metabolites

Review

Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources

Tara Eicher [1,2], Garrett Kinnebrew [1,3,4], Andrew Patt [5,6], Kyle Spencer [1,6,7], Kevin Ying [3,8], Qin Ma [1], Raghu Machiraju [1,2,9,10] and Ewy A. Mathé [1,5,*]

[1] Biomedical Informatics Department, The Ohio State University College of Medicine, Columbus, OH 43210, USA; eicher.33@osu.edu (T.E.); garrett.kinnebrew@osumc.edu (G.K.); spencer.698@osu.edu (K.S.); Qin.Ma@osumc.edu (Q.M.); machiraju.1@osu.edu (R.M.)

[2] Computer Science and Engineering Department, The Ohio State University College of Engineering, Columbus, OH 43210, USA

[3] Comprehensive Cancer Center, The Ohio State University and James Cancer Hospital, Columbus, OH 43210, USA; Kevin.Ying@osumc.edu

[4] Bioinformatics Shared Resource Group, The Ohio State University, Columbus, OH 43210, USA

[5] Division of Preclinical Innovation, National Center for Advancing Translational Sciences, NIH, 9800 Medical Center Dr., Rockville, MD, 20892, USA; patt.14@buckeyemail.osu.edu

[6] Biomedical Sciences Graduate Program, The Ohio State University, Columbus, OH 43210, USA

[7] Nationwide Children's Research Hospital, Columbus, OH 43210, USA

[8] Molecular, Cellular and Developmental Biology Program, The Ohio State University, Columbus, OH 43210, USA

[9] Department of Pathology, Wexner Medical Center, The Ohio State University, Columbus, OH 43210, USA

[10] Translational Data Analytics Institute, The Ohio State University, Columbus, OH 43210, USA

[*] Correspondence: ewy.mathe@nih.gov; Tel.: +1-301-402-8953

Received: 15 April 2020; Accepted: 13 May 2020; Published: 15 May 2020

Abstract: As researchers are increasingly able to collect data on a large scale from multiple clinical and omics modalities, multi-omics integration is becoming a critical component of metabolomics research. This introduces a need for increased understanding by the metabolomics researcher of computational and statistical analysis methods relevant to multi-omics studies. In this review, we discuss common types of analyses performed in multi-omics studies and the computational and statistical methods that can be used for each type of analysis. We pinpoint the caveats and considerations for analysis methods, including required parameters, sample size and data distribution requirements, sources of a priori knowledge, and techniques for the evaluation of model accuracy. Finally, for the types of analyses discussed, we provide examples of the applications of corresponding methods to clinical and basic research. We intend that our review may be used as a guide for metabolomics researchers to choose effective techniques for multi-omics analyses relevant to their field of study.

Keywords: multi-omics integration; dimensionality reduction; co-regulation; pathway enrichment; clustering; machine learning; deep learning; network analysis; visualization; biological pathways

1. Introduction

Biomedical researchers are increasingly relying on metabolomics and other omics data types to study and evaluate disease mechanisms and phenotypes. Omics data include, but are not limited to, measurements of the metabolome, proteome, transcriptome, genome, microbiome, and exposome. These measurements include the presence (binary), quantification (abundance), and/or characterization (chemical or biological function) of molecules or entities, such as metabolites, proteins, microbial taxa, genes, or transcripts. For simplicity, we refer to these molecules or entities as "analytes" throughout this work. Multi-omics data may also include descriptors from multiple timepoints in one or more omic modalities, phenotype information such as case/control labels, and relevant clinical variables such as age and sex. Collectively, these data provide holistic insights into disease-driven biological pathway dysregulation, which in turn provides preliminary evidence to drive the identification of new targets or intervention strategies [1]. While the utility of assessing multi-omics data is clear, the integration of metabolomic data with other omic data poses significant computational challenges that range from the need for developing statistical methods that are appropriately adapted to multi-omics integration, to the need for providing comprehensive open-source resources that provide validated relationships between omics types, biological pathways, and diseases. Multi-omics integration typically follows the general workflow depicted in Figure 1.

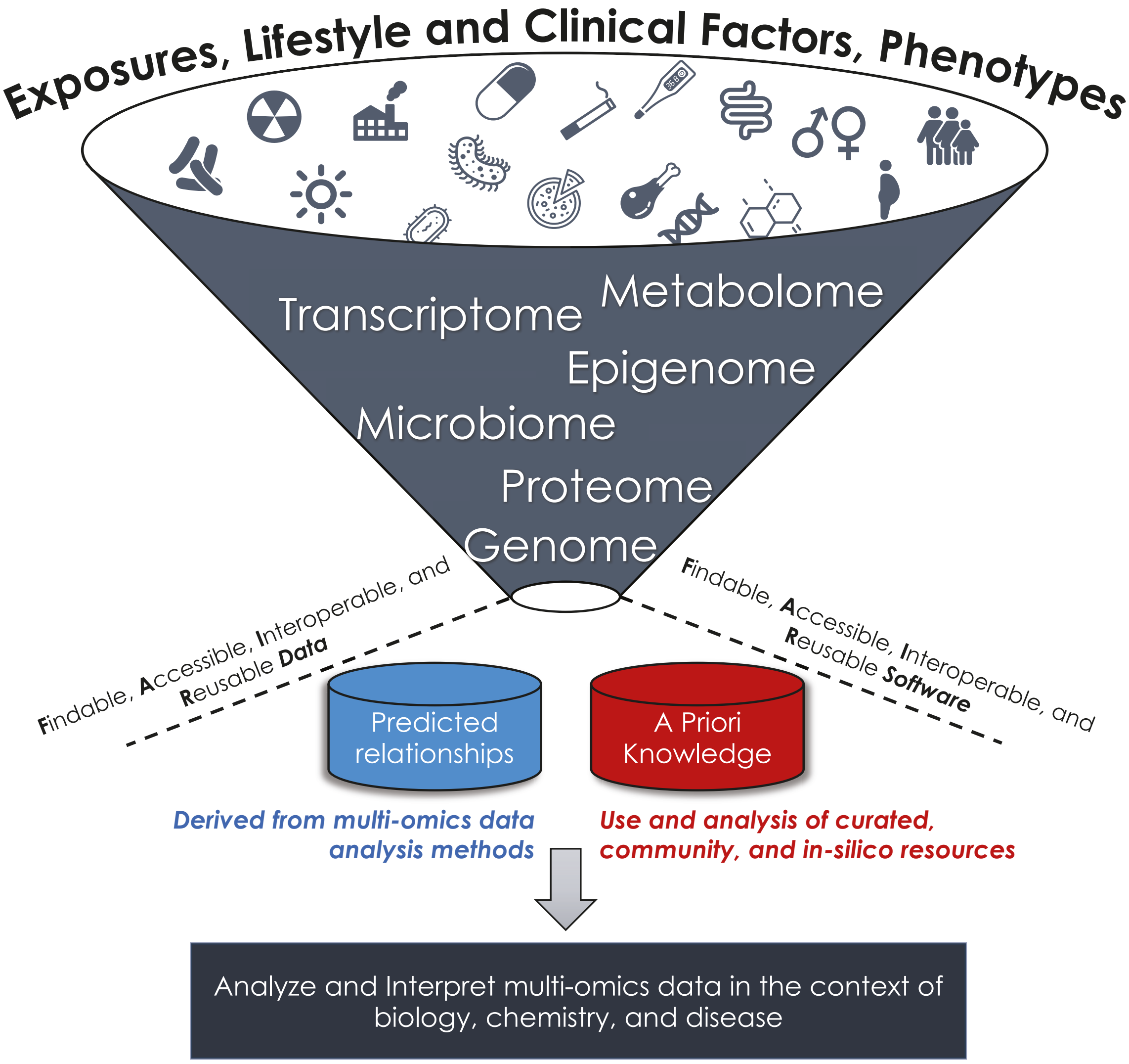
Exposures, Lifestyle and Clinical Factors, Phenotypes
Transcriptome
Metabolome
Epigenome
Microbiome
Proteome
Genome
Findable, Accessible, Interoperable, and Reusable Data
Findable, Accessible, Interoperable, and Reusable Software
Predicted relationships
A Priori Knowledge
Derived from multi-omics data analysis methods
Use and analysis of curated, community, and in-silico resources
Analyze and Interpret multi-omics data in the context of biology, chemistry, and disease

Figure 1. The metabolome in the context of other omics data types and broad approaches for their integration.

Ongoing efforts to support the integrative analysis of multi-omics data include the development of statistical methods, computational tools, and pipelines/workflows. Statistical and computational methods comprise novel metrics or novel applications of metrics that describe the relationship between multiple omic data. These include univariate and multivariate analyses, correlation networks, and traditional machine learning and deep learning techniques. A tool is an implementation of a method with proven utility, many of which are designed to be user-friendly software. Tools are often downloadable as an executable file or stored in a public code repository. A series of methods and tools can be combined into workflows to perform an analysis. Supported analyses could span the conversion of raw data (e.g., direct output from instruments or a matrix of un-normalized metabolite and gene levels) to interpretable data that explain a biological system under study. Workflows are particularly useful for conducting repetitive tasks, and typically provide default parameters that are globally applicable, hence making them user-friendly. Examples of open-source and user-friendly workflows include MetaboAnalyst [2], XCMS [3], mixomics [4], miodin [5], and many others reviewed elsewhere [1]. These workflows are particularly useful for end users that may not have a strong data analysis or computational background and are invaluable for outputting reproducible results. Ideally, methods, tools, and workflows provide up-to-date, publicly available datasets that can be used as an input for testing and benchmarking, allowing users to readily evaluate the utility of these workflows for their own purpose.

Many high-level analytical concepts are employed across workflows and data modalities. Understanding the general steps taken by many workflows is crucial to compare the different resources available for performing these tasks. Starting with raw data, prior to analysis, the data quality of each omic data must be carefully assessed to ensure that measurements are reproducible. This step typically requires a comparison of analyte measurements across technical replicates using metrics such as standard deviation or the coefficient of variation. Samples should also be evaluated, making sure that the overall distribution of analyte measurements is consistent across samples. We note that identification of potential outliers (analytes or samples) is critical, as some analytical models for multi-omics analyses (e.g., Principal Components Analysis and Student's *t*-test) could be strongly affected by the presence of outliers [6,7]. Other preprocessing steps may include normalization to account for differences in experimental effects such as differences in amounts of starting material and batch effects. Data are then typically transformed so that they follow a Gaussian or "Normal" distribution, which is commonly used for statistical analyses. Importantly, as some analyses will not work on missing data, missing values can be imputed. We note that the imputation method used can affect downstream analysis results [8,9], and thus imputation is still an active area of research [10,11]. Finally, noting that the range of values may differ between omic modalities, appropriate scaling (e.g., to a standard deviation of 1, *z*-scores) within and across omic datasets is critical for ensuring that each omic modality contributes to the analyses and that the effect of one omic modality does not dominate all analyses performed [12]. Special scaling considerations should also be taken for time-series data [13].

After preprocessing, omics data can be integrated in multiple ways. One can analyze or model each omic modality separately and then integrate results (a posteriori integration) or one can integrate data for all omic modalities before any statistical or computational modeling (a priori integration) [14]. Depending on which integration approach is utilized, data may be pretreated differently. For example, scaling analyte measurements appropriately within each omic modality is particularly critical when applying a priori integration. Additionally, the sample origin of the multi-omics datasets dictates which integration approach can be used. For example, a priori integration requires the measurements to be collected in the same biospecimens (tissue, blood, etc.) or individuals to allow measurements to be matched to the same sample, while a posteriori integration does not. When the analysis is performed on the same individual but different biospecimens, e.g., genomic data from blood and metabolomic data from urine, we note that it is not possible to evaluate direct relationships between genes and metabolites and how they may relate to phenotype. However, it is feasible to evaluate whether one omic modality (e.g., metabolites) could act as a biomarker for what is occurring at another level (e.g., genome), or one omic modality can be used to corroborate findings (e.g., biological pathways) uncovered in another omic modality [12].

Recognizing that other reviews provide a comprehensive list of available methods, software, and/or workflows [1,15–18], we instead focus on providing concepts and considerations that are useful when choosing a method or tool appropriate for one's desired data types and analyses. As such, we discuss existing guidelines in data curation and tool development and describe the building blocks that are used for developing computational tools and workflows, namely unsupervised clustering approaches to assess data quality or separation by sample type, approaches for modeling co-regulations between multiple omic modalities, approaches for identifying multi-omics factors associated with a phenotype (supervised methods), and methods that provide a biological, chemical, and/or disease context to multi-omics data (e.g., pathway analysis). We provide a summary illustration of these approaches in Figure 2.

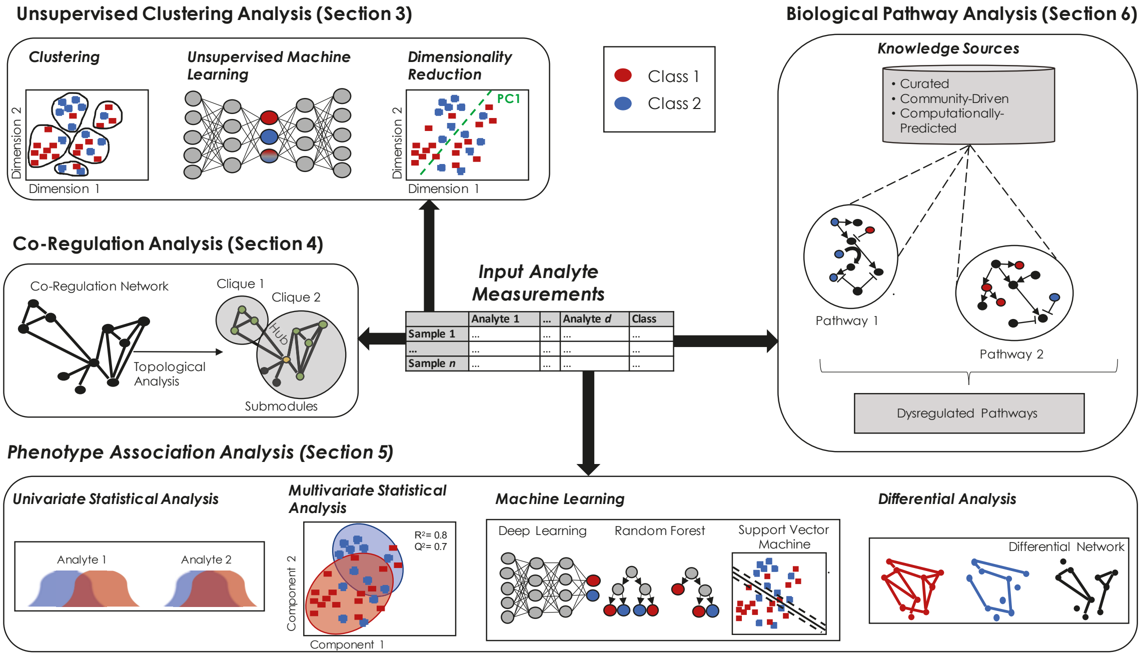

Figure 2. Analysis techniques on a dataset with n samples and d analytes in two classes. Blue represents one class of samples, and red represents another class of samples. Class typically corresponds to phenotype but could also correspond to batch or another variable of interest. Analyses include unsupervised clustering approaches (Section 3), modeling co-regulation (Section 4), approaches for identifying analytes associated with class (Section 5), and pathway analysis (Section 6).

Further, the review focuses on open-source resources, and we reference example research projects that make use of these resources in the context of metabolomic and other omic data, denoting whether and to what extent the data used is publicly available.

2. Open-Source Tool Development and Data Guidelines

The publication of the Findable, Accessible, Interoperable, Reproducible (FAIR) guiding principles for biomedical research data [19] and their adaptation to software development [20] provides clear guidelines to improve data and software infrastructure. Following these guidelines is critical to ensure the reproducibility and reuse of data and software resources. Currently, guidelines for producing tools and workflows exist, yet they are not widely adopted by the bioinformatics community as the tools and workflows are seldom developed by professional software developers [21,22]. Producing detailed documentation, example sets, and maintenance of tools and resources requires substantial resources and efforts, which are difficult to obtain through large biomedical research funding agencies. Open-source code should ideally include clear documentation detailing how to use the tool and providing example inputs to easily test the software. An analysis of publications in Oxford Presses' Bioinformatics revealed that roughly half of all publications examined had links to a code repository in their abstracts, mostly GitHub [23]. However, it is unknown how many of these repositories include documentation.

We note that the availability of high-quality, well documented, publicly available omics data to use as input for new proposed methods, tools, and workflows is critical to optimize the multi-omics research process. It has been shown that the majority of studies are not reproducible given the information disclosed in manuscripts. This could lead to intangible results. For instance, Begley and Ellis conducted an analysis of 53 high-impact pre-clinical cancer research papers and showed that only six were reproducible given the information provided in the paper, due either to lack of documentation or unpublished data [24]. While standards for reporting data analysis approaches are not well established, standards do exist for disclosing experimental data. Data standards include Minimum Information About a Microarray Experiment (MIAME) [25], the National Center for Biotechnology Information (NCBI)'s Minimum Information About a Next-generation Sequencing Experiment (MINSEQE), Metabolomics Standards Initiative (MSI) [26,27], and Minimum Information About a Proteomics Experiment (MIAPE) [28]. However, compliance is variable. For example, <50% of manuscripts published in journals requiring MIAME compliance actually met compliance [29]. For MSI compliance of metabolomics datasets, 90%–100% of clinical datasets include tissue or biofluid information, yet < 10% of these datasets include information about the ethnicity of the patient, location of collection, and/or volume of sample collection, and reporting of quality control metrics for in vitro datasets are largely missing [30].

Ideally, all multi-omics data from an experiment should be made accessible from the same location; however, the researcher must be aware of the constraints within a repository when submitting data. For instance, some of the most well-known repositories supporting multiple omic modalities are cancer-specific. These repositories include The Cancer Genome Atlas [31], the Cancer Cell Line Encyclopedia (CCLE) [32], the Therapeutically Applicable Research to Generate Effective Treatments Data Matrix [33], the Clinical Proteomic Tumor Analysis Consortium [34], and the NCI-60 Human Tumor Cell Line Screen (NCI-60) [35]. Of these, CCLE and NCI-60 support metabolomics. In scenarios where the data is not appropriate for an existing multi-omic repository, storing the data across multiple single-omic repositories can present an alternative solution. Single-omic repositories include MetaboLights [36] and Metabolomics Workbench [37] for metabolome data, the PRoteomics IDEntifications Database [38] for proteome data, Gene Expression Omnibus [39] and Sequence Read Archive [40] for genomic data, the Encyclopedia of DNA Elements [41] for functional elements in the genome, and the Human Microbiome Project Data Portal [42], MicrobiomeDB [43], and the Human Oral Microbiome Database [44] for microbiome data. When data is spread across multiple repositories, it can be challenging to identify and organize multi-omic datasets that are collected as part of the same study. Efforts that aim to consolidate different data sources as they pertain to publications, such as the Biostudies database [45], are useful for such multi-omic data identification.

3. Unsupervised Clustering of Samples to Assess Data Quality or Separation by Sample Type

Unsupervised analyses use algorithms that are agnostic to phenotype to learn the inherent distributions of the underlying data, discover relationships between analytes (regardless of phenotypes), or assess the overall quality of the data. These analyses may be executed on each omic modality separately and the results integrated using a posteriori techniques, or they may be executed on multi-omics data that has been integrated a priori. We describe methods that are commonly used for multi-omics data, especially focusing on the key role of metabolomics data. It should be noted that this list is in no way exhaustive. Examples of applications corresponding to these methods are given in Table 1.

Table 1. Examples of multi-omics applications using unsupervised analysis.

Type of Method		Functionality	Reference
Dimensionality Reduction	t-Distributed Stochastic Neighbor Embedding (t-SNE)	Visualize gut microbial communities and serum metabolites by diet and supplements.	[46]
Clustering	Hierarchical Clustering	Visualize prefrontal cortex metabolites and lipids by human population group.	[47] †
		Identify multi-omic molecular subtypes in hepatocellular carcinoma.	[48] ‡
		Identify multi-omic clusters in breast tumor tissue associated with prognosis.	[49] †‡
	k-means	Identify lipid–protein–metabolite clusters associated with diabetes and periodontal disease.	[50]
	Partitioning Around Medoids (PAM)	Identify microbial–metabolite clusters associated with diarrhea.	[51] *†‡
	Gaussian Mixture Modeling (GMM)	Identify clinical depression score clusters associated with blood metabolomic and genomic data in blood to predict drug response.	[52] ‡
	Density-Based Spatial Clustering of Applications with Noise (DBSCAN)	Evaluate the impact of bacterial metabolism on mucosal immunity.	[53]
Other Machine Learning Methods	Random Forest	Identify clusters of histological stromal features associated with prognosis and metabolites in cancer-associated fibroblasts.	[54] ‡
	Autoencoder	Cluster plasma protein and metabolite levels to identify temporal trends in murine cardiac remodeling.	[55]

* Raw data are available in the supplementary of the referenced manuscript, or a public repository. † Preprocessed data are available in the supplementary of the referenced manuscript, or a public repository. ‡ Descriptive statistics are available in a table or supplementary materials of referenced manuscript. Unmarked data are available upon request from the authors or from a consortium.

3.1. Dimensionality Reduction

The number of dimensions in a dataset generally refers to the number of analytes measured. In multi-omics workflows, the number of dimensions is typically much larger than the number of samples, a phenomenon known as the curse of dimensionality problem. This phenomenon can lead to overfitting in downstream models, where models may not reproduce in other datasets. Reducing the number of dimensions in the dataset can help mitigate this issue. Principal Components Analysis (PCA) is a commonly used method that accepts a matrix of analytes and samples as input, and reduces the dataset to fewer dimensions, or components, that capture the largest variance in the data. The data can be projected on the first two or more components, thereby potentially revealing clusters of samples. By labeling the samples, for example by batch or phenotype, one can identify clusters that may help evaluate data quality. When using PCA, it is important to report the percent variance explained by each component so that the number of components that capture most of the variance can readily be determined. Sample loadings can be incorporated into PCA that reflect the extent of a variable's contribution to a component. Noting that a component may separate samples by a phenotype or other metric of interest, loadings can be used to identify phenotype-associated analytes in an unsupervised fashion. When the goal is to assess data quality, PCA is performed on individual omic types, and for metabolomics, data collected from different instruments or ionization modes should be evaluated separately. When the goal is to look for separation and loadings, data can be combined, although care should be taken to appropriately scale the data, particularly since the dynamic ranges of metabolomics data can vary greatly and few datapoints could easily dominate the components. Because of potential differences in the variance of analytes from different omic modalities, it is pertinent to factor in contributions of individual modalities to the final loadings. Multi-Omics Factor Analysis (MOFA) [56] and Multiple Co-Inertia Analysis (MCIA) [57] are two techniques for doing so.

Multi-Dimensional Scaling (MDS) is another related, and commonly used unsupervised dimensionality reduction technique. The input for MDS is a distance matrix representing pairwise distances between samples (or analytes) [58]. Examples of commonly used distance metrics include Euclidean distance and 1-correlation for relative abundance data (transcriptomics, metabolomics, etc.), and the Jaccard and the Bray-Curtis for binary data (e.g., microbiome, genomic variants) [59].

T-distributed Stochastic Neighbor Embedding (t-SNE) is another technique used for dimensionality reduction and visualization [60]. Like MDS, t-SNE attempts to produce a lower dimensional embedding of high-dimensional data where distances in the embedding represent similarities between samples. However, rather than using linear correspondence of similarities directly, t-SNE adopts a non-linear adaptive approach based on matching Gaussian probability distributions over similarities. t-SNE can be very helpful for visualizing complex geometry observed in higher-dimensional spaces. However, t-SNE should be used and interpreted with caution, because changes in input hyperparameters (e.g., perplexity) can produce radically different plots, and misleading apparent clusters can result from random data [61].

3.2. Clustering

Clustering methods are often used to group samples and/or analytes together by shared characteristics (e.g., abundances, presence/absence). Hierarchical clustering assembles clusters of related samples using either an agglomerative “bottom-up”, or divisive “top-down” methodology [62]. An agglomerative clustering algorithm starts with each observation in the dataset belonging to separate clusters. Each iteration combines clusters based on their similarity, and the algorithm stops when all observations belong to one cluster [62]. In contrast, a divisive algorithm starts with one cluster, which is iteratively divided into many. Hierarchical clusters can be visualized as dendrograms, and users can “cut” the dendrogram to produce a desired and relevant number of clusters.

Other methods require that the user specify the number of expected output clusters prior to running the algorithm. One of these is *k*-means [63], which aims to divide all samples into well-separated clusters, where the number of clusters is specified by the input *k*. A related method, Partitioning Around Medoids (PAM) [64] is similar to *k*-means, but can take dissimilarity matrices as input. Silhouette plots, which measure the ratio of within-cluster similarity to between-cluster similarity for differing values of *k*, can be used to help the user determine the number of clusters that best matches the data [65]. Another metric that can be used to determine *k* is the gap statistic, which computes within-cluster dispersion for each value of *k* and subtracts from it the expected within-cluster dispersion for the same value *k* on a uniform distribution of observations [66]. The optimal value of *k* is the one that maximizes this difference.

Another method, the Self-Organizing Map (SOM) is a single-layer neural network with nodes laid out in a grid [67]. Each node has weights that are trained to emulate the analyte abundance/expression values of a set of observations that are similar to one another. Training is done by a mapping process where each observation is mapped to its best matching unit/node in the grid, and that unit and its neighbors in the grid are updated to reflect the features observed. After training is complete, each unit represents the centroid of its own cluster. SOM requires specification of the grid size (number of units) by the user. The user must also specify the learning rate, or the rate at which node weights update in each iteration of the algorithm, and neighborhood size, or the size of a node’s sphere of influence on its neighbors [67]. Specialized metrics for evaluating the results of SOM include map embedding accuracy [68] and topographic accuracy [69]. We note that Milone et al. developed a specialized tool for integrating the metabolome and transcriptome in plant studies using SOM [70].

While the methods described above result in all points being assigned to a cluster and do not allow overlap between clusters, this is not the case for all clustering methods. Gaussian Mixture Models (GMM) assume that all the data are generated from a mixture of Gaussians, thereby allowing overlap between clusters [71]. GMM require an upper bound on k, the number of clusters, and can be evaluated for performance (in terms of information captured) using the Bayesian Information Criterion [72] or the Akaike Information Criterion [73]. Another method, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), is designed to detect clusters of arbitrary shape and automatically excludes some observations from clustering by designating them as noise [74]. DBSCAN does not require specification of the expected number of clusters a priori, although it requires the user to specify the minimum number of observations per cluster and a minimum distance between adjacent observations within a cluster.

3.3. Other Machine Learning Methods

Other machine learning methods, such as random forest (RF) and support vector machines (SVM), that are typically run in supervised modes can be used in an unsupervised manner to characterize data structure. RF algorithms combine many different decision trees that are built on a subset of the samples and features (e.g., analytes) [75]. In unsupervised mode, synthetic data are generated, and RF is trained to differentiate between the real data and the synthetic data, where class (e.g., phenotype) labels are randomized. The algorithm then tracks the number of times two samples are placed into the same terminal node by the trees, which is then weighted and used as a distance metric between all pairs of samples. This distance metric can then be used as input to the clustering methods described above. Support Vector Machines (SVM) [76] can also be run in unsupervised mode, reviewed in [77]. The algorithm learns to separate samples with random class assignments and then iteratively changes the class assignments to optimize the classification accuracy. This results in an optimally separated set of clusters, where cluster membership corresponds to the random class assignment learned by the algorithm.

Autoencoders are another type of unsupervised learning method used in omics data exploration; they are used for learning a set of latent variables that can be used to reconstruct data [78]. Autoencoders are a type of neural network in which the number of nodes per layer is highest in the first and last layer and lowest in the middle, where the middle layer is called the encoding and is the representation of latent variables. The encoding can be used for clustering. The reconstruction error can then be used as a measure of performance. While autoencoders can learn the complex latent variables underlying the data, a downside is that autoencoders, like neural networks in general, can be difficult to interpret [79].

Although other types of deep learning approaches to clustering also exist, these have not been used in multi-omics applications, to our knowledge. However, their usage in general bioinformatics research is reviewed in [80].

3.4. Time-Series Data

Clustering of time-series omic data is most easily accomplished when aggregating the individual feature profiles into clusters (e.g., clustering individuals which have similar time-profiles of a single metabolite or clustering genes in a single individual based on similar expression changes). When using multi-omic datasets, features from separate modalities can be concatenated. In this case, any of the above methods can be straightforwardly applied to the vector representations of the time-series. Unfortunately, the simplicity of this technique comes at the cost of discarding information which could be gained by considering time as a continuously varying, or even ordered, dimension.

A common way that a continuous notion of time can be directly incorporated into clustering techniques is by fitting a model and then clustering on model parameters. Fitting splines, a type of piecewise polynomial curve, is a popular approach [81,82]. Models based on the assumed statistical properties of the processes which create the time series have also been explored, such as auto-regressive moving average (ARMA) models [83,84] or hidden Markov models (HMM) [85,86]. Machine learning techniques developed to process data collected in one timepoint can similarly process time-series into latent vectors which can then be clustered [55].

Notably, the choice of distance metric used may greatly affect clustering [87], and metrics that consider possible time-lags between series can provide a more biologically relevant notion of profile similarity. For example, Dynamic Time Warping (DTW) [88] aligns timepoints so that the distance between the aligned samples is minimized and Lag Penalized Weighted Correlation (LPWC) [89] includes a penalty based on the length of lag between series.

4. Identifying Groups of Multi-Omics Analytes that are Co-Regulated

Assessing relationships between metabolites and other analytes could shed light on the mechanisms that underlie a given phenotype. Researchers may wish to assess relationships within a single omic modality (integrating modalities using further analytical methods) or across multiple omic modalities. These relationships are typically not causative, and statistical associations between analytes do not necessarily capture direct, physical relationships, and often ignore complex relationships such as post-translational modifications and non-linear reaction kinetics. For instance, Camacho et al. observed that correlations between metabolite abundance levels could arise from both metabolites being near chemical equilibrium or from a large concentration response to a common enzyme, whereas negative correlations could result from two metabolites being part of the same moiety-conserved cycle [90]. Nonetheless, it is feasible that associative networks capture associations that are functionally relevant. Examples of applications that assess co-regulations of metabolites are provided in Table 2.

Table 2. Examples of multi-omics applications using co-regulation analysis.

Type of Method		**Functionality**	**Reference**
Associative Networks	Correlation Networks	Find metabolite–metabolite associations specific to or shared across blood, urine, and saliva.	[91] †
		Find modules of blood metabolites and genes associated with body weight change.	[92] ‡
		Find associations between serum, blood, and gut antibodies, metabolites, and microbiome and patient disease activity reports in inflammatory bowel disease.	[93] *†‡
		Find associations between metabolites, transcripts, cytokines, and cell frequencies in plasma and whole blood associated with adaptive immune response to *Herpes zoster* vaccine.	[94] †‡
	Partial Correlation Networks	Visualize associations between sleep survey responses and levels of serum cytokines, metabolites, lipids, proteins, and genes.	[95] *‡
		Visualize associations between metabolites and lipids associated with metabolic disease treatment in rat liver tissue and clinical chemistry measurements from serum.	[96] †
	Weighted Gene Co-Expression Network Analysis (WGCNA)	Characterize complex transcriptomic and metabolic traits in major depressive disorder.	[97] ‡
		Identify co-regulated modules of blood metabolites and transcripts in children with asthma.	[98] ‡
		Identify co-regulated modules of metabolites and transcripts in glioblastoma multiforme.	[99]
Topological Analysis of Networks	Subnetworks	Identify subnetworks of correlated proteins and metabolites in adrenocorticotropic hormone-secreting pituitary adenomas.	[100]
		Identify subnetworks of correlated genetic, proteomic, metabolomic, clinical, and microbiome data from multiple biofluids in cardiometabolic disease.	[101] ‡

* Raw data are available in the supplementary of the referenced manuscript, or a public repository. † Preprocessed data are available in the supplementary of the referenced manuscript, or a public repository. ‡ Descriptive statistics are available in a table or supplementary materials of referenced manuscript. Unmarked data are available upon request from the authors or from a consortium.

4.1. *Associative Networks*

Associative networks are built to model relationships between analytes. One example of such networks, correlation networks, or "relevance networks" [102], can be used to evaluate relationships between differentially expressed analytes. In these networks, nodes represent analytes, and edges represent significant correlations between analytes (for example, based on a p-value and/or effect size cutoff). Such networks can be built to infer groups of analytes within or across omic modalities that could regulate one another.

Alternatively, partial correlations can be used in lieu of correlations, where "partial" refers to the correlation between analytes, while removing the effect of other covariables. For example, partial correlations can be used to identify the major influencers (e.g., enzymes) of metabolites whose levels are highly correlated [103]. They can also be used to sparsify relevance networks by conditioning correlations between each pair of analytes on the values of other analytes in the dataset and retaining only those "relevant" and "independent" correlations that do not depend on the values of other analytes [104].

Identifying clusters of coregulated metabolites can be used as a feature selection step to select analytes of interest. For example, Weighted Gene Co-Expression Network Analysis (WGCNA) is commonly applied to study relationships between clusters of samples in high-dimensional datasets [105]. WGCNA identifies clusters of highly correlated analytes in samples and builds dendrograms of these clusters for the user to investigate further [105]. These clusters can then be evaluated for biological relevance.

4.2. *Topological Analysis of Networks*

Given a large number of analytes in omic datasets, networks produced can be very complex, oftentimes described as 'hairballs' [106]. To simplify interpretation, the topology of the network can be evaluated. For example, one can evaluate the significance of nodes by identifying hub nodes or nodes that have many connections and thereby contribute considerably to the topology of the graph. The identification of hubs has been used to study gene essentiality [107] and protein robustness to knockdown [108]. Metrics such as degree of a node (the number of nodes to which the node is connected) or betweenness-centrality of a node (the number of node pairs whose shortest path passes through the node) can be used to find hubs, as discussed in Jalili et al.'s review on the topic [109]. However, the existence of hubs must be interpreted carefully. Hubs may represent analytes that are abundant in a cell or that have interactions with or relationships to many other analytes, but this does not necessarily mean that they are relevant to the experimental context of the study [110].

Global network characteristics are also potentially useful. The small-world structure, i.e., the tendency of nodes to be connected by paths of short lengths, of analyte networks can also be informative for inferring the evolutionary history of a metabolic network [111] and the level of communication between substructures in the network [110]. In addition, topological analysis can produce submodules, which represent tightly connected substructures which are oftentimes biologically relevant [112]. A strict definition of a submodule is a clique, which refers to a group of analytes that all share an edge with all other analytes in that group. This definition has been used in the single-omic context to find groups of correlated microbiome samples in Crohn's disease [113]. Methods for the detection of submodules that are tightly connected, but that need not be cliques as such, include module graphical Least Absolute Shrinkage and Selection Operator (LASSO), which has been used for gene expression data but can be extended to multi-omics contexts [114] and Louvain community detection [115]. Submodules can also be detected by spectral clustering, which has been applied to expression quantitative trait loci (eQTL) data but could be extended to multi-omics contexts [116], and by clique conductance, defined by the sizes of and connections between cliques in the graph [117]. Multilayer N-Cut (MuNCut) [118] was developed specifically for multi-omics networks, and aims to optimize submodule detection by minimizing the "cut" (i.e., the sum of weights of the edges removed) as compared to the size of the submodules (i.e., the sum of weights of all edges in the submodules) across multiple omic modalities.

5. Identifying Multi-Omics Analytes Associated with Phenotype

Identifying analytes associated with a phenotype can be done in a variety of ways, including by finding differentially expressed analytes between sample groups, by exploring relationships between analytes and how these relationships differ between phenotypes, and by modeling the direct relationship between sets of (possibly related) analytes and a phenotype. These methods are described below, and applications corresponding to these methods are given in Table 3.

Table 3. Examples of multi-omics applications that identify analytes associated with phenotype.

Type of Method		Functionality	Reference
Univariate Statistical Methods	Student's *t*-test and effect size	Identify metabolites, miRNAs, mRNAs, and lncRNAs altered by exposure to benzo[a]pyrene to identify mechanisms of toxicity.	[119]
Multivariate Statistical Methods	Partial Least Squares Discriminant Analysis (PLS-DA) (and variants)	Identify breast tumor tissue metabolites that differentiate MRI features.	[120] ‡
		Identify metabolites that differentiate normal and tumor tissue in the prostate.	[121] ‡
		Identify differences between fibromyalgia and control groups in gut microbes, serum metabolites, miRNA, and cytokine levels.	[122] *‡
		Discover temporal changes in plasma lipid and metabolite patterns from normal and hyperlipidemic patients.	[123] †
	Linear Models (and variants)	Identify metabolites from bronchial alveolar lavage associated with continuous CT scan features in cystic fibrosis.	[124] ‡
		Identify serum metabolites associated with visceral adipose tissue features from MRI and tomography.	[125] ‡
		Identify plasma metabolites and proteins associated with prognosis in septic shock patients.	[126] ‡
		Find associations between blood DNA methylation and metabolite levels in smokers.	[127] ‡
Identifying Analyte Relationships that Differ by Phenotype	DiffCorr	Identify differences in metabolite-metabolite correlations between traumatic brain injury and control groups.	[128]
	IntLIM	Identify synovial fluid metabolites and blood and bone marrow transcripts that differentiate between osteoarthritis and rheumatoid arthritis.	[129] *
Machine Learning Methods for Predicting Phenotype	Random Forest	Identify serum metabolites, proteins, and peptides differentiating between metabolic syndrome and control groups.	[130]
		Identify metabolites and other analytes predictive of weight gain and loss.	[131] *‡
		Identify metabolites, transcripts, and proteins predictive of potato quality traits.	[132] †
		Identify metabolites and transcripts predictive of heat stress in the liver.	[133] †
	Support Vector Machine (SVM)	Predict metabolite levels using genes and metabolites in breast and hepatocellular carcinoma.	[134]
	Multilayer Perceptron (MLP)	Predict early and late stage bladder cancer using urinary metabolites and genes.	[135]
		Predict early renal injury using serum metabolites and lipids.	[136] †‡
	Convolutional Neural Network (CNN)	Predict early renal injury using serum metabolites and lipids.	[136] †‡
	Recurrent Neural Network (RNN)	Integrate transcript and metabolite levels to predict cellular state in *Escherichia coli.*	[137] *†

* Raw data are available in the supplementary of the referenced manuscript, or a public repository. † Preprocessed data are available in the supplementary of the referenced manuscript, or a public repository. ‡ Descriptive statistics are available in a table or supplementary materials of referenced manuscript. Unmarked data are available upon request from the authors or from a consortium.

5.1. Identifying Differentially Expressed Analytes (Univariate Statistical Methods)

Hypothesis testing methods determine, for each analyte, whether to accept the null hypothesis that a statistic of the analyte's distribution (such as mean or variance) is unrelated to the phenotype. These methods can be used for analytes across multiple omic modalities or within a single omic modality. These methods can be either parametric or non-parametric in nature. Parametric tests require the user to input data that fit a distribution, such as a "Normal" or Gaussian distribution, whereas non-parametric tests do not impose requirements on the underlying data distributions. Well-known examples of parametric and non-parametric hypothesis testing include the Student's *t*-test and the Wilcoxon Rank-Sum test, respectively. While these methods are restricted to comparing two phenotypes, one-way analysis of variance (ANOVA) (parametric) or a Kruskal–Wallis test (non-parametric) can handle multiple phenotypic groups.

Univariate tests result in *p*-values, which quantify the probability of the null hypothesis being correct given the observed data. When evaluating many analytes, *p*-values must be corrected to account for multiple comparisons to reduce the number of false positives. Examples of methods that can be used for multiple comparison correction include the Family-Wise Error Rate (FWER), such as Bonferroni correction, and False Discovery Rate (FDR), which includes the Benjamini–Hochberg procedure [138]. For a review of these methods and the challenges inherent in multiple comparisons in the omics space, we direct the reader to [139]. Particularly relevant for multi-omics analyses, Karathanasis et al. [140] developed a method for combining results of hypothesis testing applied to separate omic modalities (referred to as "partial p-values") in which an underlying permutation test is used that addresses correlations between the omic modalities. Other methods tailored to multiple comparison corrections in omics data include the tail statistic, which is based on an expected distribution of *p*-values [141,142], and an Empirical Bayes method originally developed for microarrays but applicable to other omic data [143].

It is well known that using a *p*-value cutoff of 0.05 (or other values) is subjective and that the interpretation of *p*-values has been mishandled [144]. It is useful to also consider effect size, such as fold changes, when determining which analytes are most relevant to a phenotype of interest. A common visualization method, the volcano plot, combines both *p*-values and fold change, highlighting analytes that have both high fold change and low *p*-values between two phenotypic groups.

Univariate analysis of omic data in a time series, as opposed to a single timepoint scenario, is complicated by the need to account for multiple timepoints. One approach to dealing with multiple timepoints is to collapse them into a single summary value before testing for differences between groups. Examples of this technique include calculating a per-individual mean, area-under-curve [145], time-at-maximum [145], or slope in linear regression [146]. The summary value can then be tested for significance using any of the above techniques. By treating time as a discrete effect, the existence of differential time profiles can be tested for directly with a two-way extension of ANOVA [147]. Two-way ANOVA directly tests for statistically different distributions by either experimental condition or time point, or, importantly, for a statistically significant interaction between the condition and timepoint. MetATT [148] supports the comparison of time-course profiles while allowing for variability both within and between timepoints, thereby reducing false positives and false negatives. The fitting of more complex curves to time-course data will generally require statistical tests specific to the type of curve fit. For instance, Berk et al. developed a modified F-statistic for significance testing in their smoothing splines mixed effects (SME) model [149].

5.2. Multivariate Statistical Methods

Multivariate methods are slightly more complex (and more informative) than univariate methods in that they consider possible dependencies between analytes and the effects of possible confounders. These methods may either be run on each omic modality separately or on integrated omic data. One common class of multivariate method is Partial Least Squares Discriminant Analysis (PLS-DA). PLS-DA can be thought of as a supervised variation of PCA, where instead of projecting the data to dimensions that maximize the overall variance, data are projected to maximize the covariance between the projected data and phenotype. Notably, PLS-DA is prone to overfitting [150], as it will always find a projection that separates phenotypes, even with randomized data [151,152]. Metrics that evaluate overfitting, such as R^2, Q^2, number of misclassifications (NMC), and Area Under the Receiver Operating Characteristic (AUROC), must be then be evaluated carefully. These metrics are compared for statistical significance in [153].

Variations of PLS-DA include the Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) [154], which removes variation in the set of analyte abundances that is unrelated to the phenotype. Another variation is Sparse Partial Least Squares Discriminant Analysis (SPLS-DA) [155], which ensures that the number of analytes contributing to the model is relatively small compared to the total number of analytes in the input. We note that both OPLS-DA and SPLS-DA suffer from the same overfitting drawback as PLS-DA.

Another class of multivariate methods uses linear models [156]. Linear models assume a linear relationship between a continuous response variable (e.g., analyte levels) and one or more independent variables or covariates (e.g., phenotype). These models are learned by minimizing the error between a predicted response variable and the true response variable, and produce weights, also called coefficients, for each independent variable that indicates the influence or significance of independent variables to the response variable. An extension of this type of model is the linear mixed effects model, which assumes that some independent variables contribute random effects to the model, where the coefficient of that independent variable is randomly drawn from a distribution rather than being a fixed coefficient. As with linear models, linear mixed effects models assume that the relationship between a combination of covariates (e.g., age, gender, or batch) and expression or abundance of each analyte is linear. The effect sizes (fold change) between the actual expression levels and those predicted by the linear model can be computed across sample groups to obtain a list of differentially expressed analytes.

Linear models can be extended to include other types of regression models with non-linear functions. In logistic regression models, such as Semi-Parametric Differential Abundance analysis [157], a sigmoidal function, rather than a line, is learned to fit the data. Linear models also may include regularization for enforcing sparsity (i.e., reducing the total number of analytes determined to differentiate between groups) or discouraging overfitting of the model to the data. One such method developed for multi-omics data is collaborative regression [158], in which the model is learned by minimizing error between each omic modality and the phenotype and between linear combinations of separate omic modalities. Additional standard regularization terms include ridge regression [159], which minimizes the sum of squared coefficients, and Least Absolute Shrinkage and Selection Operator (LASSO) [160], which minimizes the sum of absolute valued coefficients; the elastic net combines both regularization terms [161].

We note that, like PLS-DA, linear models and their variations will always learn a model that separates phenotypes, but the model may not necessarily be robust. The coefficient of determination (the proportion of output variance determined by input) or the root-mean square error (the standard deviation of output prediction error) can be used to measure the robustness of the model.

5.3. Identifying Analyte Relationships that Differ by Phenotype

Identifying analyte relationships that differ by phenotype can shed light on phenotype-specific mechanisms. Various methods and tools aim to identify phenotype-specific pairs of analytes within one or more omic modalities. For example, DiffCorr calculates correlation coefficients between pairs of analytes within each group and compares correlation coefficients between categorical phenotypes. It does this by transforming correlation differences between phenotypes into *z*-scores to test their statistical significance [162]. The Discordant Method [163] bins analyte pairs with discordant relationships, identified through a mixture model, into categories based on the type of differential relationship (e.g., positively correlated in Group 1 and negatively correlated in Group 2, positively correlated in Group 1 and no correlation in Group 2). Differential Network Enrichment Analysis [164] computes partial correlation networks across multiple phenotypes, and then finds network modules that differentiate between phenotypes. Finally, Sparse Multiple Canonical Correlation Network Analysis (SmCCNet) extends canonical correlation analysis methods by considering phenotypes when evaluating relationships between two omics modalities [165]. Other methods, such as IntLIM, capture phenotype-specific analyte relationships based on linear models that test interactions between a phenotype and an independent variable (e.g., analyte) [166].

5.4. Causative (Flux-Balance) Networks

Flux-balance networks are built on experimentally derived associations between analytes for predicting biomass. These networks are particularly useful for predicting the effects of perturbing certain nodes/analytes of the network. Flux-balance analysis relies on the principle that modeling the concentration of biomolecules is mathematically equivalent to modeling flux [167]. These networks thus use a set of equations relating reaction to biomass that can be solved using linear programming to determine which reactions are essential given quantitative (not relative) analyte abundances. While relative abundances are not used to construct these causative networks, they have recently been shown to enhance flux prediction [168]. The current state of flux-balance analysis in the multi-omics space and software implementations for such analyses are reviewed elsewhere [169,170].

5.5. Machine Learning Methods for Predicting Phenotype

Machine learning methods can be used to predict phenotype given analytes from a single omic modality or multiple omic modalities that have been integrated a priori. Unlike in multivariate statistics, machine learning methods do not require a priori selection of confounders or multi-analyte dependencies, as they model dependencies directly from the data. Like statistical models, machine learning models make assumptions that differ based on the model type. The machine learning models used in multi-omics integration include both traditional machine learning and deep learning models.

Traditional machine learning methods include Support Vector Machines (SVM) [76] and Random Forests [75]. While the application of these methods can also uncover global data structures in their unsupervised forms, as described in Section 3, we describe here their supervised functionality. SVM assumes that two phenotypes are separated by a hyperplane, which is a linear combination of analyte characteristics (e.g., levels). SVM algorithms learn the hyperplane that optimally separates two phenotypes. SVMs can also be extended to learn non-linear separators using the kernel trick [76], yet we note that these are more difficult to interpret than linear hyperplanes [171]. From SVM models, one can also evaluate the contribution of each analyte to separating the optimal hyperplane, which is calculated as the magnitude of the linear weights of the hyperplane [172–174]. Another approach to decipher the analytes most relevant to the model is to consider both the weight and margin between the hyperplane [175].

Another popular machine learning model is the Random Forest (RF) [75], which is based on decision trees. Each tree represents a branched chain of “decisions”, where the decision to branch right or left in the tree is based on one feature (e.g., one analyte) and an optimal cutoff for that feature (e.g., abundance level cutoff). Each decision tree is optimized using metrics such as Gini impurity [176] or information gain [177]. Each RF model constructs many decision trees using subsets of the input samples and subsets of analytes. The predictions of all decision trees are combined into an ensemble to obtain the final output. Like SVM, various metrics exist to determine the influence of each analyte in the RF model. Generally, metrics either evaluate the decrease in model fitness (e.g., Mean Decrease Accuracy [178] or Mean Decrease Gini [179,180]) when features are removed, or compare the influence of analytes on the model using the original values and shuffled values [181].

To evaluate the extent of separation by phenotype, one can use the information retrieval metrics precision, recall, percent accuracy, and F-score, reviewed in [182]. To visualize overall model performance, receiver operating characteristic (ROC) can be plotted [183]. Lastly and like other statistical learning approaches, such as PLS-DA, all models are prone to overfitting the data. To ensure models are not overfitting, samples should be split into training and testing datasets, where the training samples are used to fit the model and the testing set is used to test predictions made by the model. Ideally, a completely independent validation set of samples should be used as an additional model evaluation.

Deep learning models consist of multiple functions of the input data or subsets thereof, that feed into each other in a series of layers, with the outcome (e.g., phenotype) being predicted in the final layer. There are many neural network architectures that can be used for deep learning, including multilayer perceptrons (MLP) [184], convolutional neural networks (CNN) [185], and recurrent neural networks (RNN) [186]. The breadth of each layer, the number of layers, and the function of inputs used at each layer are customizable to a large extent. Neural networks are sometimes referred to as universal function approximators because they can approximate a broad spectrum of underlying models. However, these deep learning models are difficult to interpret and require many samples to accurately train: for example, recent estimations on Monte Carlo simulated data for MLP estimate a requirement of 50 times the number of adjustable parameters in the network (e.g., number of analytes, number of nodes per layer, and number of layers) [187]. Because omic data typically suffer from the curse of dimensionality problem, where the number of samples is far lower than the number of analytes and hence the dimensions, thereby increasing the risk of overfitting, the application of neural networks is a challenge in multi-omics contexts [188]. Nonetheless, neural networks have been successfully applied in some multi-omics studies, as shown in Table 3. Additionally, Yu et al. explored the use of MLP and CNN architectures for classification on 37 transcriptomic and metabolomic The Cancer Genome Atlas (TCGA) datasets, finding that MLP outperformed CNN in this case [189]. Although other deep learning architectures exist in addition to those described here, they have not been applied in the multi-omics context, to our knowledge. However, the use of other deep learning methods in bioinformatics research in general is reviewed in [190].

6. Interpreting a List of Phenotype-Related Analytes in the Context of Biology, Diseases, or Chemistry

Identifications of analytes or analyte relationships that reflect a phenotype of interest are typically not useful unless the biological, disease, or other relevant contexts are considered. Common methods to guide the biological interpretation of these data include identifying enriched pathways, and visualizing relationships between analytes. Applications using these methods are outlined in Table 4.

Table 4. Multi-omics applications using biological or visual interpretation methods.

Type of Method		Functionality	Reference
Pathway enrichment methods	Overrepresentation Analysis (ORA)	Identify dysregulated pathways in prostate tumor tissue using metabolite and transcript data.	[191]
		Identify dysregulated pathways in the murine hippocampus and left ventricle during proton irradiation using metabolite and DNA methylation data.	[192]
		Identify dysregulated pathways in cationic liposome treatment of human hepatocyte cells using metabolomic and proteomic data.	[193]
		Identify dysregulated pathways in kidney disease in the rat serum metabolome and proteome.	[194] ‡
		Identify dysregulated gut microbial pathways in gastrectomy patients.	[195] *‡
		Identify dysregulated gut microbial pathways in sports classification groups of Irish athletes.	[196] *‡
		Identify dysregulated gut microbial pathways as a result of whey protein supplementation.	[197] *‡
	Topological Scoring	Identify functional connections between dysregulated pathways in Alzheimer's using genes, metabolites, miRNA, and proteins from multiple sources.	[198]
Visualization of biological pathways and networks		Visualize metabolic networks in drug-susceptible and drug-resistant strains of *Acinetobacter baumannii.*	[199]
		Visualize interactions between metabolites and genes in non-small cell lung cancer.	[200]

* Raw data are available in the supplementary of the referenced manuscript, or a public repository. † Preprocessed data are available in the supplementary of the referenced manuscript, or a public repository. ‡ Descriptive statistics are available in a table or supplementary materials of referenced manuscript. Unmarked data are available upon request from the authors or from a consortium.

6.1. Pathway Enrichment Analysis

Identifying enriched pathways is a common approach to the biological interpretation of differentially expressed analytes, and numerous approaches exist for this type of analysis, each with advantages and disadvantages [201]. Not only do enriched pathways add interpretability to the data, but dysregulations of pathways are also more reproducible across samples than altered levels of individual analytes. [1,202,203]. While published pathways remain the gold standard for context relevance, pathway analysis tools produce relevant results when publications are sparse [204].

Overrepresentation Analyses (ORA) are based on the Fisher's or Hypergeometric test and are commonly used to identify enriched pathways. Broadly, these methods test the hypothesis that a given pathway is associated more frequently with analytes in the list of interest than would be expected by chance. A major caveat in ORA is the dependence of the result on the background set of analytes (e.g., all analytes measured, or all analytes in a pathway database) used for each pathway [205,206]. In metabolomics, pathway coverage of different metabolite classes is unequal [207]. For example, lipids suffer from lower pathway annotation coverage than other metabolite classes due to the structural complexity of lipid species [207]. Unequal coverage leads to issues in conventional enrichment testing, because the test result is biased towards annotations that are uncommon in the database. When multiple types of analytes are input into ORA analyses, the *p*-values resulting from analysis of each analyte independently can be combined using Fisher's method [208] or Stouffer's method [209], which do not penalize an analyte type that has fewer annotations given a particular pathway. This approach is readily available in various software [2,210,211], and an evaluation of both methods is provided in [212]. Other issues of ORA include the erroneous assumption that pathways are independent from one another, and the reliance on an arbitrary statistical cutoff (e.g., *p*-value) to identify enriched pathways.

Another set of pathway-enrichment methods uses Functional Set Enrichment Analysis (FSEA), which is based on the Kolmogorov–Smirnov (KS) test. FSEA methods were developed to address two drawbacks of ORA: statistical cutoffs and sensitivity to background distribution. Rather than using a list of altered analytes as inputs, FSEA takes the entire panel of analytes as input, usually as a ranked list of fold changes, and scores each pathway by an empirically determined weighted Kolmogorov–Smirnov-like statistic. FSEA, however, is more computationally intensive than ORA, and the statistical hypothesis being assessed is less straightforward to interpret. While there are no publicly available implementations of FSEA that simultaneously test multiple omic modalities, the method could theoretically be extended to this application. Examples of FSEA available in single omic modalities include Gene Set Enrichment Analysis (GSEA) [213], Metabolite Set Enrichment Analysis (MSEA) [214], and the Lipid Ontology web-based interface (LION/Web) [215].

Topological scoring techniques that use the structure of networks to infer pathway associations for altered analytes can also be applied. Enriched pathways can be found by mapping differentially expressed analytes onto individual metabolic pathway networks derived from biological pathway databases to determine the global perturbation of the pathway. In these metabolic pathway networks, nodes represent analytes, and edges represent physical or chemical interactions between analytes (e.g., catalyzation, inhibition) as part of a pathway. Each pathway is its own subnetwork. Global perturbation is measured using a combination of standard pathway enrichment and the topology of the network, such as the length of the path between altered analytes and other analytes in the network, betweenness centrality of analytes, and the degree of an analyte. Methods that fall into this category include Signaling Pathway Impact Analysis [216], Pathway Regulation Score [217], Centrality-Based Pathway Enrichment [218], Topological Analysis of Pathway Phenotype Association [219], Topology Gene Set Analysis [220], Clipper [221], and DEGraph [222], which are often used in pathway analysis of gene sets but can also be applied to other analytes. Ihnatova et al. found that the results of these methods differ in sensitivity and specificity when simulated data vary by topological motif size and size of the overexpressed gene set [223]. We note that MetaboAnalyst also incorporates a form of topological analysis called the Pathway Impact Score, which is based on betweenness centrality of differentially expressed analytes in a pathway [2].

Another type of topological scoring method represents pathways as nodes and relationships between pathways as edges in a network. In these networks, differentially expressed analytes are mapped onto pathway nodes. Then, the topology of the network is evaluated to find sets of related enriched pathways. An example of this type of analysis was performed in Zachariou et al. [198]. Lastly, topological scoring can be applied to networks where both analytes and pathways are represented as nodes, and analyte membership in a pathway is represented by edges. Then, analytes of interest are mapped onto this network, and related pathways are found using known membership. This approach is used by FELLA [224], where the authors demonstrate that biologically relevant pathways not found using other pathway analysis approaches can be highlighted using this approach in epithelial cells, ovarian cancer cells, and blood samples of malaria patients [224].

6.2. Visualization of Biological Pathways and Networks

Many pathway visualization tools are embedded into biological pathway databases and are designed to visualize one specific pathway at a time. Users can map their analytes of interest, along with analyte abundances or other characteristics, onto these pathways for further investigation. Examples of these types of visualization include OmicsViewer [225], Visualization and Analysis of Networks Containing Experimental Data (VANTED) [226], and PaintOmics [227]. PathMe [228] provides additional flexibility as it incorporates multiple sources related to biological pathways and evaluates crosstalks between these sources. Other tools provide additional flexibility in that they provide a framework for visualizing pathways and/or networks. For example, Cytoscape [229], GraphViz [230], and igraph [231] are very flexible and allow users to upload custom analytes or pathways along with their relationships. PathVisio provides a user-friendly way to draw pathways and to visualize experimental data on these pathways [232].

Other visualization tools represent analyte–analyte interactions outside of the pathway context. For example, OmicsNet [233] combines protein–protein interactions, miRNA–target interactions, transcription factor–target interactions, and enzyme–metabolite interactions from multiple annotation databases to generate a composite network given a list of analytes.

We note that standard formats exist for networks in the multi-omics space. One of these formats is Systems Biology Graphical Notation (SBGN), which includes three languages used for network representation: Activity Flow, Process Description, and Entity Relationship. Each SBGN language includes standardized glyphs and types of information that can be represented in textual annotations [234]. VANTED follows SBGN specifications. Another format is GenMAPP Pathway Markup Language (GPML), which is an Extensible Markup Language (XML)-based format with graphical elements used for storing pathways. GPML is used by some knowledge bases containing graphical information and by PathVisio. Finally, WikiPathways uses the World Wide Web Consortium's Resource Description Framework (RDF), which facilitates the integration of structured and semi-structured data by creating links between resources [235].

6.3. Sources of A Priori Knowledge

For analyses involving pathway enrichment or related analyses, it is important to consider which biological, biochemical, and disease pathway databases should be used. In fact, the coverage of analytes and analyte types differ greatly between databases [236], and the choice of the database used for analysis, such as pathway enrichment analyses, affects results [237]. For this reason, databases that integrate information from multiple sources and for multiple types of analytes have thus been developed. These are particularly useful for performing pathway enrichment using multiple types of analytes as input, as they maximize coverage of pathway annotations [210,211,237–240]. In addition, the incorporation of biological context (e.g., biospecimen type, species) into pathway analyses has been shown to yield increased specificity of results when compared to functional analysis without incorporated context [241].

6.3.1. Curated and Community Resources

Various efforts are underway to collect, organize, and disseminate information about metabolites and their association with other analytes or types of information (e.g., diseases, biospecimen location, chemical information). Most open-source resources are maintained by curators who manually input and/or review information about analytes and their annotations from the literature. The database curation process relies on domain experts to ensure the accuracy of the information contained in the database. Prominent examples of databases integrating metabolite pathway annotations with other analyte annotations (genes, proteins, microbes, etc.) include the Human Metabolome Database (HMDB) [242], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [243], the BioCyc database series including MetaCyc [244], and Metabolic Pathways Database for Microbial Taxonomic Groups (MACADAM) [245]. Other resources aim to be comprehensive and incorporate information from many database sources, including MetaCyc [244], Pathway Commons [240], PathBank [246], Relational database of Metabolomics Pathways (RaMP) [210], and Pharmacogenomics Knowledgebase (PharmGKB) [247]. We note that the most widely used resources are actively maintained. Therefore, new and updated versions of the database are deployed at varying frequencies, from every several months to yearly. Over time, the accuracy of the databases increases, as new and corroborating knowledge is incorporated.

The integration of multiple sources is challenging, particularly for metabolites, since it is highly dependent on the coverage of information, the confidence in analyte identity, and the accuracy of mapping analyte IDs across databases. Coverage of analytes and other knowledge, such as pathway membership, is important for analyte identification and the retrieval of metadata. However, recent analysis of metabolic networks revealed that mass spectral libraries only covered 40% of these networks [236]. In biological pathway databases, coverage of genes and metabolites also varies, where 12% of metabolites and 67% of coding genes are mappable to pathways [210]. Additionally, the confidence of identity and the level of resolution (e.g., location of double bonds, strain vs. species) available may also affect mapping to pathways [26,27]. To improve the confidence of annotations, users should use IDs, rather than names, to retrieve information on analytes, and should evaluate mapping results for accuracy. In addition, mapping IDs across databases is also a challenge. Different databases use different IDs, with varying levels of information. For example, metabolites represented by International Chemical Identifier Keys (InChIKeys) [248] uniquely map to one metabolite while the commonly used Chemical Entities of Biological Interest (ChEBI) [249] IDs do not [250]. While most databases include links to commonly used ID types, errors could be introduced when mapping IDs from one database to another because of this discrepancy [251]. Standardization of IDs is thus a major challenge that is not completely solved, although it is being addressed by large community-driven initiatives [26,250,252]. Until nomenclatures are fully converged, the community relies on metabolite naming translation services [253]. This issue could be further mitigated by the use of text mining algorithms [254,255].

6.3.2. Computationally Predicted Resources

Natural Language Processing (NLP) methods can be applied to automatically extract knowledge that can be incorporated in multi-omics data analyses and resources. NLP methods mine information from the literature, where tokens (individual or compound words in a document) are analyzed individually and within sentence structures to extract relevant information. One point of consideration in NLP is the dictionary of terms, i.e., the set of possible tokens. The dictionary of terms may be created manually or from existing knowledgebases or literature, if they are available. For instance, the NLP R package Onassis for omics data uses Open Biomedical Ontologies to build its dictionary [256]. In contrast, the Indian Medicinal Plants, Phytochemistry and Therapeutics dictionary was built manually, as there was no electronic resource containing the names of all Indian medicinal plants and their synonyms [257]. In addition to the dictionary of terms, the sources mined must be appropriate. Sources may include abstracts or full-text articles from multiple journals or journal repositories, in addition to other sources, including curated databases, online encyclopedias, patents [258], drug reviews [259], and lab protocols [260]. When considering journals, choosing to use abstracts only may reduce the number of true relationships found [261], but it is often done because of limited accessibility to full text articles. In addition, both the trustworthiness of the source and the context of the article (e.g., toxicology or immunology) should be relevant to the type of information the researcher wishes to extract. Finally, regular expressions must be correctly formatted to define extraction rules for text. Regular expressions are complex patterns of text that can be matched in a document, and they are used to define rules for extracting text to build the knowledgebase. These should be specialized for the task at hand. For instance, Ben Abdessalem Karaa et al. created separate regular expressions to extract causal, preventative, and associative relationships between types of food, genes, and diseases [262], Nikfarjam et al. used a list of key phrases to describe patient responses to drugs in social health networks [263], and Fan and Zhang created several regular expressions to extract patient dietary supplement use from clinical notes [264].

We note that using NLP in the context of biomedical research offers unique challenges that must be considered, which are reviewed in [265]. One of these is resolving words that co-reference the same analyte; Cohen et al. have worked toward solutions specific to biomedical journals [266]. Other work focuses on the task of associating genes with diseases [267] or finding associations between metabolites, proteins, genes, and diseases [258]. Another challenge is document triage, such as finding documents relevant to a context or field of study [268–270]. For a thorough review of NLP techniques, as applied to the biomedical literature, we invite the reader to reference [271].

Several NLP-based resources relevant to multi-omics data have been developed. The NJS16 database is a literature-derived database which contains information on the import, export, and macromolecular degradation of metabolites by 570 microbial (bacterial and archaeal) species in three colonic and liver cells [272]. MACADAM, which also contains functional links between microbial species and metabolites, incorporates information from the International Journal of Systematic and Evolutionary Microbiology and Functional Annotation of Prokaryotic Taxa databases, which are derived from the literature [245]. The Drug-Gene Interaction Database (DGIdb) is an NLP-curated database that stores information about mutated genes that could be useful to identify targets for drug development [273]. Another NLP-based resource, which is specifically focused on liver tissue, is LiverWiki [274], a wiki-based knowledgebase containing liver-related genes, metabolites, proteins, protein interactions, pathways, post-translational modifications, and diseases.

Other computational applications aim to predict analyte ontologies, synonym resolution, prediction of molecule interactions/effects, and pathway prediction. For example, ClassyFire [275] provides a taxonomy where compounds are automatically classified into appropriate taxa using a rule-based classification, based on the Simplified Molecular-Input Line-Entry System Arbitrary Target Specification (SMARTS) string and the Markush format. To resolve the many synonyms that can describe one metabolite and mitigate the duplication of analytes used for downstream statistical analyses, PubChem [276] uses an automated standardization technique. This technique works by computing the similarity between two compounds using multiple chemical properties (e.g., atom valence, functional group, and stereochemistry) and merging metabolites with significantly similar properties.

The prediction of a molecule's interactions and/or effects (e.g., toxicity) can also be automated by using chemical or molecular similarity between the compound in question and another, more well-characterized compound, as is done by Super Natural II [277]. Lastly, to fill the knowledge gap of unknown biological and chemical pathways in all organisms, pathways for unexplored organisms can be predicted using PathoLogic [278]. Specifically, pathways are predicted by first inferring the reactions present based on the identification of enzymes in the organism and then by associating key reactions with pathways.

In some cases, computational prediction and text mining have been used to enrich experiments without inclusion in knowledgebases. In one study focused on finding associations across omic modalities, Fadason et al. found interactions between metabolite-associated single nucleotide polymorphisms (SNPs), metabolites, and chromatin loops (i.e., physical contact between enhancers and promoters) in human blood by combining literature text mining, known drug interactions, Hi-C chromatin interactions, eQTL, and gene ontologies [279]. Additionally, computational predictions can be used to infer associations between analytes. For example, a study by Le et al., used an encoder–decoder neural network to learn novel functional associations between the metabolome and the microbiome in a cohort of paired Inflammatory Bowel Disease (IBD) patients. Using the weights learned using the encoder-decoder neural network to indicate the strength of the relationship between analytes, they uncovered relationships between known IBD biomarkers, such as between *Ruminococcus* and ropane alkaloids and steroidal saponins, and between *Fusobacterium* and bile acids, alcohols, and derivatives [280]. In addition, Morton et al. developed microbe–metabolite vectors (mmvec), a variation of the NLP method word2vec [281], to embed co-occurrence patterns between microbes and metabolites and then infer interactions using the embeddings [282].

We note that as with any computational prediction approaches, there is a margin of error. In this case, this error is difficult to determine because "we don't know what we don't know". Therefore, it is advisable to understand whether resources used are computationally driven, rather than curated through existing and validated experiments. Community-driven resources also bring into question confidence in the user's entry, as the user's expertise is often unknown. We note that many resources lack confidence metrics. Lastly, the context of the experiment (e.g., biospecimen location, disease type, etc.) can be utilized to help prioritize analysis results. One example in pathway enrichment analysis is a tool that uses the literature evidence to prioritize the enriched pathways that are returned [204]. Specifically, the algorithm prioritizes pathways that are supported by multiple articles that are related to the same experimental context in the study. When little or no relevant information from the literature exists, then the statistical significance returned from the pathway enrichment analysis method is given more weight in comparison to the literature evidence.

6.3.3. Metrics Used to Define Confidence in Annotations

Confidence in the correctness of an annotation in a knowledgebase can depend on whether there are unknown enzymes or reactions in a pathway or whether a curated annotation has been verified by multiple experts. MACADAM [245] seeks to address the problem of unknown enzymes and reactions using its Pathway Score, based on the percentage of reactions in a pathway that are annotated with an enzyme, and the Pathway Frequency Score, the ratio of annotated enzymes to total reactions. Several other tools include metadata describing the verification of annotations contained therein. In HumanCyc [283], PathoLogic is used to predict metabolic pathways, and pathways are associated with tiers indicating whether they have also undergone manual review. ChEBI has a similar system based on starring: one star means the metabolite entry was automatically curated from a data source, two stars means the metabolite entry was manually processed by a depositor, and three stars means the metabolite entry was manually curated by the ChEBI organization [284]. Finally, WikiPathways [238] has quality tags that can be used to indicate confidence (e.g., *ProposedDeletion*, *WormBase_Approved*, *Reactome_Approved*, and *Hypothetical)*.

Of note, many databases do not include confidence metrics. In this case, the user could map analytes or annotations back to databases which do contain confidence metadata. This can be done using identifiers for analytes or pathways. Alternatively, users can examine the supporting literature for analytes of interest, which can be a tedious process, particularly when many analytes are being considered. Finally, users could assume the same level of confidence for all analytes and annotations in the databases.

7. Discussion

Given the complexity and heterogeneity of multi-omics data and experiments, data analysis and interpretation are collaborative efforts, involving biostatisticians, bioinformaticians, molecular biologists, and domain experts (e.g., clinicians, immunologists). Further, given the large number of available methods, tools, and workflows, it is sometimes difficult to select which approach to consider. We recognize that no single method or approach is comprehensive, but rather, approaches and methods are complementary. Applying multiple methods to the same datasets is thus advisable and may corroborate findings or identify novel analyte patterns or relationships.

Aspects that should be considered when selecting methods are: 1) the requirements of input data, including data distributions and types; 2) the biological question being addressed, noting that each tool typically aims to answer a specific question; 3) the availability and metrics of confidence of external resources; 4) ease of use (some methods are hard to implement without computational or statistical expertise); and 5) reproducibility of the results (some methods are stochastic and yield different results when run on the same dataset). Once a method is selected, it is also important to consider which parameters can be modified. To help with parameter selection, most tools provide default parameters to guide the users, although we note that these parameters may not be optimal for all cases. A balance must then be struck between the ease of use of approaches, and the selection of appropriate parameters given the input datasets.

Currently, computational solutions are still lagging behind the rapid influx of molecular data being generated [285]. As computational methods, tools, and workflows emerge, it is important to compare their utility on benchmarking datasets. At present, well curated, publicly available benchmarking datasets are uncommon. Further, emerging computational approaches may not test their performance or complementarity with other tools on the same datasets, making the comparison of multi-omics approaches challenging. Efforts to create readily available data, including different formats of the same data (e.g., before and after data preprocessing) for direct input into new developments would facilitate comparison and an understanding of which approaches to use for which contexts. We also note that developers of methods should disclose the data and code used, along with their publications, to mitigate current deficiencies in reproducibility [286].

One area that we do not delve into deeply in this review is uncertainty in the identification of analytes, which is prevalent for metabolites, but is also relevant for genes and microbes, which rely on accurate sequence reads and alignments. It is feasible for algorithms to take identification uncertainty into account. In computer science, systems with uncertainty in the input, output, or parameters are often called "fuzzy systems"; machine learning methods for fuzzy systems are reviewed in [287]. The representation of uncertainty in visualization tools has also been explored and is reviewed in [288].

The granularity of information that is being received for metabolomics analysis is increasing, as annotations for analytes, experimentally validated or in silico, is ever-growing. This increasing granularity in turn enables the development of more context-specific analyses. For example, multi-omics data-analysis of colon samples can be restricted to analytes that are known to be colonic, thereby removing potential artifacts and false positives in the data. Similarly, metabolic network models can be built per organism to better capture the underlying biology.

8. Conclusions

The development of computational approaches that support multi-omics data analysis is still an active area of research. Given the large number of available approaches to such data analyses, the identification of appropriate method(s) for a researcher's needs is challenging. In this review, we describe concepts and considerations to be made when performing multi-omics analyses, particularly from the viewpoint of which methods, tools, workflows, and resources are available. We discuss the statistical and computational approaches to the tasks of unsupervised clustering, identification of co-regulated groups of analytes, the identification of associations between analytes or groups of analytes and phenotype, and biological interpretation of the relationships between analytes and phenotype, highlighting methods with examples from real-world multi-omics applications in various domains. In addition, we describe a priori sources of knowledge that can be used in biological interpretation analysis as well as points of consideration regarding these sources. Globally, we anticipate a growth in multi-omics data analysis approaches to meet the demands of biomedical research. Such analyses present a unique opportunity for collaborative work amongst different fields, providing multiple viewpoints and knowledge on the same biological system.

Author Contributions: Conceptualization, E.A.M., R.M., and T.E.; writing—original draft preparation, all authors; writing—review and editing, all authors; funding acquisition, E.A.M.; investigation, all authors; project supervision, T.E. and E.A.M.; figures—E.A.M., Q.M., and T.E. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: This work was supported by funding from The Ohio State University Translational Data Analytics Institute and startup funds from The Ohio State University to Ewy A. Mathé and was supported in part by the Intramural/Extramural research program of the National Center for Advancing Translational Sciences, National Institutes of Health.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pinu, F.R.; Beale, D.J.; Paten, A.M.; Kouremenos, K.; Swarup, S.; Schirra, H.J.; Wishart, D. Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. *Metabolites* **2019**, *9*, 76. [CrossRef] [PubMed]
2. Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D.S.; Xia, J. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **2018**, *46*, W486–W494. [CrossRef] [PubMed]
3. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787. [CrossRef] [PubMed]
4. Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef] [PubMed]
5. Ulfenborg, B. Vertical and horizontal integration of multi-omics data with miodin. *BMC Bioinform.* **2019**, *20*, 649. [CrossRef]
6. Kumar, N.; Hoque, M.A.; Sugimoto, M. Robust volcano plot: Identification of differential metabolites in the presence of outliers. *BMC Bioinform.* **2018**, *19*, 128. [CrossRef]
7. Greco, L.; Luta, G.; Krzywinski, M.; Altman, N. Analyzing outliers: Robust methods to the rescue. *Nat. Methods* **2019**, *16*, 275–276. [CrossRef]
8. Taylor, S.L.; Ruhaak, L.R.; Kelly, K.; Weiss, R.H.; Kim, K. Effects of imputation on correlation: Implications for analysis of mass spectrometry data from multiple biological matrices. *Brief. Bioinform.* **2017**, *18*, 312–320. [CrossRef]
9. Hughes, R.A.; Heron, J.; Sterne, J.A.C.; Tilling, K. Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *Int. J. Epidemiol.* **2019**, *48*, 1294–1304. [CrossRef]
10. Lin, D.; Zhang, J.; Li, J.; Xu, C.; Deng, H.-W.; Wang, Y.-P. An integrative imputation method based on multi-omics datasets. *BMC Bioinform.* **2016**, *17*, 247. [CrossRef]
11. Zhu, H.; Li, G.; Lock, E.F. Generalized integrative principal component analysis for multi-type data with block-wise missing structure. *Biostatistics* **2020**, *21*, 302–318. [CrossRef] [PubMed]
12. Chu, S.H.; Huang, M.; Kelly, R.S.; Benedetti, E.; Siddiqui, J.K.; Zeleznik, O.A.; Pereira, A.; Herrington, D.; Wheelock, C.E.; Krumsiek, J.; et al. Integration of Metabolomic and Other Omics Data in Population-Based Study Designs: An Epidemiological Perspective. *Metabolites* **2019**, *9*, 117. [CrossRef] [PubMed]
13. Tarazona, S.; Balzano-Nogueira, L.; Conesa, A. Multiomics Data Integration in Time Series Experiments. In *Comprehensive Analytical Chemistry*; Elsevier B.V.: Amsterdam, The Netherlands, 2018; Volume 82, pp. 505–532. ISBN 9780444640444.
14. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.A.; Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **2015**, *16*, 85–97. [CrossRef] [PubMed]
15. 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]
16. Cavill, R.; Jennen, D.; Kleinjans, J.; Briedé, J.J. Transcriptomic and metabolomic data integration. *Brief. Bioinform.* **2016**, *17*, 891–901. [CrossRef] [PubMed]

17. Stanstrup, J.; Broeckling, C.D.; Helmus, R.; Hoffmann, N.; Mathé, E.; Naake, T.; Nicolotti, L.; Peters, K.; Rainer, J.; Salek, R.M.; et al. The metaRbolomics Toolbox in Bioconductor and beyond. *Metabolites* **2019**, *9*, 200. [CrossRef]
18. Liu, Z.; Ma, A.; Mathé, E.; Merling, M.; Ma, Q.; Liu, B. Network analyses in microbiome based on high-throughput multi-omics data. *Brief. Bioinform.* **2020**. [CrossRef]
19. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 1–9. [CrossRef]
20. Lamprecht, A.-L.; Garcia, L.; Kuzak, M.; Martinez, C.; Arcila, R.; Martin Del Pico, E.; Dominguez Del Angel, V.; van de Sandt, S.; Ison, J.; Martinez, P.A.; et al. Towards FAIR principles for research software. *Data Sci.* **2019**, 1–23. [CrossRef]
21. Silva, L.B.; Jimenez, R.C.; Blomberg, N.; Luis Oliveira, J. General guidelines for biomedical software development. *F1000Research* **2017**, *6*, 273. [CrossRef]
22. Jiménez, R.C.; Kuzak, M.; Alhamdoosh, M.; Barker, M.; Batut, B.; Borg, M.; Capella-Gutierrez, S.; Chue Hong, N.; Cook, M.; Corpas, M.; et al. Four simple recommendations to encourage best practices in research software. *F1000Research* **2017**, *6*, 876. [CrossRef] [PubMed]
23. Russell, P.H.; Johnson, R.L.; Ananthan, S.; Harnke, B.; Carlson, N.E. A large-scale analysis of bioinformatics code on GitHub. *PLoS ONE* **2018**, *13*, e0205898. [CrossRef] [PubMed]
24. Begley, C.G.; Ellis, L.M. Drug development: Raise standards for preclinical cancer research. *Nature* **2012**, *483*, 531–533. [CrossRef] [PubMed]
25. Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C.A.; Causton, H.C.; et al. Minimum information about a microarray experiment (MIAME) - Toward standards for microarray data. *Nat. Genet.* **2001**, *29*, 365–371. [CrossRef]
26. Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; Fan, T.W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J.L.; et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **2007**, *3*, 211–221. [CrossRef]
27. Castle, A.L.; Fiehn, O.; Kaddurah-Daouk, R.; Lindon, J.C. Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. *Brief. Bioinform.* **2006**, *7*, 159–165. [CrossRef]
28. Taylor, C.F.; Paton, N.W.; Lilley, K.S.; Binz, P.A.; Julian, R.K.; Jones, A.R.; Zhu, W.; Apweiler, R.; Aebersold, R.; Deutsch, E.W.; et al. The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **2007**, *25*, 887–893. [CrossRef]
29. Ochsner, S.A.; Steffen, D.L.; Stoeckert, C.J.; McKenna, N.J. Much room for improvement in deposition rates of expression microarray datasets. *Nat. Methods* **2008**, *5*, 991. [CrossRef]
30. Spicer, R.A.; Salek, R.; Steinbeck, C. Comment: A decade after the metabolomics standards initiative it's time for a revision. *Sci. Data* **2017**, *4*, 170138. [CrossRef]
31. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkol.* **2015**, *19*, A68–A77. [CrossRef]
32. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehár, J.; Kryukov, G.V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603–607. [CrossRef] [PubMed]
33. National Cancer Institute Office of Cancer Genomics TARGET: Therapeutically Applicable Research to Generate Effective Treatments. Available online: https://ocg.cancer.gov/programs/target (accessed on 1 May 2020).
34. Edwards, N.J.; Oberti, M.; Thangudu, R.R.; Cai, S.; McGarvey, P.B.; Jacob, S.; Madhavan, S.; Ketchum, K.A. The CPTAC data portal: A resource for cancer proteomics research. *J. Proteome Res.* **2015**, *14*, 2707–2713. [CrossRef] [PubMed]
35. Shoemaker, R.H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **2006**, *6*, 813–823. [CrossRef] [PubMed]
36. Haug, K.; Cochrane, K.; Nainala, V.; Williams, M.; Chang, J.; Jayaseelan, K.; O'Donovan, C. MetaboLights: A resource evolving in response to the needs of its scientific community. - PubMed - NCBI. *Nucleic Acids Res.* **2020**, *48*, D440–D444.

37. Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K.S.; et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **2016**, *44*, D463–D470. [CrossRef]
38. Vizcaíno, J.A.; Côté, R.; Reisinger, F.; Foster, J.M.; Mueller, M.; Rameseder, J.; Hermjakob, H.; Martens, L. A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* **2009**, *9*, 4276–4283. [CrossRef]
39. Clough, E.; Barrett, T. The Gene Expression Omnibus database. *Methods Mol. Biol.* **2016**, *1418*, 93–110.
40. Leinonen, R.; Sugawara, H.; Shumway, M. The Sequence Read Archive. *Nucleic Acids Res.* **2011**, *39*, D19–D21. [CrossRef]
41. Feingold, E.A.; Good, P.J.; Guyer, M.S.; Kamholz, S.; Liefer, L.; Wetterstrand, K.; Collins, F.S.; Gingeras, T.R.; Kampa, D.; Sekinger, E.A.; et al. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* **2004**, *306*, 636–640. [CrossRef]
42. Methé, B.A.; Nelson, K.E.; Pop, M.; Creasy, H.H.; Giglio, M.G.; Huttenhower, C.; Gevers, D.; Petrosino, J.F.; Abubucker, S.; Badger, J.H.; et al. A framework for human microbiome research. *Nature* **2012**, *486*, 215–221.
43. Oliveira, F.S.; Brestelli, J.; Cade, S.; Zheng, J.; Iodice, J.; Fischer, S.; Aurrecoechea, C.; Kissinger, J.C.; Brunk, B.P.; Stoeckert, C.J.; et al. MicrobiomeDB: A systems biology platform for integrating, mining and analyzing microbiome experiments. *Nucleic Acids Res.* **2018**, *46*. [CrossRef] [PubMed]
44. Chen, T.; Yu, W.-H.; Izard, J.; Baranova, O.V.; Lakshmanan, A.; Dewhirst, F.E. The Human Oral Microbiome Database: A web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)* **2010**, *2010*, baq013. [CrossRef] [PubMed]
45. Sarkans, U.; Gostev, M.; Athar, A.; Behrangi, E.; Melnichuk, O.; Ali, A.; Minguet, J.; Rada, J.; Snow, C.; Tikhonov, A.; et al. The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* **2018**, *46*, D1266–D1270. [CrossRef] [PubMed]
46. Sreng, N.; Champion, S.; Martin, J.C.; Khelaifia, S.; Christensen, J.E.; Padmanabhan, R.; Azalbert, V.; Blasco-Baque, V.; Loubieres, P.; Pechere, L.; et al. Resveratrol-mediated glycemic regulation is blunted by curcumin and is associated to modulation of gut microbiota. *J. Nutr. Biochem.* **2019**, *72*, 108218. [CrossRef] [PubMed]
47. Tkachev, A.; Stepanova, V.; Zhang, L.; Khrameeva, E.; Zubkov, D.; Giavalisco, P.; Khaitovich, P. Differences in lipidome and metabolome organization of prefrontal cortex among human populations. *Sci. Rep.* **2019**, *9*, 18348. [CrossRef]
48. Chaisaingmongkol, J.; Budhu, A.; Dang, H.; Rabibhadana, S.; Pupacdi, B.; Kwon, S.M.; Forgues, M.; Pomyen, Y.; Bhudhisawasdi, V.; Lertprasertsuke, N.; et al. Common Molecular Subtypes Among Asian Hepatocellular Carcinoma and Cholangiocarcinoma. *Cancer Cell* **2017**, *32*, 57–70. [CrossRef]
49. Terunuma, A.; Putluri, N.; Mishra, P.; Mathé, E.A.; Dorsey, T.H.; Yi, M.; Wallace, T.A.; Issaq, H.J.; Zhou, M.; Killian, J.K.; et al. MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. *J. Clin. Invest.* **2014**, *124*, 398–412. [CrossRef]
50. Overmyer, K.A.; Rhoads, T.W.; Merrill, A.E.; Ye, Z.; Westphall, M.S.; Acharya, A.; Shukla, S.K.; Coon, J.J. Proteomics, lipidomics, metabolomics and 16S DNA sequencing of dental plaque from patients with diabetes and periodontal disease. *bioRxiv* **2020**. [CrossRef]
51. Battaglioli, E.J.; Hale, V.L.; Chen, J.; Jeraldo, P.; Ruiz-Mojica, C.; Schmidt, B.A.; Rekdal, V.M.; Till, L.M.; Huq, L.; Smits, S.A.; et al. Clostridioides difficile uses amino acids associated with gut microbial dysbiosis in a subset of patients with diarrhea. *Sci. Transl. Med.* **2018**, *10*, eaam7019. [CrossRef]
52. Athreya, A.; Iyer, R.; Neavin, D.; Wang, L.; Weinshilboum, R.; Kaddurah-Daouk, R.; Rush, J.; Frye, M.; Bobo, W. Augmentation of physician assessments with multi-omics enhances predictability of drug response: A case study of major depressive disorder. *IEEE Comput. Intell. Mag.* **2018**, *13*, 20–31. [CrossRef]
53. Schmaler, M.; Colone, A.; Spagnuolo, J.; Zimmermann, M.; Lepore, M.; Kalinichenko, A.; Bhatia, S.; Cottier, F.; Rutishauser, T.; Pavelka, N.; et al. Modulation of bacterial metabolism by the microenvironment controls MAIT cell stimulation. *Mucosal Immunol.* **2018**, *11*, 1060–1070. [CrossRef] [PubMed]
54. Knudsen, E.S.; Balaji, U.; Freinkman, E.; McCue, P.; Witkiewicz, A.K. Unique metabolic features of pancreatic cancer stroma: Relevance to the tumor compartment, prognosis, and invasive potential. *Oncotarget* **2016**, *7*, 78396–78411. [CrossRef] [PubMed]

55. Chung, N.C.; Mirza, B.; Choi, H.; Wang, J.; Wang, D.; Ping, P.; Wang, W. Unsupervised classification of multi-omics data during cardiac remodeling using deep learning. *Methods* **2019**, *166*, 66–73. [CrossRef] [PubMed]
56. 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. Boil.* **2018**, *14*, e8124. [CrossRef] [PubMed]
57. Meng, C.; Kuster, B.; Culhane, A.C.; Gholami, A.M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinform.* **2014**, *15*, 162. [CrossRef] [PubMed]
58. Hout, M.C.; Papesh, M.H.; Goldinger, S.D. Multidimensional scaling. *Wiley Interdiscip. Rev. Cogn. Sci.* **2013**, *4*, 93–103. [CrossRef]
59. Kuczynski, J.; Liu, Z.; Lozupone, C.; McDonald, D.; Fierer, N.; Knight, R. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods* **2010**, *7*, 813–819. [CrossRef]
60. Van Der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. March. Learn. Res.* **2008**, *9*, 2579–2605.
61. Wattenberg, M.; Viégas, F.; Johnson, I. How to Use t-SNE Effectively. *Distill* **2016**, *1*, e2. [CrossRef]
62. Kimes, P.K.; Liu, Y.; Neil Hayes, D.; Marron, J.S. Statistical significance for hierarchical clustering. *Biometrics* **2017**, *73*, 811–821. [CrossRef]
63. Macqueen, J.; Macqueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1965; pp. 281–297.
64. Kaufman, L.; Rousseeuw, P. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 1990.
65. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
66. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc.* **2001**, *63*, 411–423. [CrossRef]
67. Kohonen, T. The self-organizing map. *Neurocomputing* **1998**, *21*, 1–6. [CrossRef]
68. Hamel, L.; Ott, B. A Population Based Convergence Criterion for Self-Organizing Maps. In Proceedings of the 2012 International Conference on Data Mining, Brussels, Belgium, 10–13 December 2012; pp. 98–104.
69. Kiviluoto, K. Topology preservation in self-organizing maps. In Proceedings of the International Conference on Neural Networks (ICNN'96), Washington, DC, USA, 2–7 June 1996; pp. 294–299.
70. Milone, D.H.; Stegmayer, G.S.; Kamenetzky, L.; López, M.; Lee, J.M.; Giovannoni, J.J.; Carrari, F. *omeSOM: A software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. *BMC Bioinform.* **2010**, *11*, 438. [CrossRef]
71. Duda, R.O.; Hart, P.E.; Stork, D.G. Pattern Classification. In *Pattern Classification and Scene Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 1995; pp. 6–22.
72. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
73. Akaike, H. Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium on Information Theory, Akadémiai Kiadó, Budapest, Hungary, 2–8 September 1971; pp. 267–281.
74. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996.
75. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
76. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. Training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory; Publ by ACM, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
77. Winters-Hilt, S.; Merat, S. SVM clustering. In Proceedings of the BMC Bioinformatics, BioMed Central, New Orleans, LA, USA, 1–3 Febuary 2007; p. S18.
78. Ballard, D.H. Modular Learning in Neural Networks. In Proceedings of the Association for the Advancement of Artificial Intelligence Sixth National Conference on Artificial Intelligence, Seattle, WA, USA, 13–17 July 1987.
79. Samek, W.; Wiegand, T.; Müller, K.-R. Explainable Artificial Intelligence: Understanding, Visualizing, and Interpreting Deep Learning Models. *ITU J. ICT Discov.* **2017**.

80. Karim, M.R.; Beyan, O.; Zappa, A.; Costa, I.G.; Rebholz-Schuhmann, D.; Cochez, M.; Decker, S. Deep learning-based clustering approaches for bioinformatics. *Brief. Bioinform.* **2020**. [CrossRef]
81. Bar-Joseph, Z.; Gerber, G.K.; Gifford, D.K.; Jaakkola, T.S.; Simon, I. Continuous Representations of Time-Series Gene Expression Data. *J. Comput. Biol.* **2003**, *10*, 341–356. [CrossRef]
82. Déjean, S.; Martin, P.; Baccini, A.; Besse, P. Clustering Time-Series Gene Expression Data Using Smoothing Spline Derivatives. *EURASIP J. Bioinform. Syst. Biol.* **2007**, *2007*, 70561. [CrossRef]
83. Corduas, M.; Piccolo, D. Time series clustering and classification by the autoregressive metric. *Comput. Stat. Data Anal.* **2008**, *52*, 1860–1872. [CrossRef]
84. Kalpakis, K.; Gada, D.; Puttagunta, V. Distance measures for effective clustering of ARIMA time-series. In Proceedings of the IEEE International Conference on Data Mining, San Jose, CA, USA, 29 November–2 December 2001; pp. 273–280.
85. Smyth, P. Clustering Sequences with Hidden Markov Models. *Adv. Neural Inf. Process. Syst.* **1997**, *9*, 648–654.
86. Zeng, Y.; Garcia-Frias, J. A novel HMM-based clustering algorithm for the analysis of gene expression time-course data. *Comput. Stat. Data Anal.* **2006**, *50*, 2472–2494. [CrossRef]
87. Jaskowiak, P.A.; Campello, R.J.G.B.; Costa, I.G. On the selection of appropriate distances for gene expression data clustering. *BMC Bioinform.* **2014**, *15*. [CrossRef]
88. Giorgino, T. Computing and visualizing dynamic time warping alignments in R: The dtw package. *J. Stat. Softw.* **2009**, *31*, 1–24. [CrossRef]
89. Chandereng, T.; Gitter, A. Lag penalized weighted correlation for time series clustering. *BMC Bioinform.* **2020**, *21*, 21. [CrossRef]
90. Camacho, D.; de la Fuente, A.; Mendes, P. The origin of correlations in metabolomics data. *Metabolomics* **2005**, *1*, 53–63. [CrossRef]
91. Do, K.T.; Kastenmüller, G.; Mook-Kanamori, D.O.; Yousri, N.A.; Theis, F.J.; Suhre, K.; Krumsiek, J. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. *J. Proteome Res.* **2015**, *14*, 1183–1194. [CrossRef]
92. Wahl, S.; Vogt, S.; Stückler, F.; Krumsiek, J.; Bartel, J.; Kacprowski, T.; Schramm, K.; Carstensen, M.; Rathmann, W.; Roden, M.; et al. Multi-omic signature of body weight change: Results from a population-based cohort study. *BMC Med.* **2015**, *13*, 48. [CrossRef]
93. Lloyd-Price, J.; Arze, C.; Ananthakrishnan, A.N.; Schirmer, M.; Avila-Pacheco, J.; Poon, T.W.; Andrews, E.; Ajami, N.J.; Bonham, K.S.; Brislawn, C.J.; et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **2019**, *569*, 655–662. [CrossRef]
94. Li, S.; Sullivan, N.L.; Rouphael, N.; Yu, T.; Banton, S.; Maddur, M.S.; McCausland, M.; Chiu, C.; Canniff, J.; Dubey, S.; et al. Metabolic Phenotypes of Response to Vaccination in Humans. *Cell* **2017**, *169*, 862–877. [CrossRef] [PubMed]
95. Aho, V.; Ollila, H.M.; Kronholm, E.; Bondia-Pons, I.; Soininen, P.; Kangas, A.J.; Hilvo, M.; Seppälä, I.; Kettunen, J.; Oikonen, M.; et al. Prolonged sleep restriction induces changes in pathways involved in cholesterol metabolism and inflammatory responses. *Sci. Rep.* **2016**, *6*, 24828. [CrossRef] [PubMed]
96. Acharjee, A.; Ament, Z.; West, J.A.; Stanley, E.; Griffin, J.L. Integration of metabolomics, lipidomics and clinical data using a machine learning method. *BMC Bioinform.* **2016**, *17*, 37–49. [CrossRef] [PubMed]
97. Schubert, K.O.; Stacey, D.; Arentz, G.; Clark, S.R.; Air, T.; Hoffmann, P.; Baune, B.T. Targeted proteomic analysis of cognitive dysfunction in remitted major depressive disorder: Opportunities of multi-omics approaches towards predictive, preventive, and personalized psychiatry. *J. Proteomics* **2018**, *188*, 63–70. [CrossRef] [PubMed]
98. Kelly, R.S.; Chawes, B.L.; Blighe, K.; Virkud, Y.V.; Croteau-Chonka, D.C.; McGeachie, M.J.; Clish, C.B.; Bullock, K.; Celedón, J.C.; Weiss, S.T.; et al. An Integrative Transcriptomic and Metabolomic Study of Lung Function in Children With Asthma. *Chest* **2018**, *154*, 335–348. [CrossRef]
99. Heiland, D.H.; Wörner, J.; Haaker, J.G.; Delev, D.; Pompe, N.; Mercas, B.; Franco, P.; Gäbelein, A.; Heynckes, S.; Pfeifer, D.; et al. The integrative metabolomic-transcriptomic landscape of glioblastome multiforme. *Oncotarget* **2017**, *8*, 49178–49190. [CrossRef]
100. Feng, J.; Zhang, Q.; Zhou, Y.; Yu, S.; Hong, L.; Zhao, S.; Yang, J.; Wan, H.; Xu, G.; Zhang, Y.; et al. Integration of Proteomics and Metabolomics Revealed Metabolite–Protein Networks in ACTH-Secreting Pituitary Adenoma. *Front. Endocrinol. (Lausanne)* **2018**, *9*, 678. [CrossRef]

101. Price, N.D.; Magis, A.T.; Earls, J.C.; Glusman, G.; Levy, R.; Lausted, C.; McDonald, D.T.; Kusebauch, U.; Moss, C.L.; Zhou, Y.; et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat. Biotechnol.* **2017**, *35*, 747–756. [CrossRef]
102. Butte, A.J.; Kohane, I.S. Relevance Networks: A First Step Toward Finding Genetic Regulatory Networks Within Microarray Data. In *The Analysis of Gene Expression Data*; Springer: New York, NY, USA, 2003; pp. 428–446.
103. Kayano, M.; Imoto, S.; Yamaguchi, R.; Miyano, S. Multi-omics approach for estimating metabolic networks using low-order partial correlations. *J. Comput. Biol.* **2013**, *20*, 571–582. [CrossRef]
104. Li, Z.; Zuo, Y.; Xu, C.; Varghese, R.S.; Ressom, H.W. INDEED: R package for network based differential expression analysis. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine, Madrid, Spain, 3–6 December 2018; pp. 2709–2712.
105. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [CrossRef]
106. Longabaugh, W.J.R. Combing the hairball with BioFabric: A new approach for visualization of large networks. *BMC Bioinform.* **2012**, *13*, 275. [CrossRef] [PubMed]
107. Jiang, P.; Wang, H.; Li, W.; Zang, C.; Li, B.; Wong, Y.J.; Meyer, C.; Liu, J.S.; Aster, J.C.; Liu, X.S. Network analysis of gene essentiality in functional genomics experiments. *Genome Biol.* **2015**, *16*, 239. [CrossRef] [PubMed]
108. Azevedo, H.; Moreira-Filho, C.A. Topological robustness analysis of protein interaction networks reveals key targets for overcoming chemotherapy resistance in glioma. *Sci. Rep.* **2015**, *5*, 16830. [CrossRef] [PubMed]
109. Jalili, M. Functional Brain Networks: Does the Choice of Dependency Estimator and Binarization Method Matter? *Sci. Rep.* **2016**, *6*, 29780. [CrossRef] [PubMed]
110. Waller, T.C.; Berg, J.A.; Lex, A.; Chapman, B.E.; Rutter, J. Compartment and hub definitions tune metabolic networks for metabolomic interpretations. *Gigascience* **2020**, *9*. [CrossRef] [PubMed]
111. Wagner, A.; Fell, D.A. The small world inside large metabolic networks. *Proc. R. Soc. B Biol. Sci.* **2001**, *268*, 1803–1810. [CrossRef] [PubMed]
112. Kitsak, M.; Sharma, A.; Menche, J.; Guney, E.; Ghiassian, S.D.; Loscalzo, J.; Barabási, A.-L. Tissue Specificity of Human Disease Module. *Sci. Rep.* **2016**, *6*, 35241. [CrossRef]
113. Kim, S.; Thapa, I.; Zhang, L.; Ali, H. A novel graph theoretical approach for modeling microbiomes and inferring microbial ecological relationships. *BMC Genomics* **2019**, *20*, 1–13. [CrossRef]
114. Celik, S.; Logsdon, B.; Lee, S. Efficient Dimensionality Reduction for High-Dimensional Network Estimation. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1953–1961.
115. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*. [CrossRef]
116. Gaynor, S.M.; Lin, X.; Quackenbush, J. Spectral clustering in regression-based biological networks. *bioRxiv* **2019**, 651950.
117. Lu, Z.; Wahlström, J.; Nehorai, A. Community Detection in Complex Networks via Clique Conductance. *Sci. Rep.* **2018**, *8*, 5982. [CrossRef] [PubMed]
118. Teran Hidalgo, S.J.; Ma, S. Clustering multilayer omics data using MuNCut. *BMC Genomics* **2018**, *19*, 198. [CrossRef] [PubMed]
119. Wang, J.; Li, C.-L.; Tu, B.-J.; Yang, K.; Mo, T.-T.; Zhang, R.-Y.; Cheng, S.-Q.; Chen, C.-Z.; Jiang, X.-J.; Han, T.-L.; et al. Integrated Epigenetics, Transcriptomics, and Metabolomics to Analyze the Mechanisms of Benzo[a]pyrene Neurotoxicity in the Hippocampus. *Toxicol. Sci.* **2018**, *166*, 65–81. [CrossRef] [PubMed]
120. Yoon, H.; Yoon, D.; Yun, M.; Choi, J.S.; Park, V.Y.; Kim, E.K.; Jeong, J.; Koo, J.S.; Yoon, J.H.; Moon, H.J.; et al. Metabolomics of Breast Cancer Using High-Resolution Magic Angle Spinning Magnetic Resonance Spectroscopy: Correlations with 18F-FDG Positron Emission Tomography-Computed Tomography, Dynamic Contrast-Enhanced and Diffusion-Weighted Imaging MRI. *PLoS ONE* **2016**, *11*, e0159949. [CrossRef] [PubMed]
121. Huan, T.; Troyer, D.A.; Li, L. Metabolite Analysis and Histology on the Exact Same Tissue: Comprehensive Metabolomic Profiling and Metabolic Classification of Prostate Cancer. *Sci. Rep.* **2016**, *6*, 1–13. [CrossRef]

122. Clos-Garcia, M.; Andrés-Marin, N.; Fernández-Eulate, G.; Abecia, L.; Lavín, J.L.; van Liempd, S.; Cabrera, D.; Royo, F.; Valero, A.; Errazquin, N.; et al. Gut microbiome and serum metabolome analyses identify molecular biomarkers and altered glutamate metabolism in fibromyalgia. *EBioMedicine* **2019**, *46*, 499–511. [CrossRef]
123. Lee, H.; Choi, J.M.; Cho, J.Y.; Kim, T.E.; Lee, H.J.; Jung, B.H. Regulation of endogenic metabolites by rosuvastatin in hyperlipidemia patients: An integration of metabolomics and lipidomics. *Chem. Phys. Lipids* **2018**, *214*, 69–83. [CrossRef]
124. Esther, C.R.; Turkovic, L.; Rosenow, T.; Muhlebach, M.S.; Boucher, R.C.; Ranganathan, S.; Stick, S.M. Metabolomic biomarkers predictive of early structural lung disease in cystic fibrosis. *Eur. Respir. J.* **2016**, *48*, 1612–1621. [CrossRef]
125. Neeland, I.J.; Boone, S.C.; Mook-Kanamori, D.O.; Ayers, C.; Smit, R.A.J.; Tzoulaki, I.; Karaman, I.; Boulange, C.; Vaidya, D.; Punjabi, N.; et al. Metabolomics Profiling of Visceral Adipose Tissue: Results From MESA and the NEO Study. *J. Am. Heart Assoc.* **2019**, *8*, e010810. [CrossRef]
126. Cambiaghi, A.; Díaz, R.; Martinez, J.B.; Odena, A.; Brunelli, L.; Caironi, P.; Masson, S.; Baselli, G.; Ristagno, G.; Gattinoni, L.; et al. An Innovative Approach for the Integration of Proteomics and Metabolomics Data in Severe Septic Shock Patients Stratified for Mortality. *Sci. Rep.* **2018**, *8*, 6681. [CrossRef]
127. Huang, Y.; Hui, Q.; Walker, D.I.; Uppal, K.; Goldberg, J.; Jones, D.P.; Vaccarino, V.; Sun, Y. V Untargeted metabolomics reveals multiple metabolites influencing smoking-related DNA methylation. *Epigenomics* **2018**, *10*, 379–393. [CrossRef] [PubMed]
128. McGuire, J.L.; DePasquale, E.A.K.; Watanabe, M.; Anwar, F.; Ngwenya, L.B.; Atluri, G.; Romick-Rosendale, L.E.; McCullumsmith, R.E.; Evanson, N.K. Chronic Dysregulation of Cortical and Subcortical Metabolism After Experimental Traumatic Brain Injury. *Mol. Neurobiol.* **2019**, *56*, 2908–2921. [CrossRef] [PubMed]
129. Gao, N.; Ding, L.; Pang, J.; Zheng, Y.; Cao, Y.; Zhan, H.; Shi, Y. Metabonomic-Transcriptome Integration Analysis on Osteoarthritis and Rheumatoid Arthritis. *Int. J. Genomics* **2020**. [CrossRef] [PubMed]
130. Chen, D.; Zhao, X.; Sui, Z.; Niu, H.; Chen, L.; Hu, C.; Xuan, Q.; Hou, X.; Zhang, R.; Zhou, L.; et al. A multi-omics investigation of the molecular characteristics and classification of six metabolic syndrome relevant diseases. *Theranostics* **2020**, *10*, 2029–2046. [CrossRef] [PubMed]
131. Piening, B.D.; Zhou, W.; Contrepois, K.; Röst, H.; Gu Urban, G.J.; Mishra, T.; Hanson, B.M.; Bautista, E.J.; Leopold, S.; Yeh, C.Y.; et al. Integrative Personal Omics Profiles during Periods of Weight Gain and Loss. *Cell Syst.* **2018**, *6*, 157–170. [CrossRef] [PubMed]
132. Acharjee, A.; Kloosterman, B.; Visser, R.G.F.; Maliepaard, C. Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinform.* **2016**, *17*, 180. [CrossRef] [PubMed]
133. Hubbard, A.H.; Zhang, X.; Jastrebski, S.; Singh, A.; Schmidt, C. Understanding the liver under heat stress with statistical learning: An integrated metabolomics and transcriptomics computational approach. *BMC Genomics* **2019**, *20*, 502. [CrossRef]
134. 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]
135. Kouznetsova, V.L.; Kim, E.; Romm, E.L.; Zhu, A.; Tsigelny, I.F. Recognition of early and late stages of bladder cancer using metabolites and machine learning. *Metabolomics* **2019**, *15*, 1–15. [CrossRef]
136. Guo, Y.; Yu, H.; Chen, D.; Zhao, Y.Y. Machine learning distilled metabolite biomarkers for early stage renal injury. *Metabolomics* **2020**, *16*. [CrossRef]
137. Kim, M.; Rai, N.; Zorraquino, V.; Tagkopoulos, I. Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nat. Commun.* **2016**, *7*, 13090. [CrossRef] [PubMed]
138. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.* **1995**, *57*, 289–300. [CrossRef]
139. Jafari, M.; Ansari-Pour, N. Why, when and how to adjust your P values? *Cell J.* **2019**, *20*, 604–607. [PubMed]
140. Karathanasis, N.; Tsamardinos, I.; Lagani, V. omicsNPC: Applying the Non-Parametric Combination Methodology to the Integrative Analysis of Heterogeneous Omics Data. *PLoS ONE* **2016**, *11*, e0165545. [CrossRef] [PubMed]
141. Jiang, B.; Zhang, X.; Zuo, Y.; Kang, G. A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. *J. Theor. Biol.* **2011**, *277*, 67–73. [CrossRef] [PubMed]

142. Taylor, J.; Tibshirani, R. A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics* **2006**, *7*, 167–181. [CrossRef]
143. Efron, B.; Tibshirani, R. Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **2002**, *23*, 70–86. [CrossRef]
144. Baker, M. Statisticians issue warning over misuse of P values. *Nature* **2016**, *531*, 151. [CrossRef]
145. Guo, H.; Chen, J.; Huang, Y.; Zhang, W.; Xu, F.; Zhang, Z. A pseudo-kinetics approach for time-series metabolomics investigations: More reliable and sensitive biomarkers revealed in vincristine-induced paralytic ileus rats. *RSC Adv.* **2016**, *6*, 54471–54478. [CrossRef]
146. Abadie, C.; Blanchet, S.; Carroll, A.; Tcherkez, G. Metabolomics analysis of postphotosynthetic effects of gaseous O2 on primary metabolism in illuminated leaves. *Funct. Plant Biol.* **2017**, *44*, 929. [CrossRef]
147. Yates, F. The Analysis of Multiple Classifications with Unequal Numbers in the Different Classes. *J. Am. Stat. Assoc.* **1934**, *29*, 51. [CrossRef]
148. Xia, J.; Sinelnikov, I.V.; Wishart, D.S. MetATT: A web-based metabolomics tool for analyzing time-series and two-factor datasets. *Bioinformatics* **2011**, *27*, 2455–2456. [CrossRef] [PubMed]
149. Berk, M.; Ebbels, T.; Montana, G. A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics* **2011**, *27*, 1979–1985. [CrossRef] [PubMed]
150. Gromski, P.S.; Muhamadali, H.; Ellis, D.I.; Xu, Y.; Correa, E.; Turner, M.L.; Goodacre, R. A tutorial review: Metabolomics and partial least squares-discriminant analysis—A marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* **2015**, *879*, 10–23. [CrossRef] [PubMed]
151. Brereton, R.G.; Lloyd, G.R. Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.* **2014**, *28*, 213–225. [CrossRef]
152. Rodríguez-Pérez, R.; Fernández, L.; Marco, S. Overoptimism in cross-validation when using partial least squares-discriminant analysis for omics data: A systematic study. *Anal. Bioanal. Chem.* **2018**, *410*, 5981–5992. [CrossRef]
153. Szymańska, E.; Saccenti, E.; Smilde, A.K.; Westerhuis, J.A. Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **2012**, *8*, 3–16. [CrossRef]
154. Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J.K.; Holmes, E.; Trygg, J. OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. *J. Chemom.* **2006**, *20*, 341–351. [CrossRef]
155. Chun, H.; Keleş, S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2010**, *72*, 3–25. [CrossRef]
156. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*. [CrossRef]
157. Li, Y.; Fan, T.W.M.; Lane, A.N.; Kang, W.-Y.; Arnold, S.M.; Stromberg, A.J.; Wang, C.; Chen, L. SDA: A semi-parametric differential abundance analysis method for metabolomics and proteomics data. *BMC Bioinform.* **2019**, *20*, 501–510. [CrossRef] [PubMed]
158. Gross, S.M.; Tibshirani, R. Collaborative regression. *Biostatistics* **2015**, *16*, 326–338. [CrossRef] [PubMed]
159. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
160. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc.* **1996**, *58*, 267–288. [CrossRef]
161. Zou, H.; Zou, H.; Hastie, T. Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
162. Fukushima, A. DiffCorr: An R package to analyze and visualize differential correlations in biological networks. *Gene* **2013**, *518*, 209–214. [CrossRef]
163. Siska, C.; Bowler, R.; Kechris, K. The discordant method: A novel approach for differential correlation. *Bioinformatics* **2016**, *32*, 690–696. [CrossRef]
164. Ma, J.; Karnovsky, A.; Afshinnia, F.; Wigginton, J.; Rader, D.J.; Natarajan, L.; Sharma, K.; Porter, A.C.; Rahman, M.; He, J.; et al. Differential network enrichment analysis reveals novel lipid pathways in chronic kidney disease. *Bioinformatics* **2019**, *35*, 3441–3452. [CrossRef]
165. Shi, W.J.; Zhuang, Y.; Russell, P.H.; Hobbs, B.D.; Parker, M.M.; Castaldi, P.J.; Rudra, P.; Vestal, B.; Hersh, C.P.; Saba, L.M.; et al. Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics* **2019**, *35*, 4336–4343. [CrossRef]

166. Siddiqui, J.K.; Baskin, E.; Liu, M.; Cantemir-Stone, C.Z.; Zhang, B.; Bonneville, R.; McElroy, J.P.; Coombes, K.R.; Mathé, E.A. IntLIM: Integration using linear models of metabolomics and gene expression data. *BMC Bioinform.* **2018**, *19*, 81. [CrossRef]
167. Fleming, R.M.T.; Vlassis, N.; Thiele, I.; Saunders, M.A. Conditions for duality between fluxes and concentrations in biochemical networks. *J. Theor. Biol.* **2016**, *409*, 1–10. [CrossRef] [PubMed]
168. Pandey, V.; Hadadi, N.; Hatzimanikatis, V. Enhanced flux prediction by integrating relative expression and relative metabolite abundance into thermodynamically consistent metabolic models. *PLoS Comput. Biol.* **2019**, *15*, e1007036. [CrossRef] [PubMed]
169. Angione, C. Human Systems Biology and Metabolic Modelling: A Review-From Disease Metabolism to Precision Medicine. *Biomed Res. Int.* **2019**, *2019*, 8304260. [CrossRef] [PubMed]
170. Lakshmanan, M.; Koh, G.; Chung, B.K.S.; Lee, D. Software applications for flux balance analysis. *Brief. Bioinform.* **2012**, *15*, 108–122. [CrossRef]
171. Rätsch, G.; Sonnenburg, S.; Schäfer, C. Learning interpretable SVMs for biological sequence classification. *BMC Bioinform.* **2006**, *7*, S9. [CrossRef]
172. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
173. Rasmussen, P.M.; Madsen, K.H.; Lund, T.E.; Hansen, L.K. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *Neuroimage* **2011**, *55*, 1120–1131. [CrossRef]
174. Eicher, T.; Sinha, K. A support vector machine approach to identification of proteins relevant to learning in a mouse model of Down Syndrome. In Proceedings of the International Joint Conference on Neural Networks, Anchorage, AK, USA, 14–19 May 2017.
175. Gaonkar, B.; Davatzikos, C. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. *Neuroimage* **2013**, *78*, 270–283. [CrossRef]
176. Breiman, L. Bagging Predictors. *Machin. Learn.* **1996**, *24*, 123–140. [CrossRef]
177. Quinlan, R. C4.5: Programs for Machine Learning. *Machin. Learn.* **1993**, *16*, 235–240. [CrossRef]
178. Archer, K.J.; Kimes, R.V. Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.* **2008**, *52*, 2249–2260. [CrossRef]
179. Taufik, W.M. Minimizing False Negatives of Measles Prediction Model: An Experimentation of Feature Selection Based On Domain Knowledge and Random Forest Classifier. *Int. J. Eng. Adv. Technol.* **2019**, 2249–8958.
180. Calle, M.L.; Urrea, V. Letter to the Editor: Stability of Random Forest importance measures. *Brief. Bioinform.* **2011**, *12*, 86–89. [CrossRef] [PubMed]
181. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef] [PubMed]
182. Van Rijsbergen, C.J. Foundation of evaluation. *J. Doc.* **1974**, *30*, 365–373. [CrossRef]
183. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
184. Minsky, M.; Papert, S. *Perceptrons; an introduction to computational geometry*; MIT Press: Cambridge, MA, USA, 1969; ISBN 9780262130431.
185. Lecun, Y.; Eon Bottou, L.; Bengio, Y.; Haaner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
186. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
187. Alwosheel, A.; van Cranenburgh, S.; Chorus, C.G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J. Choice Model.* **2018**, *28*, 167–182. [CrossRef]
188. Mirza, B.; Wang, W.; Wang, J.; Choi, H.; Chung, N.C.; Ping, P. Machine learning and integrative analysis of biomedical big data. *Genes (Basel)* **2019**, *10*, 87. [CrossRef]
189. Yu, H.; Samuels, D.C.; Zhao, Y.Y.; Guo, Y. Architectures and accuracy of artificial neural network for disease classification from omics data. *BMC Genomics* **2019**, *20*, 167. [CrossRef] [PubMed]
190. Tang, B.; Pan, Z.; Yin, K.; Khateeb, A. Recent Advances of Deep Learning in Bioinformatics and Computational Biology. *Front. Genet.* **2019**, *10*, 214. [CrossRef] [PubMed]

191. Ren, S.; Shao, Y.; Zhao, X.; Hong, C.S.; Wang, F.; Lu, X.; Li, J.; Ye, G.; Yan, M.; Zhuang, Z.; et al. Integration of metabolomics and transcriptomics reveals major metabolic pathways and potential biomarker involved in prostate cancer. *Mol. Cell. Proteomics* **2016**, *15*, 154–163. [CrossRef] [PubMed]
192. Torres, E.R.S.; Hall, R.; Bobe, G.; Choi, J.; Impey, S.; Pelz, C.; Lindner, J.R.; Stevens, J.F.; Raber, J. Integrated Metabolomics-DNA Methylation Analysis Reveals Significant Long-Term Tissue-Dependent Directional Alterations in Aminoacyl-tRNA Biosynthesis in the Left Ventricle of the Heart and Hippocampus Following Proton Irradiation. *Front. Mol. Biosci.* **2019**, *6*, 77. [CrossRef]
193. Yu, J.; Chen, J.; Zhao, H.; Gao, J.; Li, Y.; Li, Y.; Xue, J.; Dahan, A.; Sun, D.; Zhang, G.; et al. Integrative proteomics and metabolomics analysis reveals the toxicity of cationic liposomes to human normal hepatocyte cell line L02. *Mol. Omi.* **2018**, *14*, 362–372. [CrossRef]
194. Cao, H.; Zhang, A.; Sun, H.; Zhou, X.; Guan, Y.; Liu, Q.; Kong, L.; Wang, X. Metabolomics-proteomics profiles delineate metabolic changes in kidney fibrosis disease. *Proteomics* **2015**, *15*, 3699–3710. [CrossRef]
195. Erawijantari, P.P.; Mizutani, S.; Shiroma, H.; Shiba, S.; Nakajima, T.; Sakamoto, T.; Saito, Y.; Fukuda, S.; Yachida, S.; Yamada, T. Influence of gastrectomy for gastric cancer treatment on faecal microbiome and metabolome profiles. *Gut* **2020**. [CrossRef]
196. O'Donovan, C.M.; Madigan, S.M.; Garcia-Perez, I.; Rankin, A.; O' Sullivan, O.; Cotter, P.D. Distinct microbiome composition and metabolome exists across subgroups of elite Irish athletes. *J. Sci. Med. Sport* **2020**, *23*, 63–68. [CrossRef]
197. Cronin, O.; Barton, W.; Skuse, P.; Penney, N.C.; Garcia-Perez, I.; Murphy, E.F.; Woods, T.; Nugent, H.; Fanning, A.; Melgar, S.; et al. A Prospective Metagenomic and Metabolomic Analysis of the Impact of Exercise and/or Whey Protein Supplementation on the Gut Microbiome of Sedentary Adults. *mSystems* **2018**, *3*. [CrossRef]
198. Zachariou, M.; Minadakis, G.; Oulas, A.; Afxenti, S.; Spyrou, G.M. Integrating multi-source information on a single network to detect disease-related clusters of molecular mechanisms. *J. Proteomics* **2018**, *188*, 15–29. [CrossRef]
199. Maifiah, M.H.M.; Cheah, S.E.; Johnson, M.D.; Han, M.L.; Boyce, J.D.; Thamlikitkul, V.; Forrest, A.; Kaye, K.S.; Hertzog, P.; Purcell, A.W.; et al. Global metabolic analyses identify key differences in metabolite levels between polymyxin-susceptible and polymyxin-resistant Acinetobacter baumannii. *Sci. Rep.* **2016**, *6*, 22287. [CrossRef] [PubMed]
200. Xu, H.D.; Luo, W.; Lin, Y.; Zhang, J.; Zhang, L.; Zhang, W.; Huang, S.M. Discovery of potential therapeutic targets for non-small cell lung cancer using high-throughput metabolomics analysis based on liquid chromatography coupled with tandem mass spectrometry. *RSC Adv.* **2019**, *9*, 10905–10913. [CrossRef]
201. Nguyen, T.M.; Shafi, A.; Nguyen, T.; Draghici, S. Identifying significantly impacted pathways: A comprehensive review and assessment. *Genome Biol.* **2019**, *20*, 203. [CrossRef] [PubMed]
202. Johnson, C.H.; Ivanisevic, J.; Siuzdak, G. Metabolomics: Beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 451–459. [CrossRef]
203. Ewald, J.D.; Soufan, O.; Crump, D.; Hecker, M.; Xia, J.; Basu, N. EcoToxModules: Custom Gene Sets to Organize and Analyze Toxicogenomics Data from Ecological Species. *Environ. Sci. Technol.* **2020**. [CrossRef]
204. Lee, J.; Jo, K.; Lee, S.; Kang, J.; Kim, S. Prioritizing biological pathways by recognizing context in time-series gene expression data. *BMC Bioinform.* **2016**, *17*, 477. [CrossRef]
205. Falcon, S.; Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **2007**, *23*, 257–258. [CrossRef]
206. Maere, S.; Heymans, K.; Kuiper, M. BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **2005**, *21*, 3448–3449. [CrossRef]
207. Koelmel, J.P.; Ulmer, C.Z.; Jones, C.M.; Yost, R.A.; Bowden, J.A. Common cases of improper lipid annotation using high-resolution tandem mass spectrometry data and corresponding limitations in biological interpretation. *Biochim. et Biophys. Acta (BBA) - Mol. Cell Boil. Lipids* **2017**, *1862*, 766–770. [CrossRef]
208. Fisher, R.A. *Statistical Methods for Research Workers*, 5th ed.; Oliver and Boyd: Edinburgh, UK, 1934.
209. Stouffer, S.A.; Suchman, E.A.; Devinney, L.C.; Star, S.A.; Williams, R.M. The American soldier: Adjustment during army life. In *Studies in social psychology in World War II*; Princeton University Press: Princeton, NJ, USA, 1949; Volume 1.

210. Zhang, B.; Hu, S.; Baskin, E.; Patt, A.; Siddiqui, J.K.; Mathé, E.A. RaMP: A Comprehensive Relational Database of Metabolomics Pathways for Pathway Enrichment Analysis of Genes and Metabolites. *Metabolites* **2018**, *8*, 16. [CrossRef]
211. Kamburov, A.; Cavill, R.; Ebbels, T.M.D.; Herwig, R.; Keun, H.C. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* **2011**, *27*, 2917–2918. [CrossRef] [PubMed]
212. Kaever, A.; Landesfeind, M.; Feussner, K.; Morgenstern, B.; Feussner, I.; Meinicke, P. Meta-analysis of pathway enrichment: Combining independent and dependent omics data sets. *PLoS ONE* **2014**, *9*, e89297. [CrossRef] [PubMed]
213. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
214. Xia, J.; Wishart, D.S. MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* **2010**, *38*, W71–W77. [CrossRef] [PubMed]
215. Molenaar, M.R.; Jeucken, A.; Wassenaar, T.A.; Van De Lest, C.H.A.; Brouwers, J.F.; Helms, J.B. LION/web: A web-based ontology enrichment tool for lipidomic data analysis. *Gigascience* **2019**, *8*. [CrossRef]
216. Tarca, A.; Draghici, S.; Khatri, P. A novel signaling pathway impact analysis. *Bioinformatics* **2009**, *25*, 75–82. [CrossRef]
217. Ibrahim, M.A.H.; Jassim, S.; Cawthorne, M.A.; Langlands, K. A topology-based score for pathway enrichment. *J. Comput. Biol.* **2012**, *19*, 563–573. [CrossRef]
218. Gu, Z.; Wang, J. CePa: An R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics* **2013**, *29*, 658–660. [CrossRef]
219. Gao, S.; Wang, X. TAPPA: Topological analysis of pathway phenotype association. *Bioinformatics* **2007**, *23*, 3100–3102. [CrossRef]
220. Massa, M.S.; Chiogna, M.; Romualdi, C. Gene set analysis exploiting the topology of a pathway. *BMC Syst. Biol.* **2010**, *4*, 121. [CrossRef]
221. Martini, P.; Sales, G.; Massa, M.S.; Chiogna, M.; Romualdi, C. Along signal paths: An empirical gene set approach exploiting pathway topology. *Nucleic Acids Res.* **2013**, *41*, e19. [CrossRef] [PubMed]
222. Jacob, L.; Neuvial, P.; Dudoit, S. Gains in Power from Structured Two-Sample Tests of Means on Graphs. *Ann. Appl. Stat.* **2010**, *6*, 561–600. [CrossRef]
223. Ihnatova, I.; Popovici, V.; Budinska, E. A critical comparison of topology-based pathway analysis methods. *PLoS ONE* **2018**, *13*, e0191154. [CrossRef] [PubMed]
224. Picart-Armada, S.; Fernández-Albert, F.; Vinaixa, M.; Yanes, O.; Perera-Lluna, A. FELLA: An R package to enrich metabolomics data. *BMC Bioinform.* **2018**, *19*, 538. [CrossRef]
225. Paley, S.M.; Karp, P.D. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res.* **2006**, *34*, 3771–3778. [CrossRef]
226. Junker, B.H.; Klukas, C.; Schreiber, F. Vanted: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinform.* **2006**, *7*, 109. [CrossRef]
227. Hernández-de-Diego, R.; Tarazona, S.; Martínez-Mira, C.; Balzano-Nogueira, L.; Furió-Tarí, P.; Pappas, G.J.; Conesa, A. PaintOmics 3: A web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.* **2018**, *46*, W503–W509. [CrossRef]
228. Domingo-Fernández, D.; Hoyt, C.T.; Bobis-Álvarez, C.; Marín-Llaó, J.; Hofmann-Apitius, M. ComPath: An ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst. Biol. Appl.* **2019**, *5*. [CrossRef]
229. 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*, 2498–2504. [CrossRef]
230. Gansner, E.R.; North, S.C. An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper.* **2000**, *11*, 1203–1233. [CrossRef]
231. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Sy.* **2006**.

232. Kutmon, M.; van Iersel, M.P.; Bohler, A.; Kelder, T.; Nunes, N.; Pico, A.R.; Evelo, C.T. PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLoS Comput. Biol.* **2015**, *11*, e1004085. [CrossRef] [PubMed]
233. Zhou, G.; Xia, J. OmicsNet: A web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Res.* **2018**, *46*, W514–W522. [CrossRef] [PubMed]
234. Rougny, A.; Touré, V.; Moodie, S.; Balaur, I.; Czauderna, T.; Borlinghaus, H.; Dogrusoz, U.; Mazein, A.; Dräger, A.; Blinov, M.L.; et al. Systems Biology Graphical Notation: Process Description language Level 1 Version 2.0. *J. Integr. Bioinform.* **2019**, *16*. [CrossRef] [PubMed]
235. Klyne, G.; Carroll, J.; McBride, B. Resource Description Framework (RDF): Concepts and Abstract Syntax. Available online: https://www.w3.org/TR/rdf-concepts/ (accessed on 25 March 2020).
236. Frainay, C.; Schymanski, E.; Neumann, S.; Merlet, B.; Salek, R.; Jourdan, F.; Yanes, O. Mind the Gap: Mapping Mass Spectral Databases in Genome-Scale Metabolic Networks Reveals Poorly Covered Areas. *Metabolites* **2018**, *8*, 51. [CrossRef]
237. Mubeen, S.; Hoyt, C.T.; Gemünd, A.; Hofmann-Apitius, M.; Fröhlich, H.; Domingo-Fernández, D. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Front. Genet.* **2019**, *10*. [CrossRef]
238. Slenter, D.N.; Kutmon, M.; Hanspers, K.; Riutta, A.; Windsor, J.; Nunes, N.; Mélius, J.; Cirillo, E.; Coort, S.L.; Digles, D.; et al. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **2018**, *46*, D661–D667. [CrossRef]
239. 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]
240. Cerami, E.G.; Gross, B.E.; Demir, E.; Rodchenkov, I.; Babur, O.; Anwar, N.; Schultz, N.; Bader, G.D.; Sander, C. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **2011**, *39*, D685–D690. [CrossRef]
241. Tran, V.D.T.; Moretti, S.; Coste, A.T.; Amorim-Vaz, S.; Sanglard, D.; Pagni, M. Condition-specific series of metabolic sub-networks and its application for gene set enrichment analysis. *Bioinformatics* **2019**, *35*, 2258–2266. [CrossRef]
242. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [CrossRef] [PubMed]
243. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]
244. Krieger, C.J. MetaCyc: A multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **2004**, *32*, 438–442. [CrossRef] [PubMed]
245. Le Boulch, M.; Déhais, P.; Combes, S.; Pascal, G. MACADAM database: A MetAboliC pAthways DAtabase for Microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups. *Database* **2019**, *2019*. [CrossRef]
246. Wishart, D.; Li, C.; Marcu, A.; Badran, H.; Pon, A.; Budinski, Z.; Patron, J.; Lipton, D.; Cao, X.; Oler, E.O.; et al. PathBank: A comprehensive pathway database for model organisms. *Nucleic Acids Res.* **2019**, *48*, 470–478. [CrossRef]
247. Barbarino, J.M.; Whirl-Carrillo, M.; Altman, R.B.; Klein, T.E. PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2018**, *10*, 1417. [CrossRef]
248. Heller, S.R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **2015**, *7*, 23. [CrossRef]
249. Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; et al. The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Res.* **2012**, *41*, 456–463. [CrossRef]
250. Salek, R.M.; Steinbeck, C.; Viant, M.R.; Goodacre, R.; Dunn, W.B. The role of reporting standards for metabolite annotation and identification in metabolomic studies. *Gigascience* **2013**, *2*, 13. [CrossRef]
251. Jamil, H.M. Improving Integration Effectiveness of ID Mapping Based Biological Record Linkage. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12*, 473–486. [CrossRef]

252. Rocca-Serra, P.; Salek, R.M.; Arita, M.; Correa, E.; Dayalan, S.; Gonzalez-Beltran, A.; Ebbels, T.; Goodacre, R.; Hastings, J.; Haug, K.; et al. Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* **2015**, *12*, 14. [CrossRef] [PubMed]
253. Wohlgemuth, G.; Haldiya, P.K.; Willighagen, E.; Kind, T.; Fiehn, O. The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. *Bioinformatics* **2010**, *26*, 2647–2648. [CrossRef] [PubMed]
254. Ravikumar, K.E.; Wagholikar, K.B.; Li, D.; Kocher, J.-P.; Liu, H. Text mining facilitates database curation - extraction of mutation-disease associations from Bio-medical literature. *BMC Bioinform.* **2015**, *16*, 185. [CrossRef] [PubMed]
255. Ruch, P. Text Mining to Support Gene Ontology Curation and Vice Versa. *Methods Mol. Biol.* **2017**, *1446*, 69–84. [PubMed]
256. Galeota, E.; Kishore, K.; Pelizzola, M. Ontology-driven integrative analysis of omics data through Onassis. *Sci. Rep.* **2020**, *10*, 1–9. [CrossRef]
257. Mohanraj, K.; Karthikeyan, B.S.; Vivek-Ananth, R.P.; Chand, R.P.B.; Aparna, S.R.; Mangalapandi, P.; Samal, A. IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry and Therapeutics. *Sci. Rep.* **2018**, *8*, 4329. [CrossRef]
258. Liu, Y.; Liang, Y.; Wishart, D. PolySearch2: A significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Res.* **2015**, *43*, W535–W542. [CrossRef]
259. Tutubalina, E.V.; Miftahutdinov, Z.S.; Nugmanov, R.I.; Madzhidov, T.I.; Nikolenko, S.I.; Alimova, I.S.; Tropsha, A.E. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects. *Russ. Chem. Bull.* **2017**, *66*, 2180–2189. [CrossRef]
260. Kulkarni, C.; Xu, W.; Ritter, A.; Machiraju, R. An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols. In Proceedings of the NAACL-HLT 2018, New Orleans, LA, USA, 1–6 June 2018; pp. 97–106.
261. Westergaard, D.; Staerfeldt, H.-H.; Tønsberg, C.; Jensen, L.J.; Brunak, S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLOS Comput. Biol.* **2018**, *14*, e1005962. [CrossRef]
262. Ben Abdessalem Karaa, W.; Mannai, M.; Dey, N.; Ashour, A.S.; Olariu, I. Gene-disease-food relation extraction from biomedical database. *Adv. Intell. Syst. Comput.* **2018**, *633*, 394–407.
263. Nikfarjam, A.; Ransohoff, J.D.; Callahan, A.; Jones, E.; Loew, B.; Kwong, B.Y.; Sarin, K.Y.; Shah, N.H. Early detection of adverse drug reactions in social health networks: A natural language processing pipeline for signal detection. *J. Med. Internet Res.* **2019**, *5*, e11264. [CrossRef] [PubMed]
264. Fan, Y.; Zhang, R. Using natural language processing methods to classify use status of dietary supplements in clinical notes. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 15–22. [CrossRef] [PubMed]
265. Huan, C.-C.; Lu, Z. Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Brief. Bioinform.* **2015**, *17*, 132–144. [CrossRef] [PubMed]
266. Cohen, K.B.; Lanfranchi, A.; Choi, M.J.Y.; Bada, M.; Baumgartner, W.A.; Panteleyeva, N.; Verspoor, K.; Palmer, M.; Hunter, L.E. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinform.* **2017**, *18*, 372. [CrossRef] [PubMed]
267. Pletscher-Frankild, S.; Pallejà, A.; Tsafou, K.; Binder, J.X.; Jensen, L.J. DISEASES: Text mining and data integration of disease-gene associations. *Methods* **2015**, *74*, 83–89. [CrossRef] [PubMed]
268. Fdez-Glez, J.; Ruano-Ordás, D.; Méndez, J.R.; Fdez-Riverola, F.; Laza, R.; Pavón, R. Determining the Influence of Class Imbalance for the Triage of Biomedical Documents. *Curr. Bioinform.* **2017**, *13*, 592–605. [CrossRef]
269. Wei, C.-H.; Allot, A.; Leaman, R.; Lu, Z. PubTator central: Automated concept annotation for biomedical full text articles. *Nucleic Acids Res.* **2019**, *47*, W587–W593. [CrossRef]
270. Jiang, X.; Ringwald, M.; Blake, J.A.; Arighi, C.; Zhang, G.; Shatkay, H. An effective biomedical document classification scheme in support of biocuration: Addressing class imbalance. *Database* **2019**, *2019*. [CrossRef]
271. Alshuwaier, F.; Areshey, A.; Poon, J. A comparative study of the current technologies and approaches of relation extraction in biomedical literature using text mining. In Proceedings of the 4th IEEE International Conference on Engineering Technologies and Applied Sciences, Salmabad, Bahrain, 29 November–1 December 2017; pp. 1–13.

272. Sung, J.; Kim, S.; Cabatbat, J.J.T.; Jang, S.; Jin, Y.S.; Jung, G.Y.; Chia, N.; Kim, P.J. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. *Nat. Commun.* **2017**, *8*, 1–12.
273. Griffith, M.; Griffith, O.L.; Coffman, A.C.; Weible, J.V.; McMichael, J.F.; Spies, N.C.; Koval, J.; Das, I.; Callaway, M.B.; Eldred, J.M.; et al. DGIdb: Mining the druggable genome. *Nat. Methods* **2013**, *10*, 1209–1210. [CrossRef]
274. Chen, T.; Li, M.; He, Q.; Zou, L.; Li, Y.; Chang, C.; Zhao, D.; Zhu, Y. LiverWiki: A wiki-based database for human liver. *BMC Bioinform.* **2017**, *18*, 452. [CrossRef] [PubMed]
275. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; et al. ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **2016**, *8*, 61. [CrossRef] [PubMed]
276. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; et al. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2015**, *44*, 1202–1213. [CrossRef] [PubMed]
277. Banerjee, P.; Erehman, J.; Gohlke, B.O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II-a database of natural products. *Nucleic Acids Res.* **2015**, *43*, D935–D939. [CrossRef] [PubMed]
278. Karp, P.D.; Paley, S.; Romero, P. The pathway tools software. *Bioinformatics* **2002**, *18*. [CrossRef]
279. Fadason, T.; Schierding, W.; Kolbenev, N.; Liu, J.; Ingram, J.; O'Sullivan, J.M. Reconstructing the blood metabolome and genotype using long-range chromatin interactions. *bioRxiv* **2019**, 656132. [CrossRef]
280. Le, V.; Quinn, T.P.; Tran, T.; Venkatesh, S. Deep in the Bowel: Highly Interpretable Neural Encoder-Decoder Networks Predict Gut Metabolites from Gut Microbiome. *bioRxiv* **2019**, 686394. [CrossRef]
281. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2 (NIPS'13), New York, NY, USA, 5–10 December 2013; pp. 3111–3119.
282. Morton, J.T.; Aksenov, A.A.; Nothias, L.F.; Foulds, J.R.; Quinn, R.A.; Badri, M.H.; Swenson, T.L.; Van Goethem, M.W.; Northen, T.R.; Vazquez-Baeza, Y.; et al. Learning representations of microbe–metabolite interactions. *Nat. Methods* **2019**, *16*, 1306–1314. [CrossRef]
283. Romero, P.; Wagg, J.; Green, M.L.; Kaiser, D.; Krummenacker, M.; Karp, P.D. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **2004**, *6*, R2. [CrossRef]
284. Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **2007**, *36*, D344–D350. [CrossRef]
285. Luscombe, N.M.; Greenbaum, D.; Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* **2001**, *40*, 346–358. [CrossRef] [PubMed]
286. Baggerly, K. Disclose all data in publications. *Nature* **2010**, *467*, 401. [CrossRef] [PubMed]
287. Hüllermeier, E. Fuzzy methods in machine learning and data mining: Status and prospects. *Fuzzy Sets Syst.* **2005**, *156*, 387–406. [CrossRef]
288. Bonneau, G.P.; Hege, H.C.; Johnson, C.R.; Oliveira, M.M.; Potter, K.; Rheingans, P.; Schultz, T. Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization Uncertainty, Multifield, Biomedical, and Scalable Visualization*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 37, pp. 3–27.