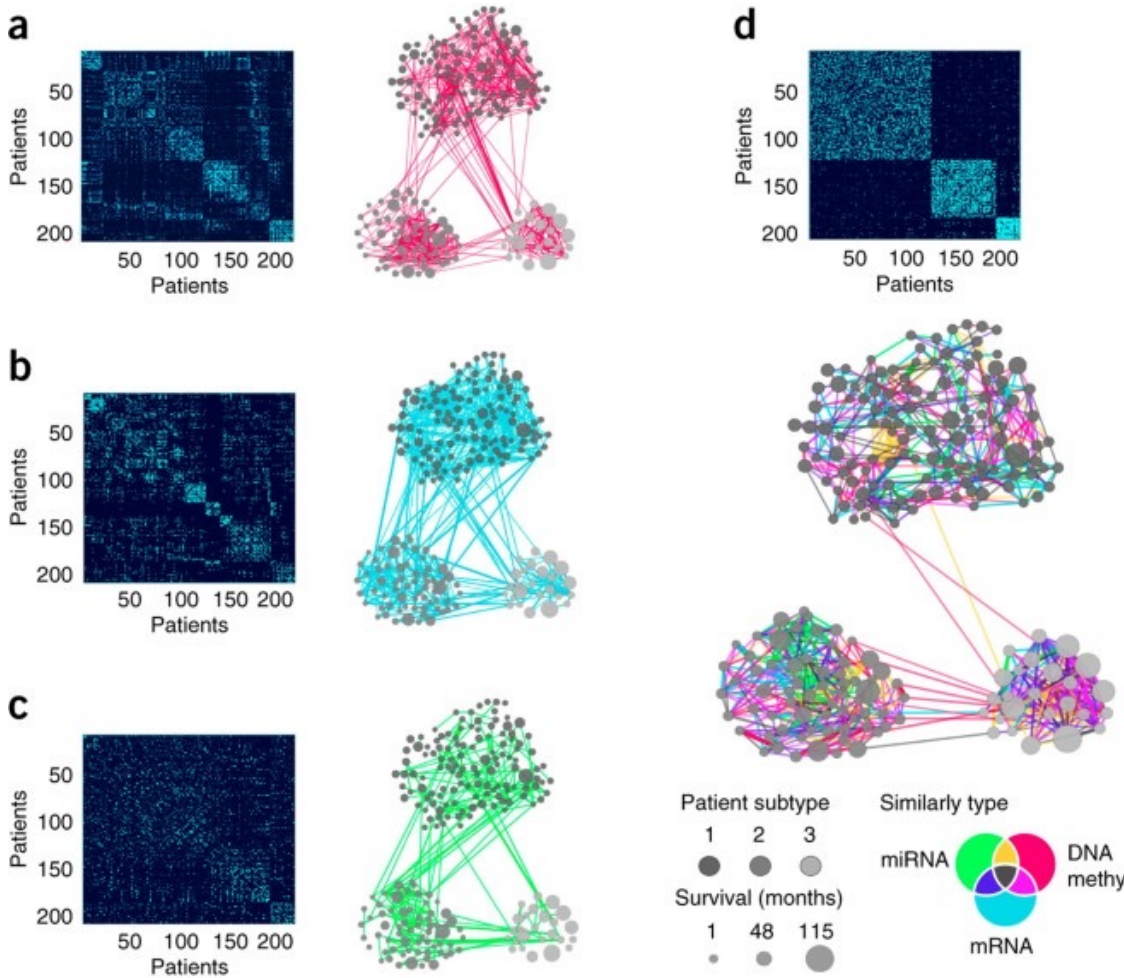
The second strategy network integrates latent representation spaces of networks into a joint representation (also known as network embedding) and feeds into machine learning models for classification purposes. In other words, each omic represented as a network is encoded into a low-dimensional space to reflect the network topology. Then, latent representations are combined to perform the downstream task.

- Network propagation methods can highlight nodes of interest in addition to already known ones
- Example algorithms: PageRank, HotNet2



SNF


a
Original data
mRNA expression
Patients
DNA methylation
Patients
b
Patient similarity matrices
Patients
Patients
Patients
Patients
c
Patient similarity networks
d
Fusion iterations
e
Fused patient similarity network
Patients
Patient similarity:
mRNA-based
DNA methylation–based
Supported by all data


a
b
c
d
Patients
50
100
150
200
Patient subtype
1
2
3
Survival (months)
1
48
115
Similarly type
miRNA
DNA methylation
mRNA

Multilayer network and network propagation

Another way to use networks for multi-omics integration is to build molecular networks. In these networks, nodes represent molecules (genes, proteins, metabolites), connected by edges that represent the pairwise relationships and interactions between two molecules.

Networks are used to represent all relevant interactions taking place in biological systems for example

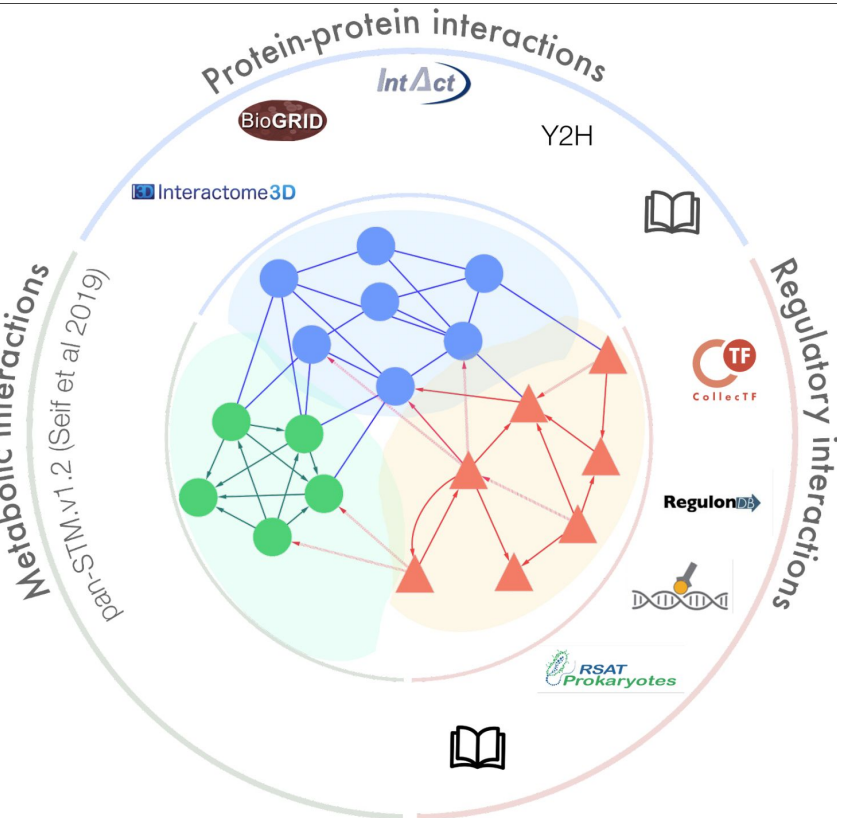
transcriptional regulation mechanisms, physical protein–protein interactions and correlation between RNA networks and protein structure or metabolic reactions (Montenegro, 2022).

Then topological features are evaluated including degree distribution to identify highly connected nodes (hubs) or shortest paths which determine the proximity between two nodes. Networks can also be divided into modules namely sub-network with highly connected nodes concerning the rest of the network. The nodes in these modules often share a similar function. Thus, by applying the ‘guilt by association’ property unknown entities in modules highly connected with known molecules are assumed to be functionally related (van Dam et al., 2018).

Inference methods to apply this principle are often applied to a single omic layer to identify interactions between molecules.
However, to elucidate interaction across multiple omics, the different layers should be connected.

Multilayered networks

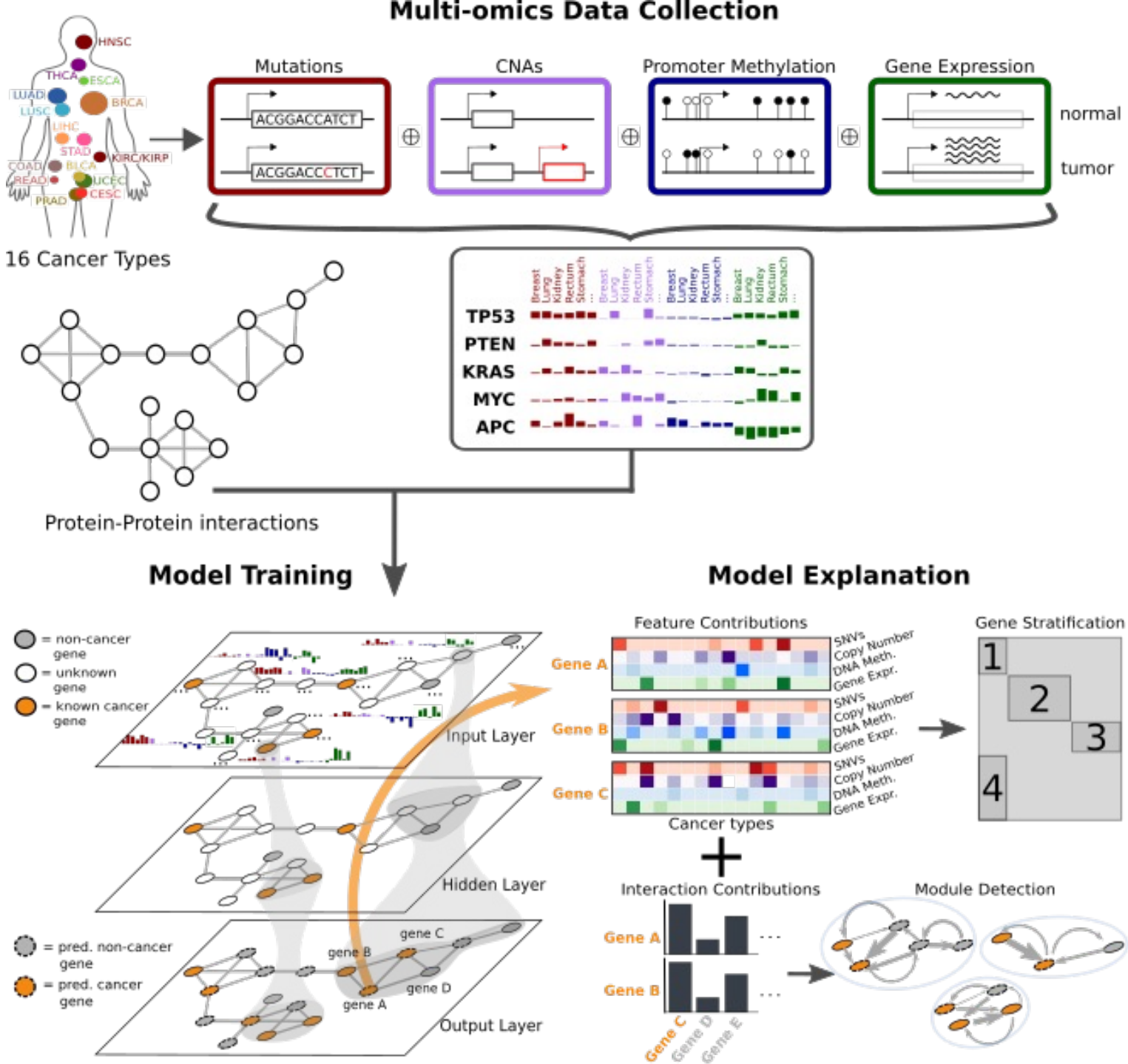
Multilayered networks contain multiple types of edges, changing the context of the interaction.



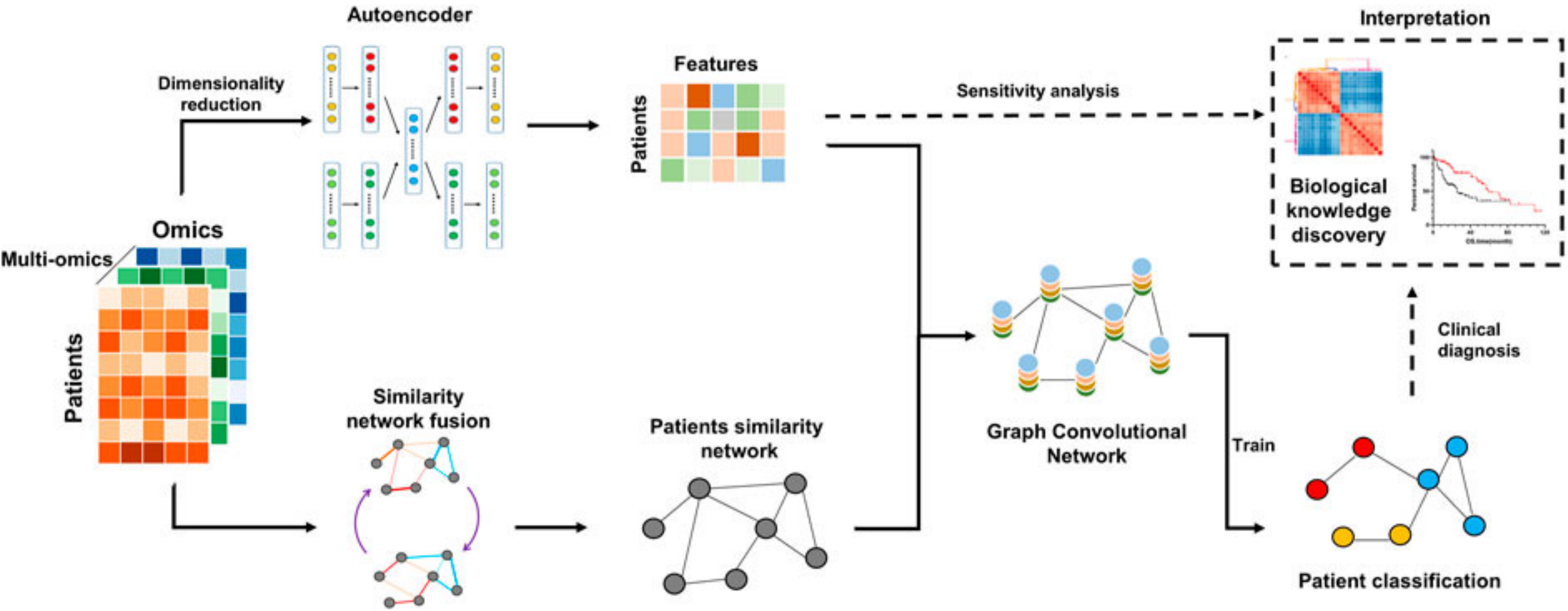
Graph convolutional network

Graph deep learning has recently emerged to incorporate graph structures into a deep learning framework. In particular, graph convolutional networks (GCNs) are able to classify unlabelled nodes in a network on the basis of both their associated feature vectors, as well as the network's topology, making it possible to integrate graph-based data with feature vectors in a natural way.
Advances in feature interpretation strategies for deep neural networks make it also possible to investigate the decision of such methods, leveraging a deep understanding of the underlying data

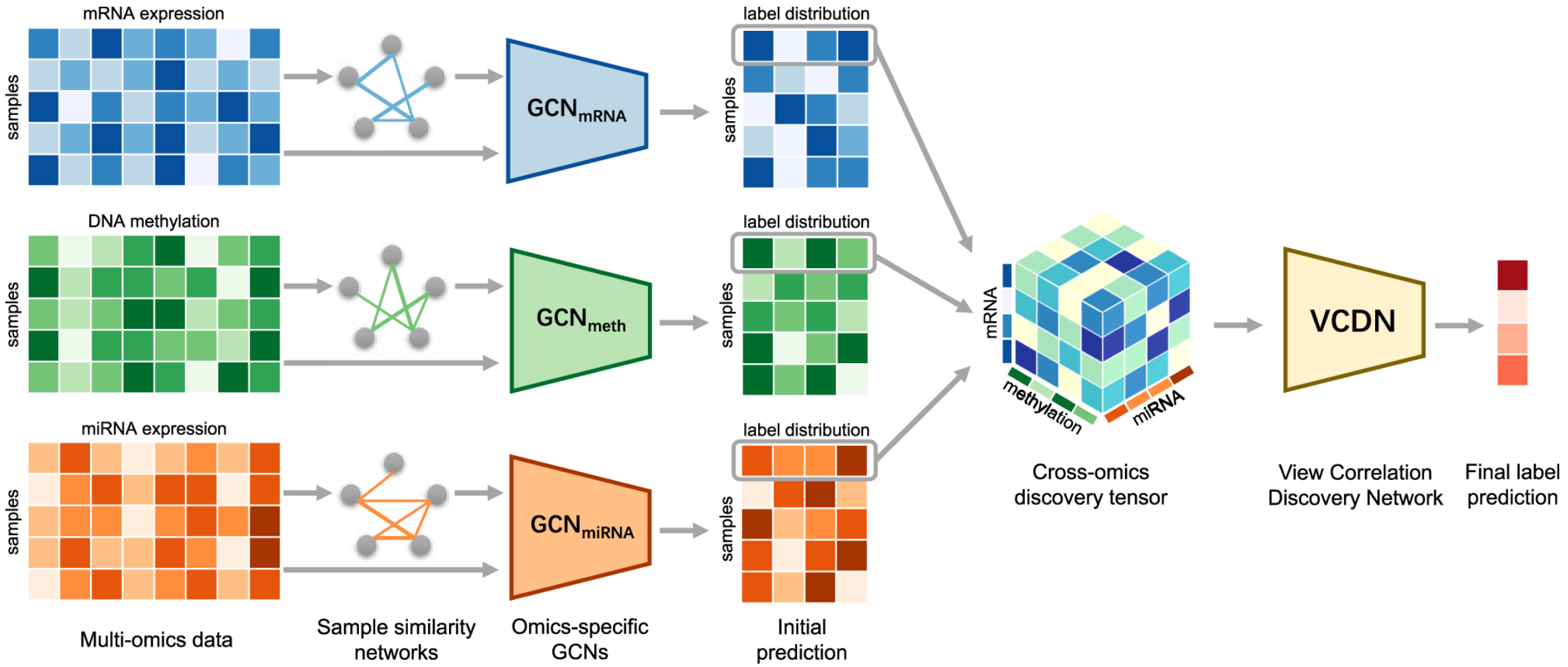
EMOGI


Multi-omics Data Collection
HNSC
THCA
ESCA
LUAD
LUSC
BRCA
UHC
STAD
KIRC/KIRP
COAD
BLCA
READ
UCEC
PRAD
CESC
16 Cancer Types
Mutations
CNAs
Promoter Methylation
Gene Expression
normal
tumor
ACGGACCATCT
ACGGACCCTCT
Protein-Protein interactions
TP53
PTEN
KRAS
MYC
APC
Breast
Lung
Kidney
Rectum
Stomach
Model Training
Model Explanation
= non-cancer gene
= unknown gene
= known cancer gene
Input Layer
Hidden Layer
Output Layer
= pred. non-cancer gene
= pred. cancer gene
gene A
gene B
gene C
gene D
Feature Contributions
Gene A
Gene B
Gene C
SNVs
Copy Number
DNA Meth.
Gene Expr.
Cancer types
Gene Stratification
1
2
3
4
Interaction Contributions
Gene A
Gene B
Gene C
Gene D
Gene E
Module Detection

MoGCN

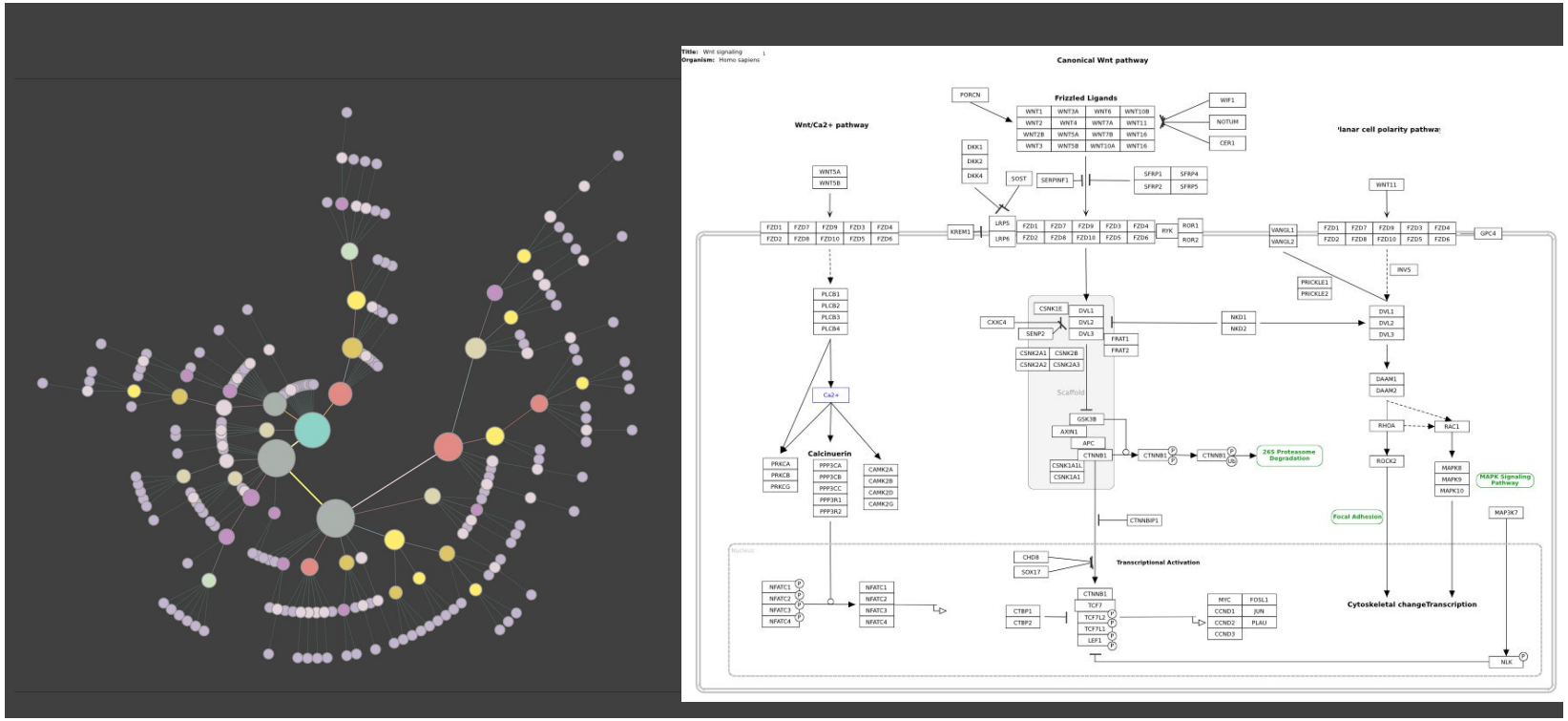

Autoencoder
Dimensionality reduction
Features
Patients
Sensitivity analysis
Interpretation
Biological knowledge discovery
Omics
Multi-omics
Patients
Similarity network fusion
Patients similarity network
Graph Convolutional Network
Clinical diagnosis
Train
Patient classification

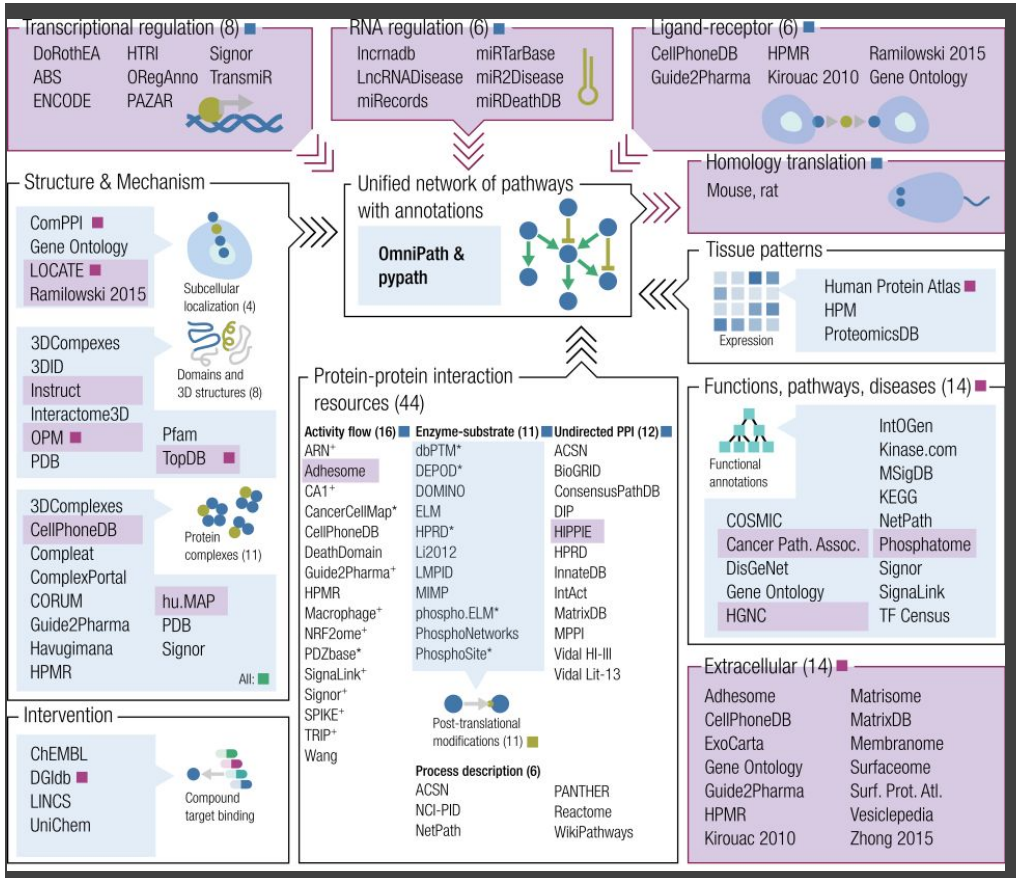
MOGONET


mRNA expression
samples
GCN$_{mRNA}$
label distribution
samples
DNA methylation
samples
GCN$_{meth}$
label distribution
samples
miRNA expression
samples
GCN$_{miRNA}$
label distribution
samples
mRNA
methylation
miRNA
VCDN
Multi-omics data
Sample similarity networks
Omics-specific GCNs
Initial prediction
Cross-omics discovery tensor
View Correlation Discovery Network
Final label prediction

Pathways as networks

These methods first construct networks by using pathways as sources of information to infer links among entities. Then each pathway network is independently analysed to obtain the summary statistics for each element. Finally, graph-based analyses are performed to calculate each pathway's P-value and network score.





Türei et al 2021

Pros:

- Database of databases
- Many interactions

Cons:

- Bias of all the various datasets
- Need to know what is the aim of your research

E.g. OmniPath, STRING , ConsensusPathDB

Omics as an image

Although these works have been devised to use embedding and conventional machine learning approaches, the use of deep learning on multi-omic data integration is still in its infancy.

Deep artificial neural networks are widely acknowledged for their ability to perform automatic feature extraction from raw data. Hence, a deep learning-based model has the capacity to develop a single-step phenotype prediction procedure where omics data would be directly used to classify the outcome without prior feature selection.

The absolute superiority of convolutional neural networks (CNNs) in image classification has been widely acknowledged but not yet sufficiently leveraged for analyzing multi-omics data.

Intuitively, CNN’s convolution operators extract local spatial features from an input image. The local information is then combined, via pooling aggregations, to higher-order special information from a large image region which would eventually be used to distinguish among different image types.

Transformers

MOT: a Multi-Omics Transformer for multiclass classification tumour types predictions