




Review

Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics

Pedro H. Godoy Sanches ^{1,†} , Nicolly Clemente de Melo ^{2,†}, Andreia M. Porcari ¹  and Lucas Miguel de Carvalho ^{3,*} 

¹ MS4Life Laboratory of Mass Spectrometry, Health Sciences Postgraduate Program, São Francisco University, Bragança Paulista 12916-900, SP, Brazil

² Graduate Program in Biomedicine, São Francisco University, Bragança Paulista 12916-900, SP, Brazil

³ Post Graduate Program in Health Sciences, São Francisco University, Bragança Paulista 12916-900, SP, Brazil

* Correspondence: lucas.miguel@usf.edu.br

† These authors contributed equally to this work.

Simple Summary: Recent high-throughput technologies such as transcriptomics, proteomics, and metabolomics have allowed progress in understanding biological systems at different levels of detail. Even so, it is necessary to integrate multiple omics data sets to achieve a comprehensive understanding of the subject under study. In this article, we review the methods used for integrating transcriptomics, proteomics, and metabolomics data and summarize them in three approaches: combined omics integration, correlation-based integration strategies, and machine learning integrative approaches. Our goal is to showcase the uses and limitations of each approach, allowing researchers to choose the more appropriate tool for each scenario to extract a comprehensive view of a biological system.



Citation: Sanches, P.H.G.; de Melo, N.C.; Porcari, A.M.; de Carvalho, L.M. Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics. *Biology* **2024**, *13*, 848. <https://doi.org/10.3390/biology13110848>

Academic Editors: Dario Di Silvestre, Min Zhao and Ruifeng Hu

Received: 30 June 2024

Revised: 19 July 2024

Accepted: 25 July 2024

Published: 22 October 2024

Abstract: With the advent of high-throughput technologies, the field of omics has made significant strides in characterizing biological systems at various levels of complexity. Transcriptomics, proteomics, and metabolomics are the three most widely used omics technologies, each providing unique insights into different layers of a biological system. However, analyzing each omics data set separately may not provide a comprehensive understanding of the subject under study. Therefore, integrating multi-omics data has become increasingly important in bioinformatics research. In this article, we review strategies for integrating transcriptomics, proteomics, and metabolomics data, including co-expression analysis, metabolite–gene networks, constraint-based models, pathway enrichment analysis, and interactome analysis. We discuss combined omics integration approaches, correlation-based strategies, and machine learning techniques that utilize one or more types of omics data. By presenting these methods, we aim to provide researchers with a better understanding of how to integrate omics data to gain a more comprehensive view of a biological system, facilitating the identification of complex patterns and interactions that might be missed by single-omics analyses.

Keywords: omics data; transcriptomics; metabolomics; proteomics; omics integration; multi-omics



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Omics data result from the use of large-scale instruments used in biology. They can contain measurements of different biomolecules, their functions, and interactions [1]. They have become an essential tool in modern biology and biomedicine [2,3]. Transcriptomics, proteomics, and metabolomics are three major omics fields that provide different types of biological information.

As an indirect measure of DNA activity [4], transcriptomics measures the expression levels of a set of RNA transcripts (mRNA, non-coding RNA, etc.) in a cell or tissue,

i.e., the transcriptome. Produced according to the instructions from mRNA, proteins and enzymes, typically > 2 kDa, are the functional products of genes and play several roles in cellular processes [5], being the macromolecule responsible for direct interactions among cells and tissues, besides maintaining the cellular structure [6]. Proteomics then focuses on the identification and quantification of a set of proteins, the *proteome* [7]. Smaller molecules (≤ 1.5 kDa), referred to as metabolites, are intermediate or end products of metabolic reactions and regulators of metabolism [8], but are not analyzed with the instrumental methodologies used in proteomics. Metabolomics comprehensively analyzes these molecules, trying to describe and quantify the molecular composition of a sample (the metabolome). Additionally, a big branch of metabolomics is the study of the lipidic composition of a sample, its lipidome, and for that, the “Lipidomics” term is used [8]. By indirectly measuring how a gene is acting, transcriptomics covers the upstream processes of metabolism, while proteomics is the intermediate step, defining protein structure and biocommunication. Metabolomics focuses on the regulators and the ultimate mediators of a metabolic process (usually smaller molecules not more than 1.5 kDa) [9]. Together, these omics technologies offer a comprehensive and streamlined view of biological processes.

Integrating multiple omics data sets is a challenging but necessary task to fully understand complex biological systems. Data integration can provide novel biological insights and reveal previously unknown relationships between different molecular components. Moreover, it can help identify biomarkers and therapeutic targets for various diseases. Several methods have been developed for integrating omics data, including correlation-based approaches, machine learning algorithms, and network-based analyses [4,10–12]

In this article, we will review and discuss different methods for integrating transcriptomics, proteomics, and metabolomics data. We will discuss the strengths and limitations of each method and provide examples of their applications in various biological contexts. Also, we will cite strategies and articles that use these omics in machine learning-based studies. By doing so, we hope to contribute to the development of effective strategies for omics data integration and pave the way for new discoveries in biomedical and biotechnology research.

2. Methods for Integrating Multi-Omics Data

Integrating omics data from several domains is critical for gaining complete knowledge of biological systems. To uncover critical regulatory pathways and networks, transcriptomics data can be combined with proteomics or metabolomics data. Many methodologies for integrating transcriptomics and proteomics data have been developed, including correlation-based approaches and pathway and co-expression analysis. Merging proteomics data with metabolomics data is also a potential strategy for biomarker development and disease diagnosis, since it can uncover alterations in metabolic pathways linked to disease states. Co-expression analysis and network-based techniques have been utilized successfully in the integration of transcriptomics and metabolomics. Overall, integrating omics data can be a significant tool for deciphering complicated biological systems and discovering novel treatment targets [3,10,12–16].

We divide the methods of omics integration into three major approaches: combined omics integration, correlation-based integration strategies, and machine learning integrative approaches (Figure 1). Combined omics integration approaches attempt to explain what occurs within each type of omics data in an integrated manner, generating independent data sets. Correlation-based strategies apply correlations between the generated omics data and create data structures to represent these relationships, such as networks. Finally, machine learning strategies utilize one or more types of omics data, potentially incorporating additional information inherent to these data sets, to comprehensively understand responses at the classification and regression levels, particularly in relation to diseases. These methods enable a comprehensive view of biological systems, facilitating the identification of complex patterns and interactions that might be missed by single-omics analyses. By leveraging these integrative approaches, researchers can achieve deeper insights into the

molecular mechanisms underlying health and disease, ultimately aiding in the discovery of novel biomarkers and therapeutic targets.

REVIEWING APPROACHES

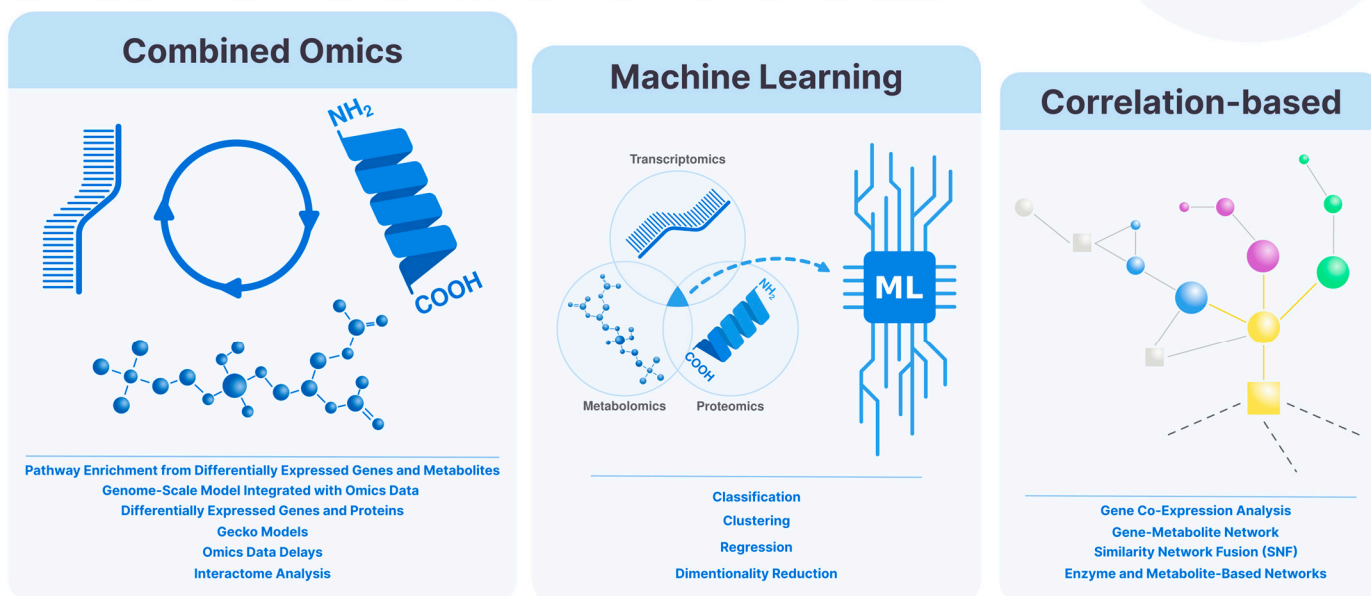
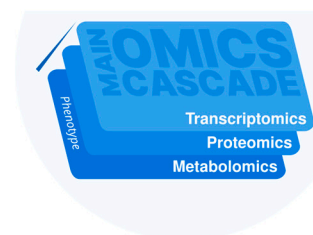


Figure 1. Strategies for integrating omics data. Methods are based on correlation-based approaches, which identify associations between different types of data; machine learning algorithms, which can predict outcomes and identify patterns across data sets; and combined individual approaches, which map the interactions and relationships between molecular components.

In this article, we focus on non-approximate strategies for cell-to-cell communication, such as single-cell RNA-seq (scRNA-seq), due to their level of resolution and ability to detect communication between individual cells. Bulk analysis assumes that cells are identical and can model the exchange between cells and the environment [17]. However, we will provide excellent recent reviews that present strategies for integrating multi-omics data from single-cell data [17–25].

2.1. Correlation-Based Methods

Correlation-based strategies involve applying statistical correlations between different types of generated omics data to uncover and quantify relationships between various molecular components. These methods, summarized in Table 1, then create data structures, such as networks, to visually and analytically represent these relationships. By mapping out these correlations, researchers can identify patterns of co-expression, co-regulation, and functional interactions that occur across different omics layers. This approach allows for the detection of complex interdependencies and the construction of interaction networks that highlight key molecules and pathways involved in biological processes.

Table 1. Summary of methods and strategies for integrating transcriptomics, proteomics, and metabolomics data using the correlation-based approach.

Integration Approach	Strategy or Method	Possible Omics Data	Main Idea
Correlation-based	Gene co-expression analysis	Transcriptomics and metabolomics	Identify co-expressed gene modules with metabolite similarity patterns under the same biological conditions
	Gene–metabolite network	Transcriptomics and metabolomics	Perform a correlation network of genes and metabolites
	Similarity Network Fusion	Transcriptomics, proteomics, and metabolomics	Builds a similarity network for each omics data separately, and subsequently, all networks are merged, and the edges with high associations in each omics network are highlighted
	Enzyme and metabolite-based network	Proteomics and metabolomics	Identify a network of protein–metabolite or enzyme–metabolite interactions using genome-scale models or pathways databases

2.1.1. Gene Co-Expression Analysis Integrated with Metabolomics Data

Co-expression analysis is a powerful approach for identifying genes with the same expression pattern that may participate in the same biological pathways or have the same biological function [26,27]. One strategy for integrating transcriptomics and metabolomics data is to perform a co-expression analysis on transcriptomics data and identify gene modules that are co-expressed. These modules can then be linked to metabolites identified from metabolomics data to identify metabolic pathways that are co-regulated with the identified gene modules [28–34].

To further understand the relationship between co-expressed genes and metabolites, the correlation between metabolite intensity patterns and the eigengenes of each co-expression module can be calculated. Eigengenes are representative expression profiles for each module that summarize the overall expression pattern of the genes within the module. By correlating these eigengenes with metabolite intensity patterns, it is possible to identify which metabolites are most strongly associated with each co-expression module [35–39]. Additionally, you can use the normalized metabolomics data directly in Weighted Correlation Network Analysis (WGCNA) [27], conducting a module–sample relationship analysis (in this case, module–metabolite relationship), and identify relationships between module eigengenes and metabolite intensity.

This approach can provide important insights into the regulation of metabolic pathways and the formation of specific metabolites. For example, if a particular co-expression module is strongly correlated with the production of a specific metabolite, it suggests that the genes within the module are involved in regulating the metabolic pathway leading to that metabolite. By combining transcriptomics and metabolomics data, it is possible to identify key genes and metabolic pathways involved in specific biological processes or disease states and to develop targeted interventions to modulate these processes, taking into account that the biological conditions of both omics analyses must be the same.

2.1.2. Gene–Metabolite Network

A gene–metabolite network is a visualization of the interactions between genes and metabolites in a biological system. Generating and analyzing these networks involves collecting gene expression and metabolite abundance data, integrating the data, constructing the network, analyzing it, and interpreting the results. Gene–metabolite networks can help

identify key regulatory nodes and pathways that are involved in metabolic processes and can be used to generate hypotheses about the underlying biology.

To generate a gene–metabolite network, researchers must first collect gene expression and metabolite abundance data from the same biological samples. These data are then integrated using Pearson correlation coefficient (PCC) analysis or other statistical methods to identify genes and metabolites that are co-regulated or co-expressed [40–42]. For example, Nikiforova et al., 2005, exhibited a systematic procedure to construct a gene–metabolite network based on the profiles of transcripts and metabolites [43]. Also, gene–metabolite networks are constructed using visualization software, such as Cytoscape [44] or igraph [45], with genes and metabolites represented as nodes in the network and connected with edges that represent the strength and direction of their interactions.

Once the network is constructed, it can be analyzed using network analysis tools to identify key regulatory nodes and pathways that are involved in metabolic processes [46–50]. Furthermore, a gene–metabolite network could be constructed with genes and metabolites specifically deregulated, to focus on the process that is modulated by each biological condition. Validation and interpretation of the network can then be performed by comparing it to known metabolic pathways and regulatory networks and using pathway analysis, gene ontology analysis, and functional enrichment analysis to identify enriched pathways and processes. Afterwards, you may select genes present in metabolic pathways or biological processes and obtain the mRNA levels of these genes by real-time qRT-PCR.

One interesting way of constructing a metabolite–gene network is based on the *corto* package developed by Mercatelli D. et al. (2020) [51]. First, the metabolite and gene data are combined into a single matrix, with only the metabolites designated as “centroids” or hubs in the resulting co-occurrence network. To determine significant edges in the network, you may set the minimum Pearson correlation coefficient p -value to 0.05 or a chosen cut-off. To test the significance of each edge in the network, it is necessary to conduct 100 or more bootstraps. The R code snippet below describes the implementation of this method, with “full_matrix” as the input matrix, and “metabolites” as a vector containing the metabolite names, as proposed by Cavicchioli M.V. et al. (2022) [52].

2.1.3. Similarity Network Fusion (SNF)

Similarity Network Fusion (SNF) [53] is a computational approach for integrating various data types [54,55]. In essence, SNF merges diverse measurements, such as mRNA expression, protein abundance, miRNA expression, metabolomics, clinical data, questionnaires, and image data, among others, for a given set of samples, like patients. Essentially, it starts by creating a sample similarity network for each data type and then combines these networks iteratively using a unique network fusion technique. By operating in the sample network space, SNF effectively circumvents issues related to different scales, collection biases, and noise in various data types.

For example, let us say that we have transcriptomics data from N patients (treated vs. control) and metabolomics data from the same patients. The SNF algorithm constructs a similarity network for each omics data set separately, where each node represents a patient, and the edge intensity connecting two patients indicates the level of association based on that omics data set. Subsequently, both networks are merged, and edges with high associations in each omics network are highlighted. Clustering can be performed to identify phenotypic associations among patients and cancer types, for instance. Furthermore, patient information (such as weight, survival time, among others) can be added to the nodes to enhance information visualization. Looking at the final network and examining each edge and its source (whether it is supported by one or more omics datasets), we can see how each omics dataset supports each group of patients. This allows for comparisons between different diseases. We highlight that there are limitations in the SNF method, such as dependence on pre-processing, noise in the input data, and computational complexity.

2.1.4. Enzyme and Metabolite-Based Network

In this approach, we utilize the available metabolomics and proteomics data to identify a network of protein–metabolite or enzyme–metabolite interactions using genome-scale models or pathways databases and then combine it with omics data such as fold changes to visualize the enriched pathways. There are two main strategies to define such networks, based on either proteomics or metabolomics data in the first step.

Let us start with the proteomics-based strategy. First, we identify the reactions in which the identified proteins participate in the genome-scale model, which can be in the GPR in AND or OR operator. Then, we determine the metabolites that are consumed (met-c) and produced (met-p) in those reactions. This information is used to construct a protein–metabolite network, where the protein is derived from the identification of the GPR, and the metabolites are the reactions that the protein is involved in (met-c and met-p). Additional details such as compartments, reactions ID, and fold changes can be included to enhance the network.

The second strategy relies on metabolomics data, which require standardization of the metabolite names or IDs in the database. Using the identified metabolites, along with the genome-scale model, we identify all the reactions in which those metabolites participate and annotate the proteins belonging to the GPR, in AND or OR operator. Then, we assemble the protein–metabolite network using this interaction information.

Both strategies are based on omics data with the addition of genome-scale models as the foundation for assembling the network structure. Although the associations from the Kyoto Encyclopedia of Genes and Genomes (KEGG) can be used directly, we expand to a whole genome-scale model. The first strategy yields a larger network due to the greater number of identifications that proteomics data provide compared to metabolomics data.

MetaBridge [56] is a powerful web-based tool that aims to integrate metabolomics with proteomics data. The tool achieves this by utilizing data from two key databases: the MetaCyc metabolic pathway database [57] and KEGG. MetaBridge maps metabolite compounds to interacting upstream or downstream enzymes in enzymatic reactions and metabolic pathways and generates a list of enzymes that can be integrated with proteomics or transcriptomics data using protein–protein interaction (PPI) networks. The resulting PPI network can be used for integrative multi-omics analyses, allowing users to identify key proteins and pathways that are differentially regulated across the data sets and also integrate the fold change per protein and select specific submodules from it. By providing a user-friendly interface and detailed protocols, MetaBridge makes it easy to perform integrative multi-omics analyses and gain insights into complex biological systems.

2.2. Combined Omics Approaches

Combined omics integration approaches, as summarized in Table 2, seek to explain the phenomena occurring within each type of omics data through a comprehensive and integrated framework. These methods generate independent data sets for each omics layer, such as transcriptomics, proteomics, and metabolomics, and then combine them to provide a full view of a biological system. This integrated analysis can reveal insights that would not be apparent when examining each omics data set in isolation, thus enabling the identification of novel interactions, pathways, and regulatory mechanisms that drive biological functions and disease states. Here, we will discuss different strategies and the necessary care for integrating omics data.

Table 2. Summary of methods and strategies for integrating transcriptomics, proteomics, and metabolomics data using the combined omics approach.

Integration Approach	Strategy or Method	Possible Omics Data	Main Idea
Combined omics	Pathway enrichment from differentially expressed genes and metabolites	Transcriptomics and metabolomics	Identify pathways enriched in both types of omics data and perform a post-analysis with these results
	Integrating genome-scale models with omics data	Transcriptomics and metabolomics	Integrate metabolic and transcriptomic data to create content-specific models and perform specific metabolic simulations
	Gecko models	Proteomics and metabolomics	Integrate proteomics data into an enzyme model, which can be validated with metabolomics data under the same biological conditions.
	Differentially expressed genes and proteins	Transcriptomics and proteomics	Identify similarities between the lists of differentials in the two omics data sets
	Observing delays between omics data	Transcriptomics and proteomics	Identify whether there is a temporal delay in the acquisition of omics data based on gene expression and protein abundance
	Interactome analysis	Transcriptomics and proteomics	Identify functional relationships between different proteins and genes using interactome databases and fold-change values

2.2.1. Pathway Enrichment from Differentially Expressed Genes and Metabolites

One strategy for integrating transcriptomics and metabolomics data is to perform KEGG enrichment analysis on differentially expressed genes (DEGs) identified from transcriptomics data and then link the enriched pathways to the metabolites identified from metabolomics data. This approach can reveal the metabolic pathways that are most affected by changes in gene expression and provide insights into the underlying biological mechanisms. Additionally, fold-change information from DEGs can be integrated with metabolomics data to identify metabolites that are significantly changed in abundance and may play a key role in the observed changes in gene expression.

DEG pathway enrichment may be performed using the clusterProfiler [58] or GSEAPy [59] packages, for example. One way to convert a gene name to Entrez Gene ID, which provides unique integer identifiers for genes and other loci, or KEGG ID to perform the enrichment through the clusterProfiler is through the genekitr package using the *transId()* function with the argument *transTo = "entrez"*. The MetaboAnalyst platform [60] may be used to enhance metabolic pathways with identified metabolites. Furthermore, information from gene expression, such as fold changes, and metabolite intensity analysis can be integrated into metabolic pathway figures using the Pathview package [61].

Additionally, it is possible to analyze the correlation between the abundance of metabolites and the expression of genes or transcripts across various biological conditions using integrated pathway analysis. To perform this analysis, it is necessary to include multiple time points or a large number of biological conditions. The association cut-off for this analysis is based on a p -value $< \alpha$ and Pearson coefficients $> \beta$, which can be plotted to determine the extent to which relevant metabolites are correlated with relevant mRNA transcripts. To carry out this analysis, Multi-Omics Factor Analysis (MOFA) [62] is a suitable approach for determining the degree to which changes in metabolite abundance and mRNA expression variables are related [63–65]. In addition, correlation analysis can be applied to

determine the degree of relationship between metabolite intensities/concentrations and gene expression, but this must be carried out with normalized (post-processing) data. Data normalization in the post-processing of metabolomics data is usually provided, especially if the statistical analyses are conducted in MetaboAnalyst.

Also, if a multivariate analysis is chosen, such as Principal Component Analysis (PCA), it is important to understand the factors that can lead to failure of the analysis of variance. We point out that data integration through concatenation can become complex when the data sets to be merged differ significantly in size. Not only do metabolomics and transcriptomics data sets differ significantly in size, but they are also generated using vastly different technologies [66]. This means that the data sets have distinct structures, unique patterns of expected values, dissimilar distributions of underlying noise, and varying levels of variance. With that in mind, the best proposal is to carry out a PCA biplot for the metabolites and another for the genes, identify the most correlated metabolites and genes either with each biological condition or with each principal component (PC1 or PC2), and then associate this result with the pathway that both belong to.

2.2.2. Integrating Genome-Scale Models with Metabolomics and Transcriptomics Data

Integrating transcriptomics data, genome-scale models (GEMs), and metabolomics data can provide a comprehensive understanding of cellular metabolism and its regulation [67]. One approach to integrating these data types is to generate condition-specific models (CSMs) that incorporate the transcriptomics data and use them to simulate metabolic fluxes in different situations [68–72]. Here, we describe a strategy for generating CSMs using GEMs and transcriptomics data and then integrating these models with metabolomics data.

The first step in this strategy is to generate a GEM that represents the metabolic network of the organism of interest. The GEM should include all the reactions and metabolites involved in cellular metabolism and should be curated and validated using experimental data. Once the GEM is generated, it can be used to simulate metabolic fluxes under different conditions.

Next, transcriptomics data can be used to generate CSMs that incorporate the expression levels of genes under different conditions. This can be achieved using constraint-based modeling techniques such as flux balance analysis (FBA) [73] or parsimonious FBA (pFBA) [74]. CSMs can be generated by constraining the fluxes through the reactions in the GEM based on the expression levels of the corresponding genes. MEWpy [75] is a package that covers a wide range of metabolic and regulatory modeling approaches, as well as phenotype simulation and Computational Strain Optimization (CSO) algorithms. This makes it a useful tool for generating transcriptome-based simulations with FBA or pFBA.

Once the CSMs are generated, they can be used to simulate metabolic fluxes in different conditions and predict the metabolic phenotypes of a cell. However, to fully understand the regulation of cellular metabolism, it is important to also integrate metabolomics data into the models.

One way to integrate metabolomics data into CSMs is to use the former to constrain the fluxes through the exchange reactions that correspond to the measured metabolites. It is essential to normalize the exchange reactions based on metabolite concentrations in a metabolic model. For example, if an organism cannot consume the entire concentration of a metabolite in 24 h, you can estimate the upper exchange flux as follows: $[\text{Concentration}]/(1 \text{ gDW} \times 24 \text{ h})$. Moreover, one possibility is using linear programming techniques to find the flux distributions that are consistent with both the transcriptomics and the metabolomics data, and this approximation is used to insert the exchange reaction fluxes. We highlight that there are limitations in the method, such as noise and data variability, differences in temporal scales between the omics data, incomplete annotations in the metabolic model, and limitations in experimental validation.

2.2.3. Gecko Models

In recent years, the integration of omics data into genome-scale metabolic models (GEMs) has become a powerful approach for exploring the relationship between genotype and phenotype for different organisms [76–78]. GEMs allow for the prediction of metabolic behavior and can be used to design experiments and engineer biological systems. One such tool that has enabled the integration of proteomics data into GEMs is the GECKO toolbox [79]. GECKO models take into account enzyme and proteomics constraints to study phenotypes that are constrained by protein limitations [80–83]. With the GECKO toolbox, it is possible to generate enzyme-constrained models (ecModels) for a variety of organisms, including budding yeasts such as *Saccharomyces cerevisiae* and humans, as well as build your own model. These models can be used to study the long-term adaptation of organisms to stress factors and nutrient-limited conditions.

The GECKO models simplify the process of limiting the metabolic fluxes in any GEM that contains enzymatic data, reducing the variability of constraint-based modeling results and improving predictions. This approach is executed by representing enzymes as entities with limited capacities in the corresponding reactions in the model, thereby extending the genome-scale modeling. In traditional genome-scale modeling, a stoichiometric matrix is defined that represents the whole metabolism, with columns indicating each reaction's stoichiometry, and rows indicating the mass balance for each metabolite. With GECKO, this approach is expanded by adding new rows to the matrix to represent the enzymes and new columns to represent each enzyme usage. Kinetic information, in the form of k_{cat} values, is included as stoichiometric coefficients to convert the metabolic flux in mmol/gDWh to the required enzyme usage in mmol/gDW. The protein level is included as an upper bound for each enzyme usage, ensuring that the desired constraint on each flux is respected.

To create your own GECKO model, k_{cat} values of reactions, the molecular weight of proteins, and protein activity information will be required and can be directly changed and included in your GEM, because different metabolic groups have different k_{cat} values and molecular weight distribution [80]. All molecular and enzymatic parameters could be automatically retrieved from the BRENDA database [84] and/or the UNIPROT database [85].

The GECKO toolbox is dependent on MATLAB and other packages. There is an option in Python to work with GECKO models using the MEWpy package. Both require ecModels and normalized proteomics data.

2.2.4. Strategies for Integrating Proteomics and Transcriptomics Data

The integration of proteomics and transcriptomics data has become a crucial part of modern Systems Biology research. The combination of these two omics data types can provide a more comprehensive understanding of biological systems. Here, we will elaborate on the three strategies for integrating proteomics and transcriptomics data.

One of the challenges in integrating transcriptomics and proteomics data is the difficulty in obtaining the same sets of differentially expressed genes and differentially expressed proteins. This is often due to differences in the timing of sample collection for RNA and protein analysis. RNA samples are typically collected at the transcription stage, whereas protein samples are collected at the translation stage. As a result, there can be significant differences in the expression patterns of genes and proteins between these two stages. Furthermore, RNA and protein stability can also differ, which can further complicate the comparison between these two types of data.

To overcome this challenge, it is important to carefully plan the experimental design and sample collection protocols. Ideally, the samples for both RNA and protein analysis should be collected at the same time point and under the same conditions. If this is not possible, researchers can try to account for the differences between RNA and protein data by using statistical methods to normalize the data or by applying machine learning algorithms to identify patterns of expression that are consistent across both data sets.

Differentially Expressed Genes and Proteins

The first strategy for integrating proteomics and transcriptomics data is Venn diagram analysis or Jaccard index calculation. In this strategy, differentially expressed genes (DEGs) and differentially abundant proteins (DAPs) are identified in the same biological conditions from both types of omics data. A Venn diagram is then used to identify overlapping genes and proteins, which can provide insights into the mechanisms underlying a certain biological condition. This strategy can be particularly useful for identifying key pathways or processes that are regulated by both proteins and transcripts. However, it is important to ensure that the transcription and translation processes are aligned at the time of collection to avoid false positives.

There are articles that show that there is not a great intersection between DEGs and DAPs [86–95], and others showing a successful intersection [96–100]. This discrepancy is based on the following: (i) induced and repressed proteins behaving differently, revealing regulatory and kinetic differences in protein synthesis and turnover; (ii) taking into account the transcription–translation delay when comparing protein and mRNA levels during dynamic adaptation; (iii) protein variation being mainly influenced by mRNA concentration in a new steady state [99]. Furthermore, if there is a significant overlap between DEGs and DAPs or if there is a GO/KEGG enrichment common to both, using the intersection list is the most appropriate approach. If there are many differences, the most common strategy is to construct a Venn diagram of the GO processes and KEGG pathways to identify the similarities and differences.

Observing Delays between Omics Data

The second strategy for integrating proteomics and transcriptomics data is scatter plot analysis. This strategy involves plotting the log of the genes fold change by the log of the proteins fold change. By observing the scatter plot, we can identify whether there is a correlation between the proteomics and the transcriptomics data. A positive correlation in the scatter plot suggests agreement in data extraction and provides a better understanding of the mechanisms underlying the biological condition. Scatter plot analysis can also be useful for identifying genes or proteins that do not have a direct correlation between their expression and protein levels based on the problems as cited before.

When we scatter plot the log fold change from a differential genes expression test (\logFC_t) versus the log fold change of the protein levels (\logFC_p), we can observe different patterns that can provide insights into the agreement or disagreement between the transcriptomics and the proteomics data. In this case, the \logFC_t and \logFC_p values increase or decrease together (Figure 2A), meaning that the genes and proteins are co-regulated in the biological system. This pattern suggests that there is a good agreement between the transcriptomics and the proteomics data and the protein abundance changes can be explained by changes in gene expression. If you see a concentration of points above (Figure 2B) and below (Figure 2C) the 45-degree line in the scatterplot of \logFC_t versus \logFC_p , it means that there is a disagreement between the changes in gene expression and protein levels for some genes/proteins. In other words, the gene expression and protein levels of these genes/proteins do not show a consistent pattern across the biological condition studied. This could be due to various factors such as post-transcriptional regulation, protein stability, or differences in the sampling and processing methods between the two omics (transcriptomics and proteomics) data sets.

The R package ReactomeGSA [101] includes a function named “plot_correlations()” which generates a comparative scatter plot of transcriptomics and proteomics data, allowing for a quick assessment of the similarity between the two data sets at the pathway level.

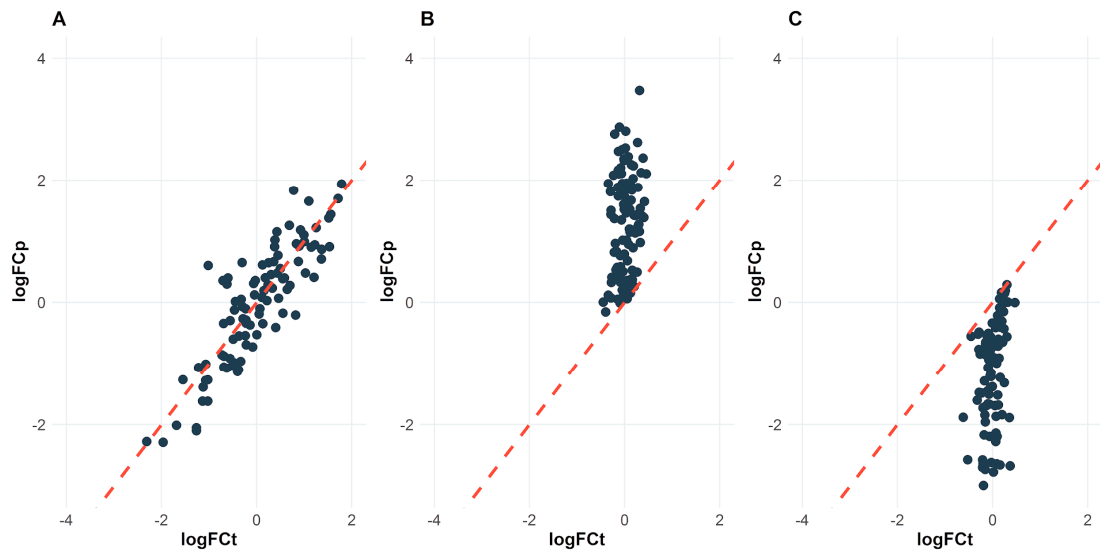


Figure 2. Scatter plot between the log fold change from a differential gene expression test (\logFCt) and the log fold change of the protein levels (\logFCp) in three scenarios: **(A)** high association between transcriptomics and proteomics data; **(B,C)** disagreement between the changes in gene expression and protein levels for some genes/proteins. The red dashed 45-degree line indicates the theoretical correspondence where changes in gene expression at the RNA-Seq level would be equally reflected at the protein level.

Interactome Analysis

The third strategy for integrating proteomics and transcriptomics data is interactome analysis. With this strategy, we could generate an interactome from differentially abundant proteins (DAPs) or differentially expressed genes (DEGs), identify submodules and hubs, and apply the fold change in gene expression to the interactome. This approach can help to identify functional relationships between different proteins and genes, providing a more comprehensive understanding of a biological system. Interactome analysis can also help to identify potential targets for further analysis, such as drugs or therapies.

First, from the list of DEGs, for example, we can generate an interactome using a protein–protein interaction (PPI) database, such as BioGRID [102] or STRING [103]. From the generated interactome, we can extract network metrics and identify hub genes and also submodules. With the application of the fold change in gene expression in the network, we can identify submodules with a predominance of a biological condition and apply biological enrichment to support such predominance.

Second, we can integrate this type of analysis based on co-expression analysis as well. Once the co-expression modules have been identified, it is possible to construct an interactome network that represents the interactions between the proteins encoded by these genes. The interactome can then be visualized and analyzed using network analysis tools such as Cytoscape, which allows for the identification of subnetworks, central nodes, and pathways that are enriched for the genes of interest, as well as of potential drug targets or biomarkers for diseases.

2.3. Machine Learning Methods Based on Omics Data

In this section, we will explore various methods that utilize omics data such as transcriptomics, proteomics, and metabolomics data for supervised and unsupervised machine learning algorithms (Figure 3). Supervised learning methods involve machine learning algorithms that use known data–outcome pairs as examples, whereas unsupervised methods operate on data sets without an outcome variable or prior knowledge of relationships between observations, dealing with unlabeled data. These strategies employ different methodologies and can involve the use of one, two, or three types of omics data. Additionally, numerous studies and reviews have extended this discussion to cover other

strategies and provide valuable tools to apply these methodologies with a greater emphasis on machine learning and multi-omics data [4,6,104–108].

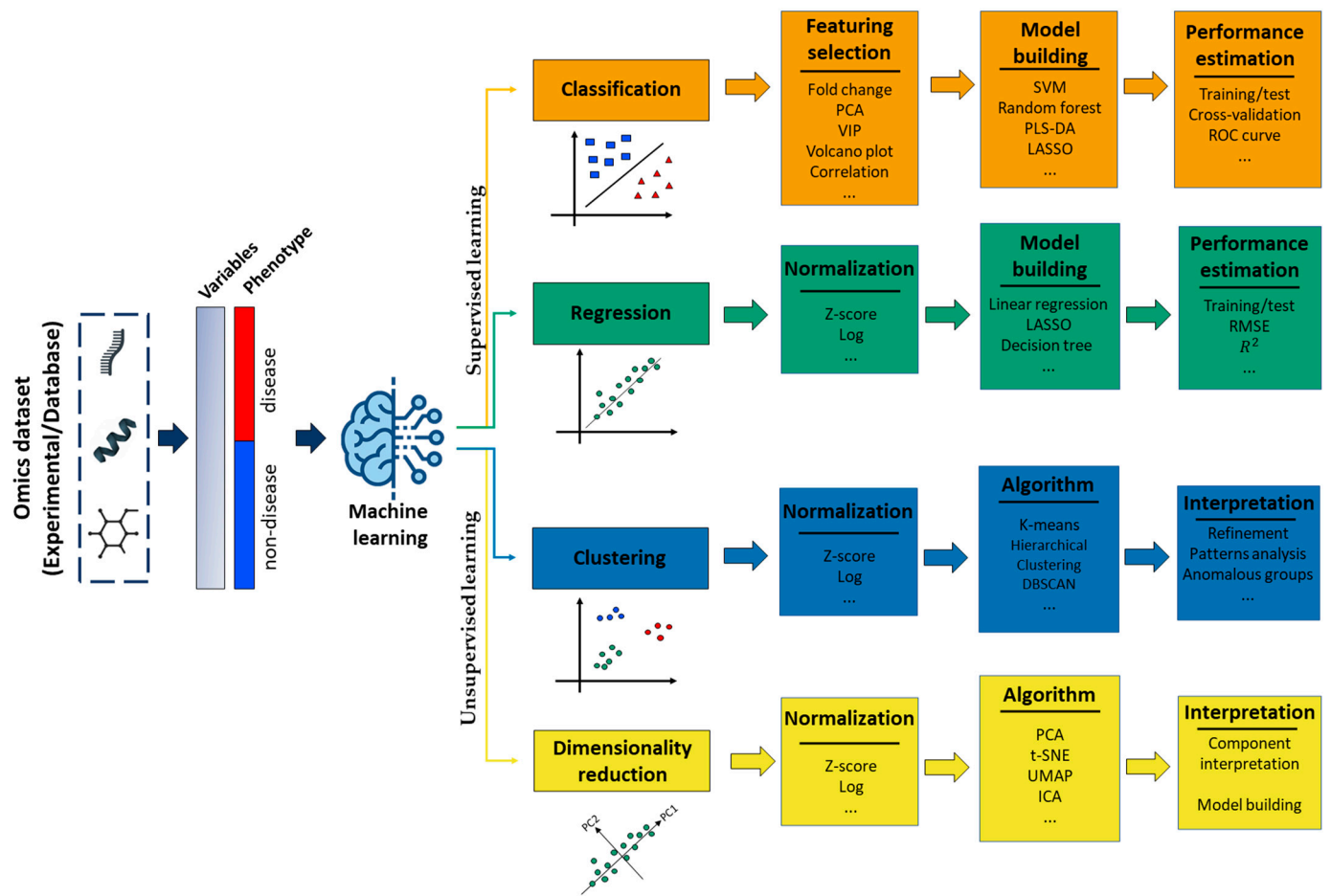


Figure 3. The pipeline illustrates the differences between supervised and unsupervised learning strategies applied to omics data. Legend: PCA: Principal Component Analysis; t-SNE: t-Distributed Stochastic Neighbor Embedding; UMAP: Uniform Manifold Approximation and Projection; ICA: Independent Component Analysis; SVM: Support Vector Machines; PLS-DA: Partial Least-Squares Discriminant Analysis; LASSO: Least Absolute Shrinkage and Selection Operator; RMSE: Root-Mean-Square Error; VIP: Variable Importance in Projection; ROC Curve: Receiver Operating Characteristic Curve.

2.3.1. Transcriptomics Data

Gene expression data can provide insights into the complex interplay between genes and cellular processes. With the advent of high-throughput technologies such as microarrays and RNA sequencing, it is now possible to generate large-scale gene expression data sets for a wide range of biological systems. In recent years, machine learning algorithms have emerged as a powerful tool for analyzing gene expression data. By leveraging the computational power of machine learning, researchers can uncover complex patterns and relationships within data that would be difficult or impossible to detect using traditional statistical methods, such as unsupervised machine learning methods.

In the literature, numerous studies have demonstrated the successful application of classification methods to predict cancer and cell types by utilizing gene expression data from microarray or bulk RNA-Seq data [109–115] and single-cell transcriptomics [116–122]. The process of employing gene expression data commences with the selection of suitable data sets, which can be accessed from databases like GEO [123], TCGA [124], and SRA [124]. Subsequently, pre-processing is carried out on the corresponding metadata for the clas-

sification algorithms, such as data from patients with or without cancer. Following this, the data set may or may not undergo a feature selection process followed by the application of a sampling technique to reach the final classification model, but a cross-validation step is essential. Each of these steps uses different data sets, pre-processing steps, model training and prediction algorithms, along with different k-fold cross-validation values, leading to varying values of accuracy, sensitivity (measuring the proportion of true positives accurately identified), and specificity (measuring the proportion of true negatives accurately identified).

It should be noted that the accuracy of machine learning classifiers using RNA-Seq data is dependent on various factors such as the type of sequencer used, the library preparation method, and the sample preparation technique. As a result, these classifiers exhibit varying levels of accuracy, with better performance observed at the transcript level compared to the gene level [125].

Biomedical researchers need to confirm the biological significance of the list or cluster of genes associated with a particular condition or developmental process that have been identified through comprehensive data analysis. To achieve this, they must evaluate the false-positive rate and conduct an autonomous biological validation. Northern blots and PCR-based methods are commonly employed to verify gene expression data, and these methods have the advantages of being able to screen through a large number of candidates relatively rapidly and perform quantitative measurements. Additionally, *in situ* hybridization and immunohistochemistry are used to determine the precise tissue in which the candidate genes are expressed [126]. Although these methods are not typically quantitative or high-throughput, they can be used to screen a large number of candidate genes and, in some cases, be performed in a quantitative manner.

2.3.2. Proteomics Data

With the advent of high-throughput technologies, such as mass spectrometry (MS) and protein microarrays, it is now possible to identify and quantify thousands of proteins in a single experiment. However, the sheer volume and complexity of proteomics data presents a challenge for traditional statistical and computational methods. Machine learning algorithms offer a promising solution for analyzing and interpreting these data, enabling researchers to extract meaningful information about protein function, interactions, and disease mechanisms. In this section, we will discuss how machine learning algorithms can be applied to proteomics data, including feature selection, classification, and clustering methods.

According to the literature, the machine learning algorithms used for proteomics data are based on retention time prediction, MS/MS spectrum prediction, the identification of peptides, biomarker identification, bias reduction during data processing, secondary structure prediction, protein toxicity prediction, protein function, and protein interactions [127–142]. Considering a comprehensive and reliable data integration, processing proteomics data and utilizing machine learning algorithms to accurately identify and analyze proteins require a consolidated pipeline. Due to the complexity of proteomics data, a comprehensive approach to data processing, normalization, quality control, feature extraction, and statistical analysis is necessary [7]. Without a consolidated pipeline, the analysis of proteomics data can be prone to errors and inconsistencies, leading to inaccurate results.

Identifying a biomarker from machine learning models based on proteomics data poses several challenges, including the need for validation in independent tests and the demonstration of clinical utility. These challenges are crucial in the translation of a promising biomarker candidate into a clinical tool. Validating the findings in independent tests is essential to ensure that the identified biomarker is reliable and reproducible. Demonstrating its clinical utility is also necessary to prove that the biomarker can effectively diagnose, monitor, or predict disease outcomes. These challenges highlight the importance of rigorous testing and validation before the implementation of any biomarker-based clinical assay.

2.3.3. Metabolomics Data

Machine learning has been applied to metabolomics data to identify biomarkers associated with diseases, understand metabolic pathways, contribute to the development of biotechnologies, and predict drug responses. PCA is one of the most widely used techniques to analyze metabolomics data, where metabolites are reduced into principal components that represent the majority of the variability in the data. PCA can identify metabolites that are significantly associated with different biological conditions, such as healthy and disease states, and can be used to cluster samples based on their metabolic profiles [143–148].

The variable in projection (VIP) score from the partial least-squares–discriminant analysis (PLS-DA) is another important method in metabolomics data analysis that can be used to identify relevant metabolites associated with a particular biological condition [149–155]. The VIP score is a measure of the contribution of each metabolite to the separation between two groups and is calculated by applying a supervised learning algorithm to the data, such as the PLS-DA algorithm. The MetaboAnalyst online platform is an example of a tool that allows researchers to apply VIP score analysis to their metabolomics data, as well as to perform other statistical and machine learning analyses.

Metabolomics data have been widely used in machine learning models to predict different types of cancer as well, as they provide valuable information about metabolic pathways altered in cancer cells. For instance, recent studies used metabolomics data in combination with machine learning algorithms to distinguish between different types of cancer, including lung, breast, non-Hodgkin’s lymphoma, and ovarian cancer, and non-cancer conditions, such as coronavirus disease (COVID-19), type-2 diabetes, acute myocardial ischemia, schizophrenia, and autism in relation to gestational age [156]. Moreover, the identification of biomarkers in several of these cited studies has the potential to improve disease diagnosis, treatment, and monitoring. It allows for the discovery of complex relationships between metabolites and biological conditions that may not be easily detected through traditional methods.

Transcriptomic data are still used to predict metabolite concentrations using machine learning models. Auslander N. et al. (2016) [157] demonstrated that the levels of a wide range of metabolites in breast cancer can be successfully predicted from the transcriptome. The authors developed a Support Vector Machine (SVM) classifier to identify reaction-gene-metabolite (RGM) triplets where the gene and metabolite involved in the same reaction showed a significant association, whether positive or negative.

Selecting the suitable ML algorithm (including linear regression, logistic regression, support vector machines, k-nearest neighbors, decision trees, random forests, neural networks, and deep learning) plays a critical role in the achievement of a metabolomics study. It is crucial for researchers to be knowledgeable about the advantages of various ML approaches and to choose the most appropriate one based on their requirements to obtain accurate and easily understandable results [156,158].

3. Conclusions

In conclusion, integrating multiple omics data types is a powerful approach that can provide a more comprehensive understanding of biological systems. Transcriptomics, proteomics, and metabolomics offer complementary views of biological processes at the RNA, protein, and metabolite levels, respectively. By combining these data types, researchers can gain insights into complex biological phenomena that may not be possible with any single omics data type alone.

There are several strategies for integrating omics data, including co-expression analysis, pathway analysis, and network analysis. Each strategy has its strengths and weaknesses, and the choice of the approach will depend on the specific research question being addressed.

Overall, the integration of omics data is a rapidly evolving field, with new methods and tools being developed to address the challenges of analyzing and interpreting large

and complex data sets. As technology continues to advance, the integration of omics data is likely to become even more important for understanding molecular mechanisms in biology.

In this article, we reviewed several methods for integrating omics data and provided examples of their application in various biological contexts. Moreover, we explored the applications of omics data in machine learning studies. We hope that this review will inspire further research in this field and lead to new insights into the complex interplay between genes, proteins, and metabolites in living systems.

Author Contributions: P.H.G.S.: conceptualization, visualization, writing—original draft, writing—review and editing. N.C.d.M.: conceptualization, writing—original draft, writing—review and editing. A.M.P.: conceptualization, writing—review and editing. L.M.d.C.: conceptualization, visualization, supervision, writing—original draft, writing—review and editing, project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financed by the São Paulo Research Foundation (FAPESP) through grant 2022/14179-1.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors have not declared any conflicts of interest.

References

- Hasin, Y.; Seldin, M.; Lusis, A. Multi-Omics Approaches to Disease. *Genome Biol.* **2017**, *18*, 83. [[CrossRef](#)]
- Hasanzad, M.; Sarhangi, N.; Ehsani Chimeh, S.; Ayati, N.; Afzali, M.; Khatami, F.; Nikfar, S.; Aghaei Meybodi, H.R. Precision Medicine Journey through Omics Approach. *J. Diabetes Metab. Disord.* **2022**, *21*, 881–888. [[CrossRef](#)] [[PubMed](#)]
- Karczewski, K.J.; Snyder, M.P. Integrative Omics for Health and Disease. *Nat. Rev. Genet.* **2018**, *19*, 299–310. [[CrossRef](#)] [[PubMed](#)]
- Picard, M.; Scott-Boyer, M.P.; Bodein, A.; Périn, O.; Droit, A. Integration Strategies of Multi-Omics Data for Machine Learning Analysis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3735–3746. [[CrossRef](#)] [[PubMed](#)]
- Rožanova, S.; Barkovits, K.; Nikolov, M.; Schmidt, C.; Urlaub, H.; Marcus, K. Quantitative Mass Spectrometry-Based Proteomics: An Overview. *Methods Mol. Biol.* **2021**, *2228*, 85–116. [[CrossRef](#)] [[PubMed](#)]
- Reel, P.S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnol. Adv.* **2021**, *49*, 107739. [[CrossRef](#)] [[PubMed](#)]
- de Carvalho, L.M.; Borelli, G.; Camargo, A.P.; de Assis, M.A.; de Ferraz, S.M.F.; Fiamenghi, M.B.; José, J.; Mofatto, L.S.; Nagamatsu, S.T.; Persinoti, G.F.; et al. Bioinformatics Applied to Biotechnology: A Review towards Bioenergy Research. *Biomass Bioenergy* **2019**, *123*, 195–224. [[CrossRef](#)]
- Wang, R.; Li, B.; Lam, S.M.; Shui, G. Integration of Lipidomics and Metabolomics for In-Depth Understanding of Cellular Mechanism and Disease Progression. *J. Genet. Genom.* **2020**, *47*, 69–83. [[CrossRef](#)] [[PubMed](#)]
- Idle, J.R.; Gonzalez, F.J. Metabolomics. *Cell Metab.* **2007**, *6*, 348–351. [[CrossRef](#)]
- Athieniti, E.; Spyrou, G.M. A Guide to Multi-Omics Data Collection and Integration for Translational Medicine. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 134–149. [[CrossRef](#)]
- Budzinski, I.G.F.; de Moraes, F.E.; Cataldi, T.R.; Franceschini, L.M.; Labate, C.A. Network Analyses and Data Integration of Proteomics and Metabolomics from Leaves of Two Contrasting Varieties of Sugarcane in Response to Drought. *Front. Plant Sci.* **2019**, *10*, 446557. [[CrossRef](#)] [[PubMed](#)]
- Dimitrakopoulos, C.; Hindupur, S.K.; Hafliger, L.; Behr, J.; Montazeri, H.; Hall, M.N.; Beerenwinkel, N. Network-Based Integration of Multi-Omics Data for Prioritizing Cancer Genes. *Bioinformatics* **2018**, *34*, 2441–2448. [[CrossRef](#)] [[PubMed](#)]
- Misra, B.B.; Langefeld, C.; Olivier, M.; Cox, L.A. Integrated Omics: Tools, Advances and Future Approaches. *J. Mol. Endocrinol.* **2019**, *62*, R21–R45. [[CrossRef](#)] [[PubMed](#)]
- Cambiaghi, A.; Ferrario, M.; Masseroli, M. Analysis of Metabolomic Data: Tools, Current Strategies and Future Challenges for Omics Data Integration. *Brief. Bioinform.* **2017**, *18*, 498–510. [[CrossRef](#)] [[PubMed](#)]
- Bersanelli, M.; Mosca, E.; Remondini, D.; Giampieri, E.; Sala, C.; Castellani, G.; Milanese, L. Methods for the Integration of Multi-Omics Data: Mathematical Aspects. *BMC Bioinform.* **2016**, *17*, 167–177. [[CrossRef](#)] [[PubMed](#)]
- Fukushima, A.; Kusano, M.; Redestig, H.; Arita, M.; Saito, K. Integrated Omics Approaches in Plant Systems Biology. *Curr. Opin. Chem. Biol.* **2009**, *13*, 532–538. [[CrossRef](#)] [[PubMed](#)]
- Hrovatin, K.; Fischer, D.S.; Theis, F.J. Toward Modeling Metabolic State from Single-Cell Transcriptomics. *Mol. Metab.* **2022**, *57*, 101396. [[CrossRef](#)] [[PubMed](#)]

18. Sun, Y.; Liu, Z.; Fu, Y.; Yang, Y.; Lu, J.; Pan, M.; Wen, T.; Xie, X.; Bai, Y.; Ge, Q. Single-Cell Multi-Omics Sequencing and Its Application in Tumor Heterogeneity. *Brief. Funct. Genom.* **2023**, *22*, 313–328. [[CrossRef](#)] [[PubMed](#)]
19. Dimitriu, M.A.; Lazar-Contes, I.; Roszkowski, M.; Mansuy, I.M. Single-Cell Multiomics Techniques: From Conception to Applications. *Front. Cell Dev. Biol.* **2022**, *10*, 854317. [[CrossRef](#)]
20. Flynn, E.; Almonte-Loya, A.; Fragiadakis, G.K. Single-cell multiomics. *Annu. Rev. Biomed. Data Sci.* **2023**, *6*, 313–337. [[CrossRef](#)]
21. Adossa, N.; Khan, S.; Rytönen, K.T.; Elo, L.L. Computational Strategies for Single-Cell Multi-Omics Integration. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 2588–2596. [[CrossRef](#)]
22. 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*, 695–713. [[CrossRef](#)] [[PubMed](#)]
23. Chen, G.; Yu, R.; Chen, X. Editorial: Integrative Analysis of Single-Cell and/or Bulk Multi-Omics Sequencing Data. *Front. Genet.* **2023**, *13*, 1121999. [[CrossRef](#)] [[PubMed](#)]
24. Huang, Y.; Mohanty, V.; Dede, M.; Tsai, K.; Daher, M.; Li, L.; Rezvani, K.; Chen, K. Characterizing Cancer Metabolism from Bulk and Single-Cell RNA-Seq Data Using METAFIux. *Nat. Commun.* **2023**, *14*, 4883. [[CrossRef](#)] [[PubMed](#)]
25. Cheng, C.; Chen, W.; Jin, H.; Chen, X. A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell-Cell Communication. *Cells* **2023**, *12*, 1970. [[CrossRef](#)] [[PubMed](#)]
26. Serin, E.A.R.; Nijveen, H.; Hilhorst, H.W.M.; Ligterink, W. Learning from Co-Expression Networks: Possibilities and Challenges. *Front. Plant Sci.* **2016**, *7*, 185898. [[CrossRef](#)]
27. Zeng, F.; Shi, M.; Xiao, H.; Chi, X. WGCNA-Based Identification of Hub Genes and Key Pathways Involved in Nonalcoholic Fatty Liver Disease. *BioMed Res. Int.* **2021**, *2021*, 5633211. [[CrossRef](#)]
28. Choi, H.; Na, K.J. Integrative Analysis of Imaging and Transcriptomic Data of the Immune Landscape Associated with Tumor Metabolism in Lung Adenocarcinoma: Clinical and Prognostic Implications. *Theranostics* **2018**, *8*, 1956–1965. [[CrossRef](#)]
29. Liu, P.; Luo, J.; Zheng, Q.; Chen, Q.; Zhai, N.; Xu, S.; Xu, Y.; Jin, L.; Xu, G.; Lu, X.; et al. Integrating Transcriptome and Metabolome Reveals Molecular Networks Involved in Genetic and Environmental Variation in Tobacco. *DNA Res.* **2020**, *27*, dsaa006. [[CrossRef](#)]
30. Amiri, F.; Moghadam, A.; Tahmasebi, A.; Niazi, A. Identification of Key Genes Involved in Secondary Metabolite Biosynthesis in *Digitalis Purpurea*. *PLoS ONE* **2023**, *18*, e0277293. [[CrossRef](#)]
31. Moghadam, A.; Foroozan, E.; Tahmasebi, A.; Taghizadeh, M.S.; Bolhassani, M.; Jafari, M. System Network Analysis of *Rosmarinus Officinalis* Transcriptome and Metabolome—Key Genes in Biosynthesis of Secondary Metabolites. *PLoS ONE* **2023**, *18*, e0282316. [[CrossRef](#)]
32. Ponsuksili, S.; Trakooljul, N.; Hadlich, F.; Methling, K.; Lalk, M.; Murani, E.; Wimmers, K. Genetic Regulation of Liver Metabolites and Transcripts Linking to Biochemical-Clinical Parameters. *Front. Genet.* **2019**, *10*, 419414. [[CrossRef](#)] [[PubMed](#)]
33. Zhu, Y.; Mordaunt, C.E.; Durbin-Johnson, B.P.; Caudill, M.A.; Malysheva, O.V.; Miller, J.W.; Green, R.; James, S.J.; Melnyk, S.B.; Fallin, M.D.; et al. Expression Changes in Epigenetic Gene Pathways Associated with One-Carbon Nutritional Metabolites in Maternal Blood from Pregnancies Resulting in Autism and Non-Typical Neurodevelopment. *Autism Res.* **2021**, *14*, 11–28. [[CrossRef](#)] [[PubMed](#)]
34. Hoang, L.T.; Domingo-Sabugo, C.; Starren, E.S.; Willis-Owen, S.A.G.; Morris-Rosendahl, D.J.; Nicholson, A.G.; Cookson, W.O.C.M.; Moffatt, M.F. Metabolomic, Transcriptomic and Genetic Integrative Analysis Reveals Important Roles of Adenosine Diphosphate in Haemostasis and Platelet Activation in Non-Small-Cell Lung Cancer. *Mol. Oncol.* **2019**, *13*, 2406–2421. [[CrossRef](#)] [[PubMed](#)]
35. Wang, Z.; Zhang, X.; He, S.; Rehman, A.; Jia, Y.; Li, H.; Pan, Z.; Geng, X.; Gao, Q.; Wang, L.; et al. Transcriptome Co-Expression Network and Metabolome Analysis Identifies Key Genes and Regulators of Proanthocyanidins Biosynthesis in Brown Cotton. *Front. Plant Sci.* **2022**, *12*, 822198. [[CrossRef](#)] [[PubMed](#)]
36. Green, R.E.; Lord, J.; Scelsi, M.A.; Xu, J.; Wong, A.; Naomi-James, S.; Handy, A.; Gilchrist, L.; Williams, D.M.; Parker, T.D.; et al. Investigating Associations between Blood Metabolites, Later Life Brain Imaging Measures, and Genetic Risk for Alzheimer’s Disease. *Alzheimer’s Res. Ther.* **2023**, *15*, 38. [[CrossRef](#)] [[PubMed](#)]
37. Green, R.; Lord, J.; Xu, J.; Maddock, J.; Kim, M.; Dobson, R.; Legido-Quigley, C.; Wong, A.; Richards, M.; Proitsi, P. Metabolic Correlates of Late Midlife Cognitive Outcomes: Findings from the 1946 British Birth Cohort. *Brain Commun.* **2022**, *4*, fcab291. [[CrossRef](#)] [[PubMed](#)]
38. Carson, C.; Lawson, H.A. Genetic Background and Diet Affect Brown Adipose Gene Coexpression Networks Associated with Metabolic Phenotypes. *Physiol. Genom.* **2020**, *52*, 223–233. [[CrossRef](#)] [[PubMed](#)]
39. Zhao, X.; Ge, W.; Miao, Z. Integrative Metabolomic and Transcriptomic Analyses Reveals the Accumulation Patterns of Key Metabolites Associated with Flavonoids and Terpenoids of *Gynostemma pentaphyllum* (Thunb.) Makino. *Sci. Rep.* **2024**, *14*, 8644. [[CrossRef](#)]
40. Xie, Z.; Wang, J.; Wang, W.; Wang, Y.; Xu, J.; Li, Z.; Zhao, X.; Fu, B. Integrated Analysis of the Transcriptome and Metabolome Revealed the Molecular Mechanisms Underlying the Enhanced Salt Tolerance of Rice Due to the Application of Exogenous Melatonin. *Front. Plant Sci.* **2021**, *11*, 618680. [[CrossRef](#)]
41. Zhou, Z.; Liu, J.; Meng, W.; Sun, Z.; Tan, Y.; Liu, Y.; Tan, M.; Wang, B.; Yang, J. Integrated Analysis of Transcriptome and Metabolome Reveals Molecular Mechanisms of Rice with Different Salinity Tolerances. *Plants* **2023**, *12*, 3359. [[CrossRef](#)] [[PubMed](#)]

42. Zhao, D.; Zhang, Y.; Ren, H.; Shi, Y.; Dong, D.; Li, Z.; Cui, G.; Shen, Y.; Mou, Z.; Kennelly, E.J.; et al. Multi-Omics Analysis Reveals the Evolutionary Origin of Diterpenoid Alkaloid Biosynthesis Pathways in *Aconitum*. *J. Integr. Plant Biol.* **2023**, *65*, 2320–2335. [[CrossRef](#)] [[PubMed](#)]
43. Nikiforova, V.J.; Daub, C.O.; Hesse, H.; Willmitzer, L.; Hoefgen, R. Integrative Gene-Metabolite Network with Implemented Causality Deciphers Informational Fluxes of Sulphur Stress Response. *J. Exp. Bot.* **2005**, *56*, 1887–1896. [[CrossRef](#)] [[PubMed](#)]
44. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498. [[CrossRef](#)]
45. Csárdi, G.; Nepusz, T.; Müller, K.; Horvát, S.; Traag, V.; Zanini, F.; Noom, D. Igraph for R: R Interface of the Igraph Library for Graph Theory and Network Analysis, Version 1.4. *Zenodo* **2023**. [[CrossRef](#)]
46. Rai, A.; Rai, M.; Kamochi, H.; Mori, T.; Nakabayashi, R.; Nakamura, M.; Suzuki, H.; Saito, K.; Yamazaki, M. Multiomics-Based Characterization of Specialized Metabolites Biosynthesis in *Cornus Officinalis*. *DNA Res.* **2020**, *27*, dsaa009. [[CrossRef](#)] [[PubMed](#)]
47. Saito, K.; Hirai, M.Y.; Yonekura-Sakakibara, K. Decoding Genes with Coexpression Networks and Metabolomics—“Majority Report by Precogs”. *Trends Plant Sci.* **2008**, *13*, 36–43. [[CrossRef](#)] [[PubMed](#)]
48. Cline, M.S.; Smoot, M.; Cerami, E.; Kuchinsky, A.; Landys, N.; Workman, C.; Christmas, R.; Avila-Campilo, I.; Creech, M.; Gross, B.; et al. Integration of Biological Networks and Gene Expression Data Using Cytoscape. *Nat. Protoc.* **2007**, *2*, 2366–2382. [[CrossRef](#)] [[PubMed](#)]
49. Liu, X.L.; Ming, Y.N.; Zhang, J.Y.; Chen, X.Y.; Zeng, M.D.; Mao, Y.M. Gene-Metabolite Network Analysis in Different Nonalcoholic Fatty Liver Disease Phenotypes. *Exp. Mol. Med.* **2017**, *49*, e283. [[CrossRef](#)]
50. Mounet, F.; Moing, A.; Garcia, V.; Petit, J.; Maucourt, M.; Deborde, C.; Bernillon, S.; Le Gall, G.; Colquhoun, I.; Defernez, M.; et al. Gene and Metabolite Regulatory Network Analysis of Early Developing Fruit Tissues Highlights New Candidate Genes for the Control of Tomato Fruit Composition and Development. *Plant Physiol.* **2009**, *149*, 1505–1528. [[CrossRef](#)]
51. Mercatelli, D.; Lopez-Garcia, G.; Giorgi, F.M. Corto: A Lightweight R Package for Gene Network Inference and Master Regulator Analysis. *Bioinformatics* **2020**, *36*, 3916–3917. [[CrossRef](#)] [[PubMed](#)]
52. Cavicchioli, M.V.; Santorsola, M.; Balboni, N.; Mercatelli, D.; Giorgi, F.M. Prediction of Metabolic Profiles from Transcriptomics Data in Human Cancer Cell Lines. *Int. J. Mol. Sci.* **2022**, *23*, 3867. [[CrossRef](#)] [[PubMed](#)]
53. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat. Methods* **2014**, *11*, 333–337. [[CrossRef](#)] [[PubMed](#)]
54. Miranda, J.; Paules, C.; Noell, G.; Youssef, L.; Paternina-Cacedo, A.; Crovetto, F.; Cañellas, N.; Garcia-Martín, M.L.; Amigó, N.; Eixarch, E.; et al. Similarity Network Fusion to Identify Phenotypes of Small-for-Gestational-Age Fetuses. *iScience* **2023**, *26*, 107620. [[CrossRef](#)] [[PubMed](#)]
55. Raphael, B.J.; Hruban, R.H.; Aguirre, A.J.; Moffitt, R.A.; Yeh, J.J.; Stewart, C.; Robertson, A.G.; Cherniack, A.D.; Gupta, M.; Getz, G.; et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **2017**, *32*, 185–203.e13. [[CrossRef](#)]
56. Blimkie, T.; Lee, A.H.Y.; Hancock, R.E.W. MetaBridge: An Integrative Multi-Omics Tool for Metabolite-Enzyme Mapping. *Curr. Protoc. Bioinform.* **2020**, *70*, e98. [[CrossRef](#)]
57. Caspi, R.; Billington, R.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Midford, P.E.; Ong, W.K.; Paley, S.; Subhraveti, P.; Karp, P.D. The MetaCyc Database of Metabolic Pathways and Enzymes—A 2019 Update. *Nucleic Acids Res.* **2020**, *48*, D445–D453. [[CrossRef](#)] [[PubMed](#)]
58. Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; et al. ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation* **2021**, *2*, 100141. [[CrossRef](#)]
59. Fang, Z.; Liu, X.; Peltz, G. GSEAPy: A Comprehensive Package for Performing Gene Set Enrichment Analysis in Python. *Bioinformatics* **2023**, *39*, btac757. [[CrossRef](#)] [[PubMed](#)]
60. Pang, Z.; Chong, J.; Zhou, G.; De Lima Morais, D.A.; Chang, L.; Barrette, M.; Gauthier, C.; Jacques, P.É.; Li, S.; Xia, J. MetaboAnalyst 5.0: Narrowing the Gap between Raw Spectra and Functional Insights. *Nucleic Acids Res.* **2021**, *49*, W388–W396. [[CrossRef](#)]
61. Luo, W.; Brouwer, C. Pathview: An R/Bioconductor Package for Pathway-Based Data Integration and Visualization. *Bioinformatics* **2013**, *29*, 1830–1831. [[CrossRef](#)] [[PubMed](#)]
62. Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J.C.; Buettner, F.; Huber, W.; Stegle, O. Multi-Omics Factor Analysis—A Framework for Unsupervised Integration of Multi-Omics Data Sets. *Mol. Syst. Biol.* **2018**, *14*, 8124. [[CrossRef](#)] [[PubMed](#)]
63. Dutta, N.K.; Tornheim, J.A.; Fukutani, K.F.; Paradkar, M.; Tiburcio, R.T.; Kinikar, A.; Valvi, C.; Kulkarni, V.; Pradhan, N.; Shivakumar, S.V.B.Y.; et al. Integration of Metabolomics and Transcriptomics Reveals Novel Biomarkers in the Blood for Tuberculosis Diagnosis in Children. *Sci. Rep.* **2020**, *10*, 19527. [[CrossRef](#)]
64. Clark, C.; Dayon, L.; Masoodi, M.; Bowman, G.L.; Popp, J. An Integrative Multi-Omics Approach Reveals New Central Nervous System Pathway Alterations in Alzheimer’s Disease. *Alzheimer’s Res. Ther.* **2021**, *13*, 71. [[CrossRef](#)]
65. Wang, H.; Liu, C.; Xie, X.; Niu, M.; Wang, Y.; Cheng, X.; Zhang, B.; Zhang, D.; Liu, M.; Sun, R.; et al. Multi-Omics Blood Atlas Reveals Unique Features of Immune and Platelet Responses to SARS-CoV-2 Omicron Breakthrough Infection. *Immunity* **2023**, *56*, 1410–1428.e8. [[CrossRef](#)] [[PubMed](#)]
66. Cavill, R.; Jennen, D.; Kleinjans, J.; Briedé, J.J. Transcriptomic and Metabolomic Data Integration. *Brief. Bioinform.* **2016**, *17*, 891–901. [[CrossRef](#)]

67. Gu, C.; Kim, G.B.; Kim, W.J.; Kim, H.U.; Lee, S.Y. Current Status and Applications of Genome-Scale Metabolic Models. *Genome Biol.* **2019**, *20*, 121. [[CrossRef](#)]
68. Larsson, I.; Uhlén, M.; Zhang, C.; Mardinoglu, A. Genome-Scale Metabolic Modeling of Glioblastoma Reveals Promising Targets for Drug Development. *Front. Genet.* **2020**, *11*, 459644. [[CrossRef](#)] [[PubMed](#)]
69. Karakitsou, E.; Foguet, C.; Contreras Mostazo, M.G.; Kurrle, N.; Schnütgen, F.; Michaelis, M.; Cinatl, J.; Marin, S.; Cascante, M. Genome-Scale Integration of Transcriptome and Metabolome Unveils Squalene Synthase and Dihydrofolate Reductase as Targets against AML Cells Resistant to Chemotherapy. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4059–4066. [[CrossRef](#)]
70. Sen, P.; Orešič, M. Integrating Omics Data in Genome-Scale Metabolic Modeling: A Methodological Perspective for Precision Medicine. *Metabolites* **2023**, *13*, 855. [[CrossRef](#)]
71. Kim, M.K.; Lun, D.S. Methods for Integration of Transcriptomic Data in Genome-Scale Metabolic Models. *Comput. Struct. Biotechnol. J.* **2014**, *11*, 59–65. [[CrossRef](#)] [[PubMed](#)]
72. Roshanzamir, F.; Robinson, J.L.; Cook, D.; Karimi-Jafari, M.H.; Nielsen, J. Metastatic Triple Negative Breast Cancer Adapts Its Metabolism to Destination Tissues While Retaining Key Metabolic Signatures. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2205456119. [[CrossRef](#)] [[PubMed](#)]
73. Orth, J.D.; Thiele, I.; Palsson, B.O. What Is Flux Balance Analysis? *Nat. Biotechnol.* **2010**, *28*, 245. [[CrossRef](#)]
74. Lewis, N.E.; Hixson, K.K.; Conrad, T.M.; Lerman, J.A.; Charusanti, P.; Polpitiya, A.D.; Adkins, J.N.; Schramm, G.; Purvine, S.O.; Lopez-Ferrer, D.; et al. Omic Data from Evolved *E. coli* Are Consistent with Computed Optimal Growth from Genome-Scale Models. *Mol. Syst. Biol.* **2010**, *6*, 390. [[CrossRef](#)] [[PubMed](#)]
75. Pereira, V.; Cruz, F.; Rocha, M. MEWpy: A Computational Strain Optimization Workbench in Python. *Bioinformatics* **2021**, *37*, 2494–2496. [[CrossRef](#)] [[PubMed](#)]
76. Lu, H.; Li, F.; Sánchez, B.J.; Zhu, Z.; Li, G.; Domenzain, I.; Marcišauskas, S.; Anton, P.M.; Lappa, D.; Lieven, C.; et al. A Consensus *S. cerevisiae* Metabolic Model Yeast8 and Its Ecosystem for Comprehensively Probing Cellular Metabolism. *Nat. Commun.* **2019**, *10*, 3586. [[CrossRef](#)]
77. Robinson, J.L.; Kocabaş, P.; Wang, H.; Cholley, P.E.; Cook, D.; Nilsson, A.; Anton, M.; Ferreira, R.; Domenzain, I.; Billa, V.; et al. An Atlas of Human Metabolism. *Sci. Signal.* **2020**, *13*, eaaz1482. [[CrossRef](#)] [[PubMed](#)]
78. Monk, J.M.; Lloyd, C.J.; Brunk, E.; Mih, N.; Sastry, A.; King, Z.; Takeuchi, R.; Nomura, W.; Zhang, Z.; Mori, H.; et al. IML1515, a Knowledgebase That Computes Escherichia Coli Traits. *Nat. Biotechnol.* **2017**, *35*, 904–908. [[CrossRef](#)] [[PubMed](#)]
79. Domenzain, I.; Sánchez, B.; Anton, M.; Kerkhoven, E.J.; Millán-Oropeza, A.; Henry, C.; Siewers, V.; Morrissey, J.P.; Sonnenschein, N.; Nielsen, J. Reconstruction of a Catalogue of Genome-Scale Metabolic Models with Enzymatic Constraints Using GECKO 2.0. *Nat. Commun.* **2022**, *13*, 3766. [[CrossRef](#)]
80. Sánchez, B.J.; Zhang, C.; Nilsson, A.; Lahtvee, P.; Kerkhoven, E.J.; Nielsen, J. Improving the Phenotype Predictions of a Yeast Genome-scale Metabolic Model by Incorporating Enzymatic Constraints. *Mol. Syst. Biol.* **2017**, *13*, 935. [[CrossRef](#)]
81. Zhou, J.; Zhuang, Y.; Xia, J. Integration of Enzyme Constraints in a Genome-Scale Metabolic Model of *Aspergillus Niger* Improves Phenotype Predictions. *Microb. Cell Factories* **2021**, *20*, 125. [[CrossRef](#)] [[PubMed](#)]
82. Arend, M.; Zimmer, D.; Xu, R.; Sommer, F.; Mühlhaus, T.; Nikoloski, Z. Proteomics and Constraint-Based Modelling Reveal Enzyme Kinetic Properties of *Chlamydomonas Reinhardtii* on a Genome Scale. *Nat. Commun.* **2023**, *14*, 4781. [[CrossRef](#)] [[PubMed](#)]
83. Wu, K.; Mao, Z.; Mao, Y.; Niu, J.; Cai, J.; Yuan, Q.; Yun, L.; Liao, X.; Wang, Z.; Ma, H. EcBSU1: A Genome-Scale Enzyme-Constrained Model of *Bacillus Subtilis* Based on the ECPy Workflow. *Microorganisms* **2023**, *11*, 178. [[CrossRef](#)] [[PubMed](#)]
84. Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D. BRENDA in 2017: New Perspectives and New Tools in BRENDA. *Nucleic Acids Res.* **2017**, *45*, D380–D388. [[CrossRef](#)] [[PubMed](#)]
85. Bateman, A.; Martin, M.J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523–D531. [[CrossRef](#)]
86. Chen, J.; Luo, P.; Wang, C.; Yang, C.; Bai, Y.; He, X.; Zhang, Q.; Zhang, J.; Yang, J.; Wang, S.; et al. Integrated Single-Cell Transcriptomics and Proteomics Reveal Cellular-Specific Responses and Microenvironment Remodeling in Aristolochic Acid Nephropathy. *JCI Insight* **2022**, *7*, e157360. [[CrossRef](#)] [[PubMed](#)]
87. Alsagaby, S.A. Integration of Proteomics and Transcriptomics Data Sets Identifies Prognostic Markers in Chronic Lymphocytic Leukemia. *Majmaah J. Health Sci.* **2019**, *7*, 1. [[CrossRef](#)]
88. Higdon, R.; Kala, J.; Wilkins, D.; Yan, J.F.; Sethi, M.K.; Lin, L.; Liu, S.; Montague, E.; Janko, I.; Choiniere, J.; et al. Integrated Proteomic and Transcriptomic-Based Approaches to Identifying Signature Biomarkers and Pathways for Elucidation of Daoy and UW228 Subtypes. *Proteomes* **2017**, *5*, 5. [[CrossRef](#)]
89. Gygi, S.P.; Rochon, Y.; Franza, B.R.; Aebersold, R. Correlation between Protein and mRNA Abundance in Yeast. *Mol. Cell. Biol.* **1999**, *19*, 1720–1730. [[CrossRef](#)]
90. Yue, X.; Huan, P.; Hu, Y.; Liu, B. Integrated Transcriptomic and Proteomic Analyses Reveal Potential Mechanisms Linking Thermal Stress and Depressed Disease Resistance in the Turbot *Scophthalmus Maximus*. *Sci. Rep.* **2018**, *8*, 1896. [[CrossRef](#)]
91. Li, G.; Zhang, B.; Zhang, H.; Xu, A.; Qian, H. Integration of Transcriptomic and Proteomic Analyses Reveals New Insights into the Regulation of Immune Pathways in Midgut of *Samia Ricini* upon SariNPV Infection. *Insects* **2022**, *13*, 294. [[CrossRef](#)] [[PubMed](#)]
92. Sun, Z.; Liu, Y.; He, X.; Di, R.; Wang, X.; Ren, C.; Zhang, Z.; Chu, M. Integrative Proteomics and Transcriptomics Profiles of the Oviduct Reveal the Prolificacy-Related Candidate Biomarkers of Goats (*Capra hircus*) in Estrous Periods. *Int. J. Mol. Sci.* **2022**, *23*, 14888. [[CrossRef](#)] [[PubMed](#)]

93. Zhang, Q.; Li, Y.; Sun, L.; Chu, S.; Xu, H.; Zhou, X. Integration of Transcriptomic and Proteomic Analyses of *Rhododendron chrysanthum* Pall. in Response to Cold Stress in the Changbai Mountains. *Mol. Biol. Rep.* **2023**, *50*, 3607–3616. [[CrossRef](#)] [[PubMed](#)]
94. Miao, J.; Yang, Z.; Guo, W.; Liu, L.; Song, P.; Ding, C.; Guan, W. Integrative Analysis of the Proteome and Transcriptome in Gastric Cancer Identified LRP1B as a Potential Biomarker. *Biomark. Med.* **2022**, *16*, 1101–1111. [[CrossRef](#)] [[PubMed](#)]
95. Colak, D.; Alaiya, A.A.; Kaya, N.; Muiya, N.P.; AlHarazi, O.; Shinwari, Z.; Andres, E.; Dzimiri, N. Integrated Left Ventricular Global Transcriptome and Proteome Profiling in Human End-Stage Dilated Cardiomyopathy. *PLoS ONE* **2016**, *11*, e0162669. [[CrossRef](#)] [[PubMed](#)]
96. Du, Y.; Clair, G.C.; Al Alam, D.; Danopoulos, S.; Schnell, D.; Kitzmiller, J.A.; Misra, R.S.; Bhattacharya, S.; Warburton, D.; Mariani, T.J.; et al. Integration of Transcriptomic and Proteomic Data Identifies Biological Functions in Cell Populations from Human Infant Lung. *Am. J. Physiol. Cell. Mol. Physiol.* **2019**, *317*, L347. [[CrossRef](#)] [[PubMed](#)]
97. Peng, Z.; He, S.; Gong, W.; Xu, F.; Pan, Z.; Jia, Y.; Geng, X.; Du, X. Integration of Proteomic and Transcriptomic Profiles Reveals Multiple Levels of Genetic Regulation of Salt Tolerance in Cotton. *BMC Plant Biol.* **2018**, *18*, 128. [[CrossRef](#)] [[PubMed](#)]
98. Zheng, W.; Zhang, Y.; Sun, C.; Ge, S.; Tan, Y.; Shen, H.; Yang, P. A Multi-Omics Study of Human Testis and Epididymis. *Molecules* **2021**, *26*, 3345. [[CrossRef](#)]
99. Liu, Y.; Beyer, A.; Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **2016**, *165*, 535–550. [[CrossRef](#)]
100. Zhang, G.; Zhong, F.; Chen, L.; Qin, P.; Li, J.; Zhi, F.; Tian, L.; Zhou, D.; Lin, P.; Chen, H.; et al. Integrated Proteomic and Transcriptomic Analyses Reveal the Roles of Brucella Homolog of BAX Inhibitor 1 in Cell Division and Membrane Homeostasis of Brucella Suis S2. *Front. Microbiol.* **2021**, *12*, 632095. [[CrossRef](#)]
101. Griss, J.; Viteri, G.; Sidiropoulos, K.; Nguyen, V.; Fabregat, A.; Hermjakob, H. ReactomeGSA-Efficient Multi-Omics Comparative Pathway Analysis. *Mol. Cell. Proteom.* **2020**, *19*, 2115–2124. [[CrossRef](#)] [[PubMed](#)]
102. Stark, C.; Breitkreutz, B.J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M. BioGRID: A General Repository for Interaction Datasets. *Nucleic Acids Res.* **2006**, *34*, D535. [[CrossRef](#)]
103. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING V11: Protein–Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [[CrossRef](#)]
104. Wang, X.W.; Wang, T.; Schaub, D.P.; Chen, C.; Sun, Z.; Ke, S.; Hecker, J.; Maaser-Hecker, A.; Zeleznik, O.A.; Zeleznik, R.; et al. Benchmarking Omics-Based Prediction of Asthma Development in Children. *Respir. Res.* **2023**, *24*, 62. [[CrossRef](#)] [[PubMed](#)]
105. Arjmand, B.; Hamidpour, S.K.; Tayanloo-Beik, A.; Goodarzi, P.; Aghayan, H.R.; Adibi, H.; Larijani, B. Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer. *Front. Genet.* **2022**, *13*, 824451. [[CrossRef](#)]
106. Wekesa, J.S.; Kimwele, M. A Review of Multi-Omics Data Integration through Deep Learning Approaches for Disease Diagnosis, Prognosis, and Treatment. *Front. Genet.* **2023**, *14*, 1199087. [[CrossRef](#)]
107. Li, R.; Li, L.; Xu, Y.; Yang, J. Machine Learning Meets Omics: Applications and Perspectives. *Brief. Bioinform.* **2022**, *23*, bbab460. [[CrossRef](#)] [[PubMed](#)]
108. Leite, D.M.C.; Brochet, X.; Resch, G.; Que, Y.A.; Neves, A.; Peña-Reyes, C. Computational Prediction of Inter-Species Relationships through Omics Data Analysis and Machine Learning. *BMC Bioinform.* **2018**, *19*, 151–159. [[CrossRef](#)]
109. Alharbi, F.; Vakanski, A. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. *Bioengineering* **2023**, *10*, 173. [[CrossRef](#)]
110. Bashiri, A.; Ghazisaeedi, M.; Safdari, R.; Shahmoradi, L.; Ehtesham, H. Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. *Iran. J. Public Health* **2017**, *46*, 165.
111. Hijazi, H.; Chan, C. A Classification Framework Applied to Cancer Gene Expression Profiles. *J. Healthc. Eng.* **2013**, *4*, 255–283. [[CrossRef](#)]
112. Khalsan, M.; MacHado, L.R.; Al-Shamery, E.S.; Ajit, S.; Anthony, K.; Mu, M.; Agyeman, M.O. A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction. *IEEE Access* **2022**, *10*, 27522–27534. [[CrossRef](#)]
113. Ravindran, U.; Gunavathi, C. A Survey on Gene Expression Data Analysis Using Deep Learning Methods for Cancer Diagnosis. *Prog. Biophys. Mol. Biol.* **2023**, *177*, 1–13. [[CrossRef](#)]
114. Tabares-Soto, R.; Orozco-Arias, S.; Romero-Cano, V.; Bucheli, V.S.; Rodríguez-Sotelo, J.L.; Jiménez-Varón, C.F. A Comparative Study of Machine Learning and Deep Learning Algorithms to Classify Cancer Types Based on Microarray Gene Expression Data. *PeerJ Comput. Sci.* **2020**, *2020*, e270. [[CrossRef](#)] [[PubMed](#)]
115. Yuan, F.; Lu, L.; Zou, Q. Analysis of Gene Expression Profiles of Lung Cancer Subtypes with Machine Learning Algorithms. *Biochim. Biophys. Acta (BBA)-Mol. Basis Dis.* **2020**, *1866*, 165822. [[CrossRef](#)]
116. He, B.; Thomson, M.; Subramaniam, M.; Perez, R.; Ye, C.J.; Zou, J. CloudPred: Predicting Patient Phenotypes from Single-Cell RNA-Seq. *Pac. Symp. Biocomput.* **2022**, *2021*, 337–348. [[CrossRef](#)]
117. Ma, Y.; Chen, J.; Wang, T.; Zhang, L.; Xu, X.; Qiu, Y.; Xiang, A.P.; Huang, W. Accurate Machine Learning Model to Diagnose Chronic Autoimmune Diseases Utilizing Information From B Cells and Monocytes. *Front. Immunol.* **2022**, *13*, 870531. [[CrossRef](#)] [[PubMed](#)]
118. Galdos, F.X.; Xu, S.; Goodyer, W.R.; Duan, L.; Huang, Y.V.; Lee, S.; Zhu, H.; Lee, C.; Wei, N.; Lee, D.; et al. DevCellPy Is a Machine Learning-Enabled Pipeline for Automated Annotation of Complex Multilayered Single-Cell Transcriptomic Data. *Nat. Commun.* **2022**, *13*, 5271. [[CrossRef](#)]

119. Liu, J.; Fan, Z.; Zhao, W.; Zhou, X. Machine Intelligence in Single-Cell Data Analysis: Advances and New Challenges. *Front. Genet.* **2021**, *12*, 655536. [[CrossRef](#)]
120. Patil, A.R.; Schug, J.; Liu, C.; Lahori, D.; Descamps, H.C.; Naji, A.; Kaestner, K.H.; Faryabi, R.B.; Vahedi, G. Modeling Type 1 Diabetes Progression Using Machine Learning and Single-Cell Transcriptomic Measurements in Human Islets. *Cell Rep. Med.* **2024**, *5*, 101535. [[CrossRef](#)]
121. Hu, Y.; Hase, T.; Li, H.P.; Prabhakar, S.; Kitano, H.; Ng, S.K.; Ghosh, S.; Wee, L.J.K. A Machine Learning Approach for the Identification of Key Markers Involved in Brain Development from Single-Cell Transcriptomic Data. *BMC Genom.* **2016**, *17*, 1025. [[CrossRef](#)] [[PubMed](#)]
122. Vrahatis, A.G.; Tasoulis, S.K.; Maglogiannis, I.; Plagianakos, V.P. Recent Machine Learning Approaches for Single-Cell RNA-Seq Data Analysis. *Stud. Comput. Intell.* **2020**, *891*, 65–79. [[CrossRef](#)]
123. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for Functional Genomics Data Sets—Update. *Nucleic Acids Res.* **2013**, *41*, D991–D995. [[CrossRef](#)] [[PubMed](#)]
124. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Sander, C.; Stuart, J.M.; Chang, K.; Creighton, C.J.; et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **2013**, *45*, 1113–1120. [[CrossRef](#)] [[PubMed](#)]
125. Johnson, N.T.; Dhroso, A.; Hughes, K.J.; Korkin, D. Biological Classification with RNA-Seq Data: Can Alternatively Spliced Transcript Expression Enhance Machine Learning Classifiers? *RNA* **2018**, *24*, 1119–1132. [[CrossRef](#)] [[PubMed](#)]
126. Kuo, W.P.; Kim, E.Y.; Trimarchi, J.; Jenssen, T.K.; Vinterbo, S.A.; Ohno-Machado, L. A Primer on Gene Expression and Microarrays for Machine Learning Researchers. *J. Biomed. Inform.* **2004**, *37*, 293–303. [[CrossRef](#)] [[PubMed](#)]
127. Kelchtermans, P.; Bittremieux, W.; De Grave, K.; Degroeve, S.; Ramon, J.; Laukens, K.; Valkenburg, D.; Barsnes, H.; Martens, L. Machine Learning Applications in Proteomics Research: How the Past Can Boost the Future. *Proteomics* **2014**, *14*, 353–366. [[CrossRef](#)] [[PubMed](#)]
128. Swan, A.L.; Mobasher, A.; Allaway, D.; Liddell, S.; Bacardit, J. Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. *OMICS* **2013**, *17*, 595–610. [[CrossRef](#)] [[PubMed](#)]
129. Barla, A.; Jurman, G.; Riccadonna, S.; Merler, S.; Chierici, M.; Furlanello, C. Machine Learning Methods for Predictive Proteomics. *Brief. Bioinform.* **2008**, *9*, 119–128. [[CrossRef](#)]
130. Neely, B.A.; Dorfer, V.; Martens, L.; Bludau, I.; Bouwmeester, R.; Degroeve, S.; Deutsch, E.W.; Gessulat, S.; Käll, L.; Palczynski, P.; et al. Toward an Integrated Machine Learning Model of a Proteomics Experiment. *J. Proteome Res.* **2023**, *22*, 681–696. [[CrossRef](#)]
131. Desaire, H.; Go, E.P.; Hua, D. Advances, Obstacles, and Opportunities for Machine Learning in Proteomics. *Cell Rep. Phys. Sci.* **2022**, *3*, 101069. [[CrossRef](#)] [[PubMed](#)]
132. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [[CrossRef](#)] [[PubMed](#)]
133. Sonsare, P.M.; Gunavathi, C. Investigation of Machine Learning Techniques on Proteomics: A Comprehensive Survey. *Prog. Biophys. Mol. Biol.* **2019**, *149*, 54–69. [[CrossRef](#)] [[PubMed](#)]
134. Vishnoi, S.; Matre, H.; Garg, P.; Pandey, S.K. Artificial Intelligence and Machine Learning for Protein Toxicity Prediction Using Proteomics Data. *Chem. Biol. Drug Des.* **2020**, *96*, 902–920. [[CrossRef](#)]
135. Mann, M.; Kumar, C.; Zeng, W.F.; Strauss, M.T. Artificial Intelligence for Proteomics and Biomarker Discovery. *Cell Syst.* **2021**, *12*, 759–770. [[CrossRef](#)] [[PubMed](#)]
136. Li, J.; Zhou, K.; Mu, B. Machine Learning for Mass Spectrometry Data Analysis in Proteomics. *Curr. Proteom.* **2020**, *18*, 620–634. [[CrossRef](#)]
137. Wen, B.; Zeng, W.F.; Liao, Y.; Shi, Z.; Savage, S.R.; Jiang, W.; Zhang, B. Deep Learning in Proteomics. *Proteomics* **2020**, *20*, 1900335. [[CrossRef](#)] [[PubMed](#)]
138. Hindson, J. Proteomics and Machine-Learning Models for Alcohol-Related Liver Disease Biomarkers. *Nat. Rev. Gastroenterol. Hepatol.* **2022**, *19*, 488. [[CrossRef](#)] [[PubMed](#)]
139. Cox, J. Prediction of Peptide Mass Spectral Libraries with Machine Learning. *Nat. Biotechnol.* **2022**, *41*, 33–43. [[CrossRef](#)]
140. Sengupta, A.; Naresh, G.; Mishra, A.; Parashar, D.; Narad, P. Proteome Analysis Using Machine Learning Approaches and Its Applications to Diseases. *Adv. Protein Chem. Struct. Biol.* **2021**, *127*, 161–216. [[CrossRef](#)]
141. Tilocca, B.; Britti, D.; Urbani, A.; Roncada, P. Computational Immune Proteomics Approach to Target COVID-19. *J. Proteome Res.* **2020**, *19*, 4233–4241. [[CrossRef](#)]
142. Bernardes, J.S.; Pedreira, C.E. A Review of Protein Function Prediction Under Machine Learning Perspective. *Recent Pat. Biotechnol.* **2013**, *7*, 122–141. [[CrossRef](#)] [[PubMed](#)]
143. Worley, B.; Powers, R. Multivariate Analysis in Metabolomics. *Curr. Metabolomics* **2013**, *1*, 92. [[CrossRef](#)]
144. Sinha, N.; Viswan, A.; Singh, C.; Rai, R.K.; Azim, A.; Baronia, A.K. Metabolomics Based Predictive Biomarker Model of ARDS: A Systemic Measure of Clinical Hypoxemia. *PLoS ONE* **2017**, *12*, e0187545. [[CrossRef](#)]
145. Martín-Blázquez, A.; Díaz, C.; González-Flores, E.; Franco-Rivas, D.; Jiménez-Luna, C.; Melguizo, C.; Prados, J.; Genilloud, O.; Vicente, F.; Caba, O.; et al. Untargeted LC-HRMS-Based Metabolomics to Identify Novel Biomarkers of Metastatic Colorectal Cancer. *Sci. Rep.* **2019**, *9*, 20198. [[CrossRef](#)]

146. Goldberg, E.; Ievari-Shariati, S.; Kidane, B.; Kim, J.; Banerji, S.; Qing, G.; Srinathan, S.; Murphy, L.; Aliani, M. Comparative Metabolomics Studies of Blood Collected in Streck and Heparin Tubes from Lung Cancer Patients. *PLoS ONE* **2021**, *16*, e0249648. [[CrossRef](#)] [[PubMed](#)]
147. French, C.D.; Willoughby, R.E.; Pan, A.; Wong, S.J.; Foley, J.F.; Wheat, L.J.; Fernandez, J.; Encarnacion, R.; Ondrush, J.M.; Fattah, N.; et al. NMR Metabolomics of Cerebrospinal Fluid Differentiates Inflammatory Diseases of the Central Nervous System. *PLoS Negl. Trop. Dis.* **2018**, *12*, e0007045. [[CrossRef](#)] [[PubMed](#)]
148. Collakova, E.; Aghamirzaie, D.; Fang, Y.; Klumas, C.; Tabataba, F.; Kakumanu, A.; Myers, E.; Heath, L.S.; Grene, R. Metabolic and Transcriptional Reprogramming in Developing Soybean (*Glycine Max*) Embryos. *Metabolites* **2013**, *3*, 347–372. [[CrossRef](#)]
149. You, J.; Zhang, Y.; Liu, A.; Li, D.; Wang, X.; Dossa, K.; Zhou, R.; Yu, J.; Zhang, Y.; Wang, L.; et al. Transcriptomic and Metabolomic Profiling of Drought-Tolerant and Susceptible Sesame Genotypes in Response to Drought Stress. *BMC Plant Biol.* **2019**, *19*, 267. [[CrossRef](#)]
150. Alreshidi, M.M. Selected Metabolites Profiling of *Staphylococcus Aureus* Following Exposure to Low Temperature and Elevated Sodium Chloride. *Front. Microbiol.* **2020**, *11*, 512774. [[CrossRef](#)]
151. Broughton-Neiswanger, L.E.; Rivera-Velez, S.M.; Suarez, M.A.; Slovak, J.E.; Piñeyro, P.E.; Hwang, J.K.; Villarino, N.F. Urinary Chemical Fingerprint Left behind by Repeated NSAID Administration: Discovery of Putative Biomarkers Using Artificial Intelligence. *PLoS ONE* **2020**, *15*, e0228989. [[CrossRef](#)] [[PubMed](#)]
152. Xu, J.; Pan, T.; Qi, X.; Tan, R.; Wang, X.; Liu, Z.; Tao, Z.; Qu, H.; Zhang, Y.; Chen, H.; et al. Increased Mortality of Acute Respiratory Distress Syndrome Was Associated with High Levels of Plasma Phenylalanine. *Respir. Res.* **2020**, *21*, 99. [[CrossRef](#)] [[PubMed](#)]
153. Monteleone, A.M.; Troisi, J.; Serena, G.; Fasano, A.; Grave, R.D.; Cascino, G.; Marciello, F.; Calugi, S.; Scala, G.; Corrivetti, G.; et al. The Gut Microbiome and Metabolomics Profiles of Restricting and Binge-Purging Type Anorexia Nervosa. *Nutrients* **2021**, *13*, 507. [[CrossRef](#)] [[PubMed](#)]
154. Caterino, M.; Costanzo, M.; Fedele, R.; Cevenini, A.; Gelzo, M.; Di Minno, A.; Andolfo, I.; Capasso, M.; Russo, R.; Annunziata, A.; et al. The Serum Metabolome of Moderate and Severe COVID-19 Patients Reflects Possible Liver Alterations Involving Carbon and Nitrogen Metabolism. *Int. J. Mol. Sci.* **2021**, *22*, 9548. [[CrossRef](#)] [[PubMed](#)]
155. Silva, A.A.R.; Cardoso, M.R.; Cardoso De Oliveira, D.; Godoy, P.; Cecília, M.; Talarico, R.; Marrero Gutiérrez, J.; Rodrigues Peres, R.M.; De Carvalho, L.M.; Angelo Da, N.; et al. Plasma Metabolome Signatures to Predict Responsiveness to Neoadjuvant Chemotherapy in Breast Cancer. *Cancers* **2024**, *16*, 2473. [[CrossRef](#)] [[PubMed](#)]
156. Galal, A.; Talal, M.; Moustafa, A. Applications of Machine Learning in Metabolomics: Disease Modeling and Classification. *Front. Genet.* **2022**, *13*, 1017340. [[CrossRef](#)] [[PubMed](#)]
157. Auslander, N.; Yizhak, K.; Weinstock, A.; Budhu, A.; Tang, W.; Wang, X.W.; Ambs, S.; Ruppin, E. A Joint Analysis of Transcriptomic and Metabolomic Data Uncovers Enhanced Enzyme-Metabolite Coupling in Breast Cancer. *Sci. Rep.* **2016**, *6*, 29662. [[CrossRef](#)]
158. Ghosh, T.; Zhang, W.; Ghosh, D.; Kechris, K. Predictive Modeling for Metabolomics Data. *Methods Mol. Biol.* **2020**, *2104*, 313. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.