biomolecules

MDPI

Review

Statistical Methods for Multi-Omics Analysis in Neurodevelopmental Disorders: From High Dimensionality to Mechanistic Insight

Manuel Airoldi [1], Veronica Remori [1] and Mauro Fasano [1,2,*]

[1] Department of Science and High Technology, University of Insubria, 22100 Como, Italy; mairoldi@uninsubria.it (M.A.); vremori@uninsubria.it (V.R.)
[2] Center of Neuroscience, University of Insubria, 21052 Busto Arsizio, Italy
* Correspondence: mauro.fasano@uninsubria.it

Abstract

Neurodevelopmental disorders (NDDs), including autism spectrum disorder, intellectual disability, and attention-deficit/hyperactivity disorder, are genetically and phenotypically heterogeneous conditions affecting millions worldwide. High-throughput omics technologies—transcriptomics, proteomics, metabolomics, and epigenomics—offer a unique opportunity to link genetic variation to molecular and cellular mechanisms underlying these disorders. However, the high dimensionality, sparsity, batch effects, and complex covariance structures of omics data present significant statistical challenges, requiring robust normalization, batch correction, imputation, dimensionality reduction, and multivariate modeling approaches. This review provides a comprehensive overview of statistical frameworks for analyzing high-dimensional omics datasets in NDDs, including univariate and multivariate models, penalized regression, sparse canonical correlation analysis, partial least squares, and integrative multi-omics methods such as DIABLO, similarity network fusion, and MOFA. We illustrate how these approaches have revealed convergent molecular signatures—synaptic, mitochondrial, and immune dysregulation—across transcriptomic, proteomic, and metabolomic layers in human cohorts and experimental models. Finally, we discuss emerging strategies, including single-cell and spatially resolved omics, machine learning-driven integration, and longitudinal multi-modal analyses, highlighting their potential to translate complex molecular patterns into mechanistic insights, biomarkers, and therapeutic targets. Integrative multi-omics analyses, grounded in rigorous statistical methodology, are poised to advance mechanistic understanding and precision medicine in NDDs.

Keywords: neurodevelopmental disorders; multi-omics integration; study design; wide data

Academic Editor: Trupti Joshi

Received: 7 September 2025
Revised: 28 September 2025
Accepted: 1 October 2025
Published: 2 October 2025

Citation: Airoldi, M.; Remori, V.; Fasano, M. Statistical Methods for Multi-Omics Analysis in Neurodevelopmental Disorders: From High Dimensionality to Mechanistic Insight. *Biomolecules* **2025**, *15*, 1401. https://doi.org/10.3390/biom15101401

1. Introduction

Neurodevelopmental disorders (NDDs), including autism spectrum disorder (ASD), intellectual disability (ID), attention-deficit/hyperactivity disorder (ADHD), Rett syndrome, and CDKL5 deficiency disorder, represent a diverse group of conditions that affect approximately 1–3% of children worldwide. Meta-analyses estimate ASD prevalence at ~0.7–1% globally [1,2], ID around 2% [3], and ADHD near 10% in boys and 5% in girls [4]. Rett syndrome and CDKL5 deficiency disorder are much rarer, with estimated prevalences of ~1 in 10,000–15,000 and ~1 in 40,000–60,000 live births, respectively, and are typically caused

by highly penetrant monogenic mutations [5,6]. Despite their rarity, these monogenic disorders have provided critical mechanistic insights into neurodevelopmental pathways and synaptic function, serving as model systems to understand complex NDDs.

The genetic architecture of NDDs is highly heterogeneous, encompassing rare, high-penetrance variants (e.g., de novo single-nucleotide variants and copy number variants) as well as the cumulative effects of common alleles contributing to polygenic risk [7,8]. Large-scale sequencing studies, including whole-exome sequencing (WES) and whole-genome sequencing (WGS), have substantially increased diagnostic yields, reaching 20–40% in trio-based studies. Emerging long-read sequencing technologies and integrative approaches that combine WES/WGS with transcriptomic, epigenomic, or methylome profiling are further enhancing diagnostic resolution [9,10]. Nevertheless, a substantial gap remains between variant detection and functional or mechanistic interpretation, which underscores the necessity of complementary molecular and computational strategies [11].

High-throughput omics technologies provide complementary perspectives on these mechanisms. Transcriptomic profiling captures gene expression dysregulation, alternative splicing, and allele-specific expression in NDDs [12], while proteomics quantifies protein abundance, post-translational modifications (PTMs) such as phosphorylation or ubiquitination, and protein–protein interactions, providing functional insights that are not always inferable from RNA-level data [3,4]. Correlations between mRNA and protein levels in human brain tissue are often modest, reflecting complex post-transcriptional regulation, protein turnover, and tissue-specific translational control [5]. Structural proteomics, phosphoproteomics, and interactomics are increasingly leveraged to understand synaptic function, neuroplasticity, and signaling pathway alterations in ASD and related NDDs [13]. Integrating these molecular layers offers the potential to bridge genetic variation with cellular phenotypes and disease-relevant pathways.

Beyond single-layer analyses, the integration of multiple molecular layers—genomic, transcriptomic, proteomic, metabolomic, and epigenomic— provides the opportunity to bridge genetic variation with cellular phenotypes, signaling pathways, and disease-relevant networks. Such multi-omics integration is particularly valuable in NDDs, where perturbations are often subtle, distributed across interconnected pathways, and context-dependent. For example, a rare de novo mutation in a synaptic gene may trigger downstream dysregulation of both protein networks and metabolic pathways, which can only be captured through integrative approaches.

However, omics studies in NDDs face persistent statistical and computational challenges. High dimensionality, batch effects, sparsity, and multiple testing burdens complicate the analysis of genome-scale data. Proper normalization and variance modeling are critical to distinguish biological signal from technical noise. For instance, DESeq2's median-of-ratios approach addresses library size variability in RNA-seq [6], whereas proteomics datasets often rely on quantile normalization, internal reference standards, or vendor-specific algorithms to mitigate technical artifacts. Quality control (QC) steps, which include assessment of sample integrity, detection of technical outliers, and evaluation of dataset-wide metrics such as mapping rates, duplication levels, or signal-to-noise ratios, are equally important. Indeed, poor QC can severely compromise downstream inference, introducing artifacts that persist even after normalization and batch correction. For example, outlier samples due to RNA degradation or low sequencing depth can distort differential expression analyses and bias integrative modeling. In neurodevelopmental disorder studies, where phenotypic and molecular variability is already high, failure to identify and exclude low-quality data can exacerbate false discoveries or obscure true biological signals [14,15]. Additionally, the adjustment for latent or known confounders, such as age, sex, brain region, or technical covariates, is critical, particularly in NDD studies

where case–control imbalances or developmental stage effects are common [16]. Downstream, shrinkage estimation, penalized regression, and feature filtering are necessary to extract robust, biologically meaningful signals from high-dimensional datasets [7–9]. These challenges are amplified in multi-omics integration, where heterogeneous data types and differing levels of missingness necessitate careful modeling. Importantly, the complexity of NDDs also arises from phenotypic heterogeneity and overlapping comorbidities. Patients may present with combinations of cognitive, behavioral, and motor deficits, often influenced by environmental factors and developmental stages. This heterogeneity complicates cohort selection, study design, and the generalizability of findings. Moreover, sex differences and ancestry-specific genetic effects can contribute to differential risk and molecular phenotypes, emphasizing the need for stratified analyses and careful consideration of confounding factors.

Overall, advances in high-throughput omics technologies, together with sophisticated computational methods, have opened unprecedented opportunities to dissect the molecular architecture of NDDs. Integrating genomic, transcriptomic, and proteomic data allows for the mapping of disease-associated variants to functional consequences, regulatory networks, and cellular phenotypes. By leveraging these approaches, researchers can identify convergent molecular pathways, propose candidate biomarkers, and ultimately inform precision therapeutic strategies for diverse NDD populations.

2. Statistical Challenges in High-Dimensional Omics Data

High-throughput omics platforms generate so-called wide data, characterized by thousands of features measured in relatively small sample cohorts. This "large p, small n" scenario (where the number of features greatly exceeds the number of samples) increases the risk of overfitting, spurious associations, and irreproducible findings if not properly managed [17,18]. This imbalance complicates traditional statistical inference because standard methods assume that the number of observations exceeds the number of variables, a condition violated in typical omics datasets. Consequently, specialized statistical frameworks that explicitly model noise, dependence structures, and sparsity are necessary to ensure robust inference and reproducibility.

Data preprocessing procedures, such as normalization, are critical first steps to mitigate technical artifacts, addressing biases such as library size variability in RNA-seq or labeling and ionization differences in mass spectrometry-based proteomics. Common transcriptomic normalization methods include the median-of-ratios implemented in DESeq2 [19], trimmed mean of M values (TMM) from edgeR [20] and quantile normalization [21]. Proteomics normalization often relies on quantile scaling, internal reference standards, or variance-stabilizing normalization [22]. Failure to appropriately normalize data can result in confounding technical variation with biological differences, leading to false conclusions. Recent advances also include methods such as RUVSeq (Remove Unwanted Variation) that leverage control genes or samples to improve normalization accuracy [23]. Normalization strategies must be tailored to the omics platform and experimental design; no one-size-fits-all approach exists.

Batch effects and hidden confounders constitute another major challenge. Differences in sample handling, reagents, instrumentation, or even operator can introduce systematic noise that obscures true biological signals [24,25]. Surrogate variable analysis (SVA) [26] and factor-based methods [27] are widely applied in transcriptomics. In contrast, ComBat [24] and Limma's *removeBatchEffect()* [28] are widely used in proteomics, although they were originally designed for transcriptomic data. These methods aim to preserve biological heterogeneity while mitigating technical artifacts, though overcorrection can inadvertently remove relevant signals. In addition, emerging approaches such as harmonization via

mutual nearest neighbors (MNN) and deep learning-based batch correction algorithms are gaining traction for their ability to handle complex batch structures, especially in single-cell omics [29,30]. In NDD studies, batch correction is particularly critical when combining data across brain regions, developmental stages, or experimental models (e.g., cerebral organoids, iPSC-derived neurons), which may introduce subtle but biologically meaningful variance that can be mistaken for noise.

Cohort heterogeneity adds another layer of complexity. Differences in sex, age, ancestry, disease severity, comorbidities, and medication status can all influence molecular measurements, introducing variance that is not disease-related [31,32]. Study design factors, including sampling strategies, tissue type, postmortem interval, and developmental stage, further introduce variance that may obscure true disease-associated signals [33]. Longitudinal and repeated-measures designs help mitigate some of these challenges by capturing intra-individual variability over time, thereby improving statistical power to detect disease-relevant changes [34]. Integrative analyses of single-cell and spatially resolved omics can deconvolve mixed cell populations, revealing cell-type-specific effects that are otherwise hidden in bulk measurements [35]. Computational frameworks that model heterogeneity explicitly, including mixed-effects models, Bayesian hierarchical approaches, and matrix factorization methods, have been shown to improve robustness and reproducibility in NDD studies [26]. These models not only account for known sources of variability but also allow for the discovery of latent structures in the data, such as patient subgroups or tissue-specific modules, that may correspond to distinct disease mechanisms or developmental trajectories. Figure 1 provides an overview of key sources of biological and technical variability that should be considered when designing multi-omics studies of neurodevelopmental disorders, including cohort characteristics, sample-related factors, and study design structure. Addressing cohort heterogeneity and study design confounders is particularly important when datasets are integrated across multiple sites, experimental platforms, or omics layers. Failure to account for these factors can lead to biased conclusions, inflated false positives, and poor generalizability of findings, undermining the potential of integrative multi-omics approaches.

Missing values and zero inflation are particularly problematic in proteomics datasets acquired via data-dependent acquisition (DDA), where observed zeros may reflect stochastic sampling rather than true absence of expression. While the optimal solution to the missing values problem is to prevent them during data collection, by optimizing sample preparation, increasing analytical depth, or using data-independent acquisition methods, this is not always feasible, particularly in clinical or resource-limited settings. Thus, several strategies are commonly employed when this is not feasible. Listwise deletion, one of the most common procedures, involves excluding observations containing missing values, but this approach may remove a large fraction of the original data [36]. Alternatively, imputation strategies, including K-nearest neighbors, random forest-based approaches, or left-censored minimal value imputation help preserve variance structure [37]. However, each method presents specific trade-offs: K-nearest neighbors and random forest imputation can preserve local data structure but may introduce bias when missingness is not random, while minimal value imputation risks underestimating true variability and inflating false positives. Poorly chosen imputation strategies can distort the underlying variance structure of the dataset, leading to misleading clustering, spurious associations, or attenuation of biological signals. Notably, recent advances in probabilistic modeling and deep learning, such as Bayesian missing data models and autoencoder-based imputations, offer promising avenues to recover missing proteomic measurements while retaining biological variability [38,39]. Yet, these methods also require careful parameter tuning and validation, and their assumptions may not be appropriate for all omics contexts. A practical

approach should therefore involve careful assessment of the missing data mechanism and the specific characteristics of the omics layer being analyzed.

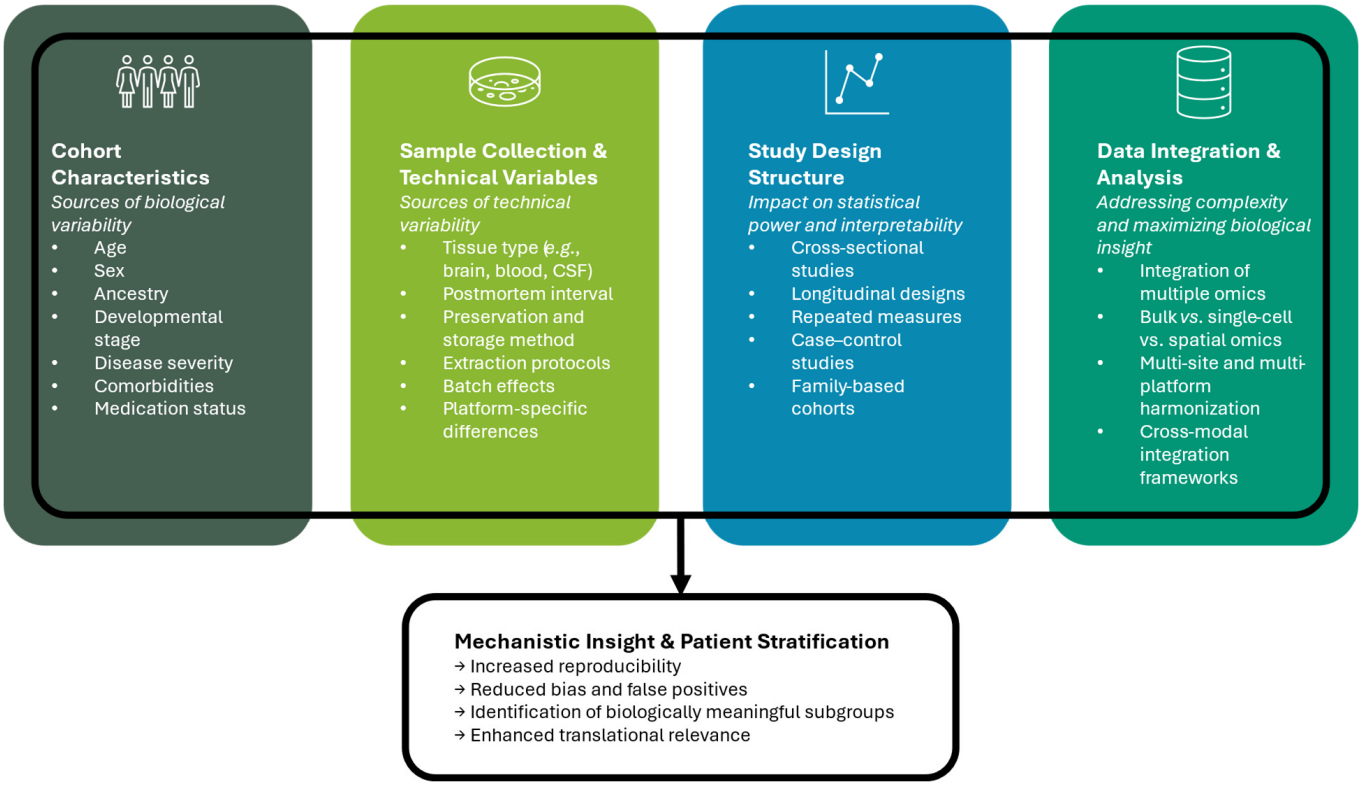

Figure 1. Overview of key factors influencing experimental design in multi-omics studies of neurodevelopmental disorders (NDDs). The figure illustrates major sources of variability that must be considered in study design, including biological characteristics of the cohort, sample-related and technical factors, and the type of study design employed. These elements contribute to both inter- and intra-subject heterogeneity and have critical implications for data quality, interpretability, and downstream analysis in multi-omics research.

The multiple testing burden is another central challenge. Tens of thousands of features are typically analyzed simultaneously, necessitating control of false positive occurrences. There is a number of existing methods addressing this problem, and one of the most commonly used and simple approach is the Bonferroni correction, which relies on the Family Wise Error Rate (FWER). The Bonferroni method controls FWER by computing the adjusted p-values, multiplying the original p-values by the number of tested hypotheses. A drawback of this procedure is that it is extremely conservative. The Benjamini–Hochberg procedure [40] increases the method's power by controlling the false discovery rate (FDR), with advanced approaches such as shrinkage-based estimators and penalized regression methods (e.g., LASSO, Elastic Net, Priority-Elastic Net) improving power while controlling false positives [41]. In addition, hierarchical testing frameworks and adaptive FDR methods have been developed to exploit the dependency structure of omics data, further enhancing detection power [42,43].

Dimensionality reduction and the management of multicollinearity are essential steps to facilitate both pattern discovery and predictive modeling in high-dimensional biological data. As said before, large scale omics datasets contained thousands of correlated features, which can obscure meaningful biological signals and compromise statistical inference. Matrix factorization methods, such as principal component analysis (PCA) [44] and independent component analysis (ICA) [45], provide a general framework to address these problems by decomposing the original dataset into a reduced set of latent variables

that capture the most relevant sources of variation. Closely related latent variable methods, including partial least squares discriminant analysis (PLS-DA), and sparse canonical correlation analysis (sCCA) [46,47], extend this principle by incorporating group information or integrating multiple datasets. Collectively, these methods provide a powerful toolkit to summarize high-dimensional datasets while retaining key biological variation. These approaches reduce overfitting, enhance interpretability, and provide a foundation for integrative multi-omics analyses. Moreover, nonlinear methods such as t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) have become popular for visualizing complex omics data in lower dimensions, facilitating the identification of subtle biological subgroups [48].

Careful attention to normalization, batch correction, missing data handling, multiple testing, and dimensionality reduction forms the backbone of robust statistical frameworks for high-dimensional omics studies [17,19,25,28], essential before applying analyses to complex biological contexts such as neurodevelopmental disorders. Equally important is the validation of findings across independent datasets or cohorts, which can strengthen biological conclusions and mitigate overfitting. Validation, benchmarking, and reproducibility are critical components to ensure the reliability of multi-omics findings. Independent dataset validation, either through external cohorts or cross-validation strategies, helps confirm that identified biomarkers or molecular signatures are robust and generalizable across different populations and technical conditions [49]. Benchmarking studies comparing different statistical methods and integration pipelines provide valuable insights into their relative performance, strengths, and limitations, guiding appropriate method selection for specific research questions [50]. Furthermore, reproducibility is challenged by variations in data preprocessing, batch effects, and analytical choices. Thus, transparent reporting, standardized workflows, and availability of code and data are essential to enable reproducibility and facilitate cumulative knowledge building in the NDD field. As omics data continue to grow in volume and complexity, ongoing methodological developments will be critical to unlock their full potential in deciphering disease mechanisms and guiding precision medicine.

3. Univariate vs. Multivariate Models in Transcriptomics and Proteomics

Omics datasets in NDDs are often high-dimensional, with thousands of genes, proteins, or metabolites measured in relatively small cohorts. These characteristics pose unique statistical and computational challenges. Selecting the appropriate analytical strategy is critical to avoid spurious associations, control overfitting, and maximize biological insight. Univariate approaches examine each feature independently to test for differences across conditions. They are computationally efficient, straightforward to interpret, and well-suited for hypothesis-driven analyses. In transcriptomics, commonly used tools include DESeq2 [19], edgeR [28], and Limma [28], which model gene expression data under assumptions appropriate for count-based or normalized data and employ robust variance estimation and empirical Bayes shrinkage. The choice between DESeq2, edgeR, and Limma depends on the specific characteristics of the dataset and the assumptions underlying each method. DESeq2 and edgeR both rely on negative binomial models but differ in their approaches to normalization and dispersion estimation. DESeq2 applies a median-of-ratios normalization and uses more conservative dispersion estimates, making it particularly suitable for datasets with moderate-to-high variance and small sample sizes, conditions frequently encountered in NDD studies. edgeR, while also robust, tends to perform better with larger sample sizes or higher counts per feature due to its more flexible estimation of dispersion. Limma, originally developed for microarray data and extended to RNA-seq through the voom transformation, is effective when working with normalized expression

values and leverages empirical Bayes shrinkage to stabilize variance estimates across features, often performing well in small-N scenarios. In NDD studies, where cohorts are often limited, heterogeneous, or noisy, DESeq2 and Limma are commonly preferred for their stability and capacity to handle heteroscedasticity. Recent studies illustrate these methodological differences in practical contexts. In a transcriptomic analysis of human neural progenitors exposed to low-grade environmental factors linked to autism, DESeq2 identified convergent upregulation of synaptic and lipid metabolism pathways, demonstrating its sensitivity in small-scale, exposure-based experimental designs [51]. edgeR, alongside DESeq2, was used in a study of patient-derived retinal organoids from individuals with optic nerve hypoplasia (ONH), where both methods revealed overlapping sets of differentially expressed genes enriched for known NDD-related pathways, supporting the genetic contribution to ONH [52]. Limma was applied in an integrated microarray analysis to construct an immune-related diagnostic model for ASD, identifying 41 differentially expressed genes and implicating specific immune cell populations such as neutrophils and CD8+ T cells [53]. These examples highlight how tool selection can influence both the sensitivity and interpretability of findings in NDD research.

In proteomics, classical statistical tests, such as Student's *t*-tests, ANOVA, and their non-parametric equivalents, are widely applied [17]. Univariate tests are straightforward to interpret and robust under multiple testing correction (e.g., Benjamini–Hochberg [40]). For example, Limma and Student's *t*-test have been used to identify differentially expressed proteins in extracellular vesicles from plasma of individuals with ASD, revealing five downregulated proteins, including CD40 and HSP27, with potential as diagnostic biomarkers [54]. One-way ANOVA and Tukey's post hoc tests were applied to quantify spatiotemporal protein expression changes in dopaminergic neurons across developmental stages, capturing molecular signatures relevant to neuronal maturation [55]. These examples illustrate how univariate frameworks remain central in proteomic studies of NDDs, particularly when combined with proper normalization and multiple testing correction.

Furthermore, in NDD studies, univariate analyses have successfully identified disease-associated molecular changes. For example, differential gene expression analyses in post-mortem ASD brain tissue revealed alterations in synaptic and immune-related pathways [8]. In Rett syndrome, univariate analyses of proteomic datasets highlighted altered abundances of proteins involved in chromatin remodeling and neuronal development, providing initial insights into molecular dysregulation [12]. However, univariate methods inherently overlook inter-feature dependencies and systems-level organization. Many molecular pathways and regulatory networks involve coordinated changes across multiple genes, proteins, or metabolites; univariate tests, by evaluating each feature in isolation, may miss these subtle, multicomponent alterations.

Multivariate approaches capture the joint behavior of multiple features, allowing the identification of covariance structures, latent variables, and correlated patterns that reflect systems-level organization. Dimensionality reduction techniques such as PCA [44], ICA [45], and PLS-DA [46] reduce dimensionality while highlighting major sources of variation. Sparse extensions, including sPLS-DA and sCCA [47,56,57], combine feature selection with dimensionality reduction, improving interpretability and reducing overfitting in small cohorts.

In NDD research, multivariate models have revealed coordinated dysregulation across multiple molecular layers. For instance, studies integrating transcriptomics and proteomics in ASD models identified simultaneous disruptions in synaptic proteins and mitochondrial pathways [58]. Moreover, integrative analyses combining metabolomics with transcriptomics uncovered immune-metabolic dysregulation mediated by transcription factors such as RARA and NFKB2 [59]. Penalized regression models, such as LASSO, Elastic Net,

and Priority-Elastic Net [41,60], further enhance predictive modeling by identifying informative features while controlling overfitting. These approaches have been successfully applied to multi-omics data from patient-derived cerebral organoids and iPSC-derived neuronal cultures, providing mechanistic insight into convergent molecular signatures across NDDs [58].

Hybrid approaches are increasingly adopted to combine the strengths of univariate and multivariate methods. In a union strategy, features identified by either approach are retained to maximize sensitivity, particularly useful when effect sizes are small or heterogeneous [61,62]. Conversely, intersection strategies, where only features detected by both univariate and multivariate models are selected, enhance specificity, reducing false positives and emphasizing robust, reproducible signals [63]. This integrative selection framework is particularly valuable in NDD research, where disease-associated molecular changes may be subtle, context-dependent, and spread across multiple omics layers. Additionally, combining feature-level statistics from univariate models with latent factors derived from multivariate models (e.g., using multivariate scores as covariates in univariate regression) can further refine associations and improve interpretability [64].

Overall, the choice between univariate and multivariate methods should be guided by the aim of research, data characteristics, and sample size. While univariate tests remain essential for detecting strong, individual-level effects, multivariate and penalized models are critical for uncovering complex, systems-level dysregulation typical of NDDs. In practice, combining both strategies offers a flexible and powerful analytic toolkit capable of adapting to the inherent complexity of multi-omics data in NDDs.

Table 1 provides a summary of the main statistical and integrative methods commonly applied in multi-omics analyses of neurodevelopmental disorders.

Table 1. Key statistical and integrative methods in multi-omics analysis of neurodevelopmental disorders.

Method	Omics Layer	Strengths	Limitations	Refs.
DESeq2	Transcriptomics	Shrinkage of fold changes; robust FDR control	Needs replicates; univariate only	[19]
edgeR	Transcriptomics	Negative binomial modeling; small sample sizes	Sensitive to outliers	[20,28]
Limma	Transcriptomics/ Proteomics	Linear modeling; empirical Bayes	Assumes log-normality	[28]
sPLS-DA	Multi-omics	Feature selection; dimensionality reduction	Sensitive to tuning; possible overfitting	[56]
sCCA	Multi-omics	Captures cross-dataset correlations	Needs regularization; computationally intensive	[47]
ComBat	Proteomics/ Transcriptomics	Corrects batch effects	Risk of overcorrection	[24,25]

Table 1. *Cont.*

Method	Omics Layer	Strengths	Limitations	Refs.
DIABLO	Multi-omics	Links features to phenotype	Requires normalized input; tuning complexity	[65]
SNF	Multi-omics	Patient stratification	Needs large cohorts; computational load	[66]
MOFA	Multi-omics	Handles missing data; interpretable latent factors	Hyperparameter sensitivity	[67]
Elastic Net/Priority-Elastic Net	Multi-omics	Feature selection; mitigates overfitting	Limited interpretability with correlated features	[41,60]

Note: columns show the method, the omics layer it applies to, main strengths, main limitations, and representative references.

4. Integrative Multi-Omics Approaches in Neurodevelopmental Disorders

Single-omics analyses, while informative, often fail to capture the full complexity of NDDs, where molecular perturbations span multiple regulatory layers. Integrative multi-omics approaches—combining transcriptomic, proteomic, metabolomic, epigenomic, and other datasets—provide a systems-level perspective that can reveal convergent disease mechanisms and identify candidate biomarkers [1,12,58,59,68]. By linking genetic variation to downstream molecular consequences, these approaches help bridge the gap between variant discovery and functional interpretation [7,8].

Integration strategies vary according to the timing and type of data combination. Early integration involves merging normalized datasets from multiple omics layers prior to statistical modeling. Techniques such as sparse canonical correlation analysis (sCCA) [47] and multi-block partial least squares (MB-PLS) [69] identify latent variables that capture covariation across data types, enabling the detection of molecular patterns invisible in single-layer analyses [46,65]. For instance, a recent work [70] demonstrates that a multi-block sPLS-DA (a discriminant version of MB-PLS with lasso penalization) can be effectively applied to multi-omics data integration. Building on this approach, they combined Polygenic Risk Scores, epigenomics and metabolomics to investigate ADHD. This integration revealed previously reported associations with ADHD, including those related to *MAD1L1* gene and glucocorticoid associations, while also highlighting *STAP2* as a possible novel gene involved in this neurodevelopmental condition.

Supervised frameworks, including DIABLO, further allow integration of omics profiles with clinical or phenotypic outcomes, facilitating the identification of molecular drivers associated with ASD, ID, or ADHD [65]. The main advantage of early integration lies in its ability to uncover coordinated molecular signals across data types, improving statistical power and biological interpretability when datasets are well matched and normalized. However, it requires rigorous preprocessing to harmonize data scales and distributions, and it can be computationally intensive. Additionally, early integration may be less robust in the presence of missing data or heterogeneous sample sets, which are common challenges in NDD research. In contrast, late integration approaches combine results from indepen-

dent analyses (e.g., intersecting differentially expressed genes with differentially abundant proteins) to highlight convergent pathways and regulatory networks [71]. Also, late integration approaches include pathway and gene set enrichment analyses that identify shared biological processes, network-based methods that integrate multi-omics data through protein–protein interaction or co-expression networks, and meta-dimensional approaches that combine independent omics results while preserving layer-specific information. Such methods allow for flexible, hypothesis-driven interpretation and can better accommodate heterogeneous datasets typical of NDD studies [72,73]. The strengths of late integration include its flexibility to incorporate diverse data types analyzed independently and its relative robustness to missing or unbalanced data. It also facilitates biological interpretation by focusing on convergent signals at the pathway or network level. However, this approach may overlook subtle, cross-omics interactions detectable only through joint modeling, and independent analyses may differ in statistical power or bias, complicating integration. By understanding the advantages and limitations of each strategy, researchers can select the most appropriate integration approach based on study design, data characteristics, and the specific hypotheses being tested.

Several studies illustrate the utility of these integrative strategies in NDD contexts. In a *Cntnap2* knockout mouse model and human cerebral organoids derived from individuals with ASD, combined transcriptomic and proteomic profiling revealed consistent alterations in synaptic function, mitochondrial activity, and axonal architecture, particularly in excitatory neurons [58]. Urine proteomics and metabolomics in children with ASD identified co-regulated changes in glutathione metabolism, xenobiotic detoxification, and immune-related pathways, suggesting non-invasive biomarkers of neuroinflammation [68]. Similarly, combined transcriptomics and metabolomics analyses demonstrated immune activation, synaptic dysfunction, and metabolic dysregulation in ASD, with transcription factors such as RARA and NFKB2 modulating these interconnected pathways [59].

A recent multi-omics study integrating metagenomics, metaproteomics, host proteomics, and metabolomics from fecal samples of children with ASD provided a comprehensive view of microbiome–host interactions. Using a customized bacterial protein database based on 16S rRNA sequencing, combined with robust normalization and statistical frameworks, this approach linked microbial and metabolic alterations to potential pathophysiological mechanisms in ASD, highlighting the value of integrative multi-omics in unraveling complex neurodevelopmental disorders [72]. The study underscored how integrating host and microbial layers can help disentangle the role of gut–brain axis dysfunctions in ASD, a crucial area given the emerging evidence of microbiota influence on neurodevelopment and behavior. Another study exemplified this by applying transcriptome-wide association modeling (FUSION), summary-based Mendelian randomization (SMR), and Bayesian colocalization (COLOC) to integrate m^6A-QTLs with GWAS, eQTL, and pQTL datasets, enabling multi-layered inference of regulatory mechanisms across neuropsychiatric disorders, including ASD [74].

Moreover, late multi-omics integration strategies are able to highlight mechanistic insights as demonstrated by recent studies. For instance, in [75] it is evident how late integration of bulk transcriptomics, proteomics and scRNA-seq data has deepened our knowledge of immune pathway alterations in ASD, revealing specific dysregulation in TRAIL, RANKL, and TWEAK pathways specifically in circulating NK cells and T cell subsets rather than general immune pathway dysfunction. Computational strategies are central to robust multi-omics integration. Penalized regression approaches, including Elastic Net and Priority-Elastic Net, are widely applied to high-dimensional datasets for feature selection and predictive modeling, mitigating overfitting while retaining biologically informative signals [41,76]. Similarity network fusion (SNF) aggregates patient similarity

matrices across omics layers, enabling molecular stratification of heterogeneous cohorts [66]. Network diffusion models, such as Markov Affinity-based Proteogenomic Signal Diffusion (MAPSD), propagate signals through protein–protein interaction networks, identifying brain-region-specific subnetworks and novel disease-associated genes [77]. Bayesian latent variable models, including Multi-Omics Factor Analysis (MOFA) and DIABLO, allow for missing data handling, uncover latent molecular factors, and provide interpretable links between omics modalities and phenotypes [65,67].

Despite these advances, challenges remain, including differences in data dimensionality, technical noise across platforms, and the need for interpretable models suited to limited sample sizes common in NDD cohorts. Additional hurdles include batch effects, heterogeneity in sample types (e.g., brain region, developmental stage), and the integration of longitudinal multi-omics data, which require sophisticated normalization and harmonization techniques to ensure robust inference. Collectively, these integrative approaches have begun to uncover shared molecular signatures in NDDs, including synaptic imbalance, mitochondrial dysfunction, and immune dysregulation, despite extensive genetic heterogeneity [58,59,68]. The success of multi-omics studies depends on careful cohort design, standardized preprocessing, and independent validation [9].

Furthermore, machine learning strategies have been increasingly applied to integrate heterogeneous datasets, linking clinical phenotypes with molecular profiles to identify robust biomarkers and disease subtypes. For example, combining detailed clinical data from the Autism Diagnostic Interview-Revised (ADI-R) with transcriptomic profiles via supervised and unsupervised machine learning methods has enabled the delineation of ASD subgroups characterized by distinct behavioral and gene expression patterns, thus illustrating the power of integrative computational approaches in neurodevelopmental disorders [78]. Deep learning frameworks and ensemble methods have also been explored to capture complex nonlinear relationships and interactions across omics layers, although their interpretability remains a key challenge that researchers are addressing through methods like SHAP values and attention mechanisms.

Indeed, future directions include the incorporation of spatially resolved and cell-type-specific omics, as well as machine learning strategies, to translate complex molecular patterns into mechanistic insights and potential biomarkers for precision medicine in neurodevelopmental disorders [38]. For instance, recent advances in single-cell multimodal omics technologies enabled simultaneous profiling of multiple molecular layers (e.g., genome, epigenome, transcriptome, proteome) within individual cells, offering unprecedented resolution to dissect cellular heterogeneity and gene regulatory networks in complex tissues such as the brain [79]. Such high-resolution approaches are particularly promising in disorders like ASD and Rett syndrome, where altered neurodevelopment may affect only specific cell types or circuits. These approaches facilitate a comprehensive understanding of cell-type-specific mechanisms underlying neurodevelopmental disorders and hold promises for identifying novel therapeutic targets. Moreover, emerging spatial transcriptomics and proteomics platforms are poised to provide critical insights into the spatial context of molecular changes, allowing researchers to map pathological signatures directly onto brain architecture, which is especially relevant given the region-specific pathology observed in many NDDs.

5. Future Directions and Translational Perspectives in Neurodevelopmental Disorders

Despite substantial progress in omics and integrative analyses, several challenges remain in translating multi-omics findings into clinical applications for NDDs. One major limitation is the genetic and phenotypic heterogeneity of NDD cohorts, which hinders the identification of reproducible biomarkers and therapeutic targets [80]. This variability extends not only across patient populations but also across tissues, developmental stages, and even experimental platforms. Future efforts will benefit from larger, deeply phenotyped cohorts combined with harmonized multi-omics pipelines, thereby improving statistical power and mechanistic insights [81]. Longitudinal study designs—where patients are followed across multiple developmental time points—are likely to be particularly powerful in capturing disease trajectories.

ASD exemplifies the complexity of NDDs, encompassing a wide array of genetic, molecular, and behavioral characteristics. Multi-omics approaches are increasingly elucidating the mechanistic underpinnings of ASD. For instance, recent studies have identified novel SHANK2 variants, highlighting disruptions in synaptic genes as contributors to ASD pathophysiology [82]. Additionally, research into neuroepitranscriptomics has revealed that RNA modifications such as m6A and m3C influence neuronal development and synaptic function, linking dysregulated RNA modification to neurodevelopmental deficits [83]. Furthermore, Granger Causality Analysis applied to functional imaging data has uncovered reduced connectivity in the medial prefrontal cortex and amygdala in children with ASD, associating neural network alterations with social cognition deficits and suggesting the potential of imaging biomarkers for diagnosis and longitudinal monitoring [84]. Moreover, AI-based literature mining systems have demonstrated the utility of computational approaches in integrating genomics, transcriptomics, and proteomics datasets, mapping the complex molecular landscape of ASD [85]. Collectively, these studies illustrate how multi-omics analyses, combined with computational and imaging tools, can capture the breadth of molecular and functional alterations in ASD.

Single-cell and spatially resolved omics further enhance our understanding by mapping dysregulation to specific cell types and brain regions, which is critical in ASD and Rett syndrome, where excitatory neurons, interneurons, and glial populations are differentially affected [31,86]. Integration of single-cell multimodal omics with spatial transcriptomics has been shown to reveal the complex cellular architecture and intercellular communication in the brain, further enabling the identification of cell-type-specific disease signatures and molecular interactions critical for neurodevelopmental disorders [87]. Such spatially resolved approaches complement bulk and single-cell data by preserving tissue context, which is essential for deciphering region-specific pathophysiology and guiding targeted interventions.

Coupling single-cell transcriptomics with spatial proteomics or metabolomics can uncover region-specific regulatory networks and post-transcriptional modifications that bulk analyses may obscure [88]. Functional validation of multi-omics findings using patient-derived models, such as induced pluripotent stem cells (iPSCs) and cerebral organoids, further illuminates convergent molecular phenotypes and allows testing of candidate therapeutic interventions [89–91]. Similarly, computational models leveraging machine learning and network-based methods can predict functional impacts of variants and prioritize targets for experimental validation [92]. Worthy of note, the network-based approach is progressing from single-layer network representations (taking into account single omics data at a time) to multilayer network architectures that integrate heterogeneous omics data types. This new framework enables a more comprehensive view of perturbed complex sys-

tems, particularly relevant for understanding the multifactorial nature of complex disorder such as NDDs [93].

Longitudinal and multi-modal data integration will be essential for capturing the dynamic nature of neurodevelopment. Static snapshots may miss critical temporal alterations in gene expression, protein abundance, or metabolite flux, whereas repeated sampling combined with integrative multi-omics can illuminate disease trajectories, identify early biomarkers, and inform precision medicine strategies [1,7,80]. Translating these discoveries into clinical practice requires standardized data sharing, rigorous validation across cohorts, and integration with electronic health records and imaging data. Regulatory frameworks for multi-omics-based diagnostics and therapeutics remain nascent, emphasizing the need for reproducibility, transparency, and robust computational pipelines. As the field advances, these strategies hold promises to bridge the gap from molecular discovery to clinically relevant precision interventions in NDDs, ultimately improving diagnosis, prognosis, and therapeutic outcomes.

6. Conclusions

Integrative multi-omics approaches are reshaping our understanding of neurodevelopmental disorders (NDDs), offering a systems-level view of the molecular alterations underlying conditions such as ASD, ADHD, and ID. By combining omics data, these strategies have revealed convergent pathways, including synaptic dysfunction, immune dysregulation, and mitochondrial impairment, despite extensive genetic heterogeneity.

Advanced computational methods, such as DIABLO, MOFA, and network-based models, are enabling the integration of high-dimensional, heterogeneous datasets, supporting molecular stratification and biomarker discovery. Notably, single-cell and spatially resolved multi-omics are allowing precise mapping of dysregulation to specific cell types and brain regions, as demonstrated in studies using cerebral organoids and scRNA-seq. Moreover, longitudinal designs and patient-derived models, including iPSCs, are beginning to capture dynamic disease trajectories and validate molecular findings in functional systems.

Despite remaining challenges, such as data harmonization, small cohort sizes, and the need for reproducibility, these integrative strategies are advancing the field toward clinically relevant applications. The convergence of multi-omics data with machine learning and imaging holds promise for precision diagnostics and targeted interventions, ultimately paving the way for personalized medicine in NDDs.

Author Contributions: Conceptualization, M.A. and M.F.; writing—original draft preparation, M.A., V.R. and M.F.; writing—review and editing, M.A., V.R. and M.F.; supervision, M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This review received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The content of this manuscript was developed by the authors based on literature searches conducted through PubMed. ChatGPT-5 was used exclusively for grammar and language revision and did not contribute to the generation of scientific content.

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

Abbreviations

The following abbreviations are used in this manuscript:

ADHD	attention-deficit/hyperactivity disorder
ADI-R	Autism Diagnostic Interview-Revised
ASD	autism spectrum disorder
DDA	data-dependent acquisition
FDR	false discovery rate
FWER	Family Wise Error Rate
ONH	optic nerve hypoplasia
ICA	independent component analysis
ID	intellectual disability
iPSCs	induced pluripotent stem cells
MAPSD	Markov Affinity-based Proteogenomic Signal Diffusion
MB-PLS	multi-block partial least squares
MNN	mutual nearest neighbors
MOFA	Multi-Omics Factor Analysis
NDD	neurodevelopmental disorders
PCA	principal component analysis
PLS-DA	partial least squares discriminant analysis
PTM	post-translational modification
QC	quality control
RUV	Remove Unwanted Variation
sCCA	sparse canonical correlation analysis
SMR	summary-based Mendelian randomization
SNF	similarity network fusion
SVA	surrogate variable analysis
TMM	trimmed mean of M values
t-SNE	t-distributed stochastic neighbor embedding
UMAP	Uniform Manifold Approximation and Projection
WGS	whole-genome sequencing
WES	whole-exome sequencing

References

1. Chui, M.M.-C.; Kwong, A.K.-Y.; Leung, H.Y.C.; Pang, C.; Scheller, I.F.; Wong, S.S.-N.; Fung, C.-W.; Yépez, V.A.; Gagneur, J.; Mak, C.C.-Y.; et al. An Outlier Approach: Advancing Diagnosis of Neurological Diseases through Integrating Proteomics into Multi-Omics Guided Exome Reanalysis. *npj Genom. Med.* **2025**, *10*, 36. [CrossRef]
2. Issac, A.; Halemani, K.; Shetty, A.; Thimmappa, L.; Vijay, V.R.; Koni, K.; Mishra, P.; Kapoor, V. The Global Prevalence of Autism Spectrum Disorder in Children: A Systematic Review and Meta-Analysis. *Osong Public Health Res. Perspect.* **2025**, *16*, 3–27. [CrossRef]
3. Nair, R.; Chen, M.; Dutt, A.S.; Hagopian, L.; Singh, A.; Du, M. Significant Regional Inequalities in the Prevalence of Intellectual Disability and Trends from 1990 to 2019: A Systematic Analysis of GBD 2019. *Epidemiol. Psychiatr. Sci.* **2022**, *31*, e91. [CrossRef] [PubMed]
4. Ayano, G.; Demelash, S.; Gizachew, Y.; Tsegay, L.; Alati, R. The Global Prevalence of Attention Deficit Hyperactivity Disorder in Children and Adolescents: An Umbrella Review of Meta-Analyses. *J. Affect. Disord.* **2023**, *339*, 860–866. [CrossRef] [PubMed]
5. Neul, J.L.; Kaufmann, W.E.; Glaze, D.G.; Christodoulou, J.; Clarke, A.J.; Bahi-Buisson, N.; Leonard, H.; Bailey, M.E.S.; Schanen, N.C.; Zappella, M.; et al. Rett Syndrome: Revised Diagnostic Criteria and Nomenclature. *Ann. Neurol.* **2010**, *68*, 944–950. [CrossRef] [PubMed]
6. Fehr, S.; Wilson, M.; Downs, J.; Williams, S.; Murgia, A.; Sartori, S.; Vecchi, M.; Ho, G.; Polli, R.; Psoni, S.; et al. The CDKL5 Disorder Is an Independent Clinical Entity Associated with Early-Onset Encephalopathy. *Eur. J. Hum. Genet.* **2013**, *21*, 266–273. [CrossRef]
7. Lan, X.; Tang, X.; Weng, W.; Xu, W.; Song, X.; Yang, Y.; Sun, H.; Ye, H.; Zhang, H.; Yu, G.; et al. Diagnostic Utility of Trio–Exome Sequencing for Children with Neurodevelopmental Disorders. *JAMA Netw. Open* **2025**, *8*, e251807. [CrossRef]

8. Tian, C.; Paskus, J.D.; Fingleton, E.; Roche, K.W.; Herring, B.E. Autism Spectrum Disorder/Intellectual Disability-Associated Mutations in Trio Disrupt Neuroligin 1-Mediated Synaptogenesis. *J. Neurosci.* **2021**, *41*, 7768–7778. [CrossRef]
9. Gilpatrick, T.; Lee, I.; Graham, J.E.; Raimondeau, E.; Bowen, R.; Heron, A.; Downs, B.; Sukumar, S.; Sedlazeck, F.J.; Timp, W. Targeted Nanopore Sequencing with Cas9-Guided Adapter Ligation. *Nat. Biotechnol.* **2020**, *38*, 433–438. [CrossRef]
10. Kaplanis, J.; Samocha, K.E.; Wiel, L.; Zhang, Z.; Arvai, K.J.; Eberhardt, R.Y.; Gallone, G.; Lelieveld, S.H.; Martin, H.C.; McRae, J.F.; et al. Evidence for 28 Genetic Disorders Discovered by Combining Healthcare and Research Data. *Nature* **2020**, *586*, 757–762. [CrossRef]
11. Heyne, H.O.; Singh, T.; Stamberger, H.; Abou Jamra, R.; Caglayan, H.; Craiu, D.; De Jonghe, P.; Guerrini, R.; Helbig, K.L.; Koeleman, B.P.C.; et al. De Novo Variants in Neurodevelopmental Disorders with Epilepsy. *Nat. Genet.* **2018**, *50*, 1048–1053. [CrossRef] [PubMed]
12. Murtaza, N.; Uy, J.; Singh, K.K. Emerging Proteomic Approaches to Identify the Underlying Pathophysiology of Neurodevelopmental and Neurodegenerative Disorders. *Mol. Autism* **2020**, *11*, 27. [CrossRef] [PubMed]
13. Deri, E.; Kumar Ojha, S.; Kartawy, M.; Khaliulin, I.; Amal, H. Multi-Omics Study Reveals Differential Expression and Phosphorylation of Autophagy-Related Proteins in Autism Spectrum Disorder. *Sci. Rep.* **2025**, *15*, 10878. [CrossRef] [PubMed]
14. Saffari, A.; Arno, M.; Nasser, E.; Ronald, A.; Wong, C.C.Y.; Schalkwyk, L.C.; Mill, J.; Dudbridge, F.; Meaburn, E.L. RNA Sequencing of Identical Twins Discordant for Autism Reveals Blood-Based Signatures Implicating Immune and Transcriptional Dysregulation. *Mol. Autism* **2019**, *10*, 38. [CrossRef]
15. Li, J.; Varghese, R.S.; Ressom, H.W. RNA-Seq Data Analysis. In *RNA Amplification and Analysis: Methods and Protocols*; Astatke, M., Ed.; Springer: New York, NY, USA, 2024; pp. 263–290. ISBN 978-1-0716-3918-4.
16. Tomaiuolo, P.; Piras, I.S.; Sain, S.B.; Picinelli, C.; Baccarin, M.; Castronovo, P.; Morelli, M.J.; Lazarevic, D.; Scattoni, M.L.; Tonon, G.; et al. RNA Sequencing of Blood from Sex- and Age-Matched Discordant Siblings Supports Immune and Transcriptional Dysregulation in Autism Spectrum Disorder. *Sci. Rep.* **2023**, *13*, 807. [CrossRef]
17. Lualdi, M.; Fasano, M. Statistical Analysis of Proteomics Data: A Review on Feature Selection. *J. Proteom.* **2019**, *198*, 18–26. [CrossRef]
18. Ioannidis, J.P.A. Why Most Published Research Findings Are False. *PLoS Med.* **2005**, *2*, e124. [CrossRef]
19. Love, M.I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
20. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef]
21. Bolstad, B.M.; Irizarry, R.A.; Åstrand, M.; Speed, T.P. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics* **2003**, *19*, 185–193. [CrossRef]
22. Välikangas, T.; Suomi, T.; Elo, L.L. A Systematic Evaluation of Normalization Methods in Quantitative Label-Free Proteomics. *Brief. Bioinform.* **2018**, *19*, 1–11. [CrossRef] [PubMed]
23. Molania, R.; Foroutan, M.; Gagnon-Bartsch, J.A.; Gandolfo, L.C.; Jain, A.; Sinha, A.; Olshansky, G.; Dobrovic, A.; Papenfuss, A.T.; Speed, T.P. Removing Unwanted Variation from Large-Scale RNA Sequencing Data with PRPS. *Nat. Biotechnol.* **2023**, *41*, 82–95. [CrossRef] [PubMed]
24. Yu, Y.; Zhang, N.; Mai, Y.; Ren, L.; Chen, Q.; Cao, Z.; Chen, Q.; Liu, Y.; Hou, W.; Yang, J.; et al. Correcting Batch Effects in Large-Scale Multiomics Studies Using a Reference-Material-Based Ratio Method. *Genome Biol.* **2023**, *24*, 201. [CrossRef] [PubMed]
25. Yu, Y.; Mai, Y.; Zheng, Y.; Shi, L. Assessing and Mitigating Batch Effects in Large-Scale Omics Studies. *Genome Biol.* **2024**, *25*, 254. [CrossRef]
26. Leek, J.T.; Storey, J.D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* **2007**, *3*, e161. [CrossRef]
27. Risso, D.; Ngai, J.; Speed, T.P.; Dudoit, S. Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples. *Nat. Biotechnol.* **2014**, *32*, 896–902. [CrossRef]
28. Law, C.W.; Alhamdoosh, M.; Su, S.; Dong, X.; Tian, L.; Smyth, G.K.; Ritchie, M.E. RNA-Seq Analysis Is Easy as 1-2-3 with Limma, Glimma and edgeR. *F1000Res* **2016**, *5*, 1408. [CrossRef]
29. Zhou, H.; Panwar, P.; Guo, B.; Hallinan, C.; Ghazanfar, S.; Hicks, S.C. Spatial Mutual Nearest Neighbors for Spatial Transcriptomics Data. *Bioinformatics* **2025**, *41*, btaf403. [CrossRef]
30. Hu, X.; Li, H.; Chen, M.; Qian, J.; Jiang, H. Reference-Informed Evaluation of Batch Correction for Single-Cell Omics Data with Overcorrection Awareness. *Commun. Biol.* **2025**, *8*, 521. [CrossRef]
31. Velmeshev, D.; Schirmer, L.; Jung, D.; Haeussler, M.; Perez, Y.; Mayer, S.; Bhaduri, A.; Goyal, N.; Rowitch, D.H.; Kriegstein, A.R. Single-Cell Genomics Identifies Cell Type–Specific Molecular Changes in Autism. *Science* **2019**, *364*, 685–689. [CrossRef]

32. Parikshak, N.N.; Luo, R.; Zhang, A.; Won, H.; Lowe, J.K.; Chandran, V.; Horvath, S.; Geschwind, D.H. Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell* **2013**, *155*, 1008–1021. [CrossRef] [PubMed]
33. Yang, X.; Zhang, J. Study Design, Sample Size Estimation, and Selection of Statistical Method. In *Textbook of Medical Statistics: For Medical Students*; Guo, X., Xue, F., Eds.; Springer Nature: Singapore, 2024; pp. 7–26. ISBN 978-981-99-7390-3.
34. Schober, P.; Vetter, T.R. Repeated Measures Designs and Analysis of Longitudinal Data: If at First You Do Not Succeed—Try, Try Again. *Anesth. Analg.* **2018**, *127*, 569. [CrossRef] [PubMed]
35. Chen, J.; Wang, Y.; Ko, J. Single-Cell and Spatially Resolved Omics: Advances and Limitations. *J. Pharm. Anal.* **2023**, *13*, 833–835. [CrossRef] [PubMed]
36. Allison, P.D. *Missing Data*; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2002; ISBN 978-1-4129-8507-9.
37. Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **2016**, *15*, 1116–1125. [CrossRef]
38. Crook, O.M.; Chung, C.; Deane, C.M. Challenges and Opportunities for Bayesian Statistics in Proteomics. *J. Proteome Res.* **2022**, *21*, 849–864. [CrossRef]
39. Webel, H.; Niu, L.; Nielsen, A.B.; Locard-Paulet, M.; Mann, M.; Jensen, L.J.; Rasmussen, S. Imputation of Label-Free Quantitative Mass Spectrometry-Based Proteomics Data Using Self-Supervised Deep Learning. *Nat. Commun.* **2024**, *15*, 5405. [CrossRef]
40. Hochberg, Y.; Benjamini, Y. More Powerful Procedures for Multiple Significance Testing. *Stat. Med.* **1990**, *9*, 811–818. [CrossRef]
41. Musib, L.; Coletti, R.; Lopes, M.B.; Mouriño, H.; Carrasquinha, E. Priority-Elastic Net for Binary Disease Outcome Prediction Based on Multi-Omics Data. *BioData Min.* **2024**, *17*, 45. [CrossRef]
42. Schweickart, A.; Chetnik, K.; Batra, R.; Kaddurah-Daouk, R.; Suhre, K.; Halama, A.; Krumsiek, J. AutoFocus: A Hierarchical Framework to Explore Multi-Omic Disease Associations Spanning Multiple Scales of Biomolecular Interaction. *Commun. Biol.* **2024**, *7*, 1094. [CrossRef]
43. Zhang, M.J.; Xia, F.; Zou, J. Fast and Covariate-Adaptive Method Amplifies Detection Power in Large-Scale Multiple Hypothesis Testing. *Nat. Commun.* **2019**, *10*, 3433. [CrossRef]
44. Ringnér, M. What Is Principal Component Analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304. [CrossRef]
45. Hyvärinen, A.; Oja, E. Independent Component Analysis: Algorithms and Applications. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef] [PubMed]
46. Saccenti, E.; Hoefsloot, H.C.J.; Smilde, A.K.; Westerhuis, J.A.; Hendriks, M.M.W.B. Reflections on Univariate and Multivariate Analysis of Metabolomics Data. *Metabolomics* **2014**, *10*, 361–374. [CrossRef]
47. Witten, D.M.; Tibshirani, R.J. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Stat. Appl. Genet. Mol. Biol.* **2009**, *8*, 28. [CrossRef]
48. ElKarami, B.; Alkhateeb, A.; Qattous, H.; Alshomali, L.; Shahrrava, B. Multi-Omics Data Integration Model Based on UMAP Embedding and Convolutional Neural Network. *Cancer Inform.* **2022**, *21*, 11769351221124205. [CrossRef] [PubMed]
49. Wu, Q.; Morrow, E.M.; Uzun, E.D.G. A Deep Learning Model for Prediction of Autism Status Using Whole-Exome Sequencing Data. *PLoS Comput. Biol.* **2024**, *20*, e1012468. [CrossRef]
50. Alqaysi, M.E.; Albahri, A.S.; Hamid, R.A. Evaluation and Benchmarking of Hybrid Machine Learning Models for Autism Spectrum Disorder Diagnosis Using a 2-Tuple Linguistic Neutrosophic Fuzzy Sets-Based Decision-Making Model. *Neural Comput. Appl.* **2024**, *36*, 18161–18200. [CrossRef]
51. Arora, A.; Becker, M.; Marques, C.; Oksanen, M.; Li, D.; Mastropasqua, F.; Watts, M.E.; Arora, M.; Falk, A.; Daub, C.O.; et al. Screening Autism-Associated Environmental Factors in Differentiating Human Neural Progenitors with Fractional Factorial Design-Based Transcriptomics. *Sci. Rep.* **2023**, *13*, 10519. [CrossRef]
52. Aparicio, J.G.; Hopp, H.; Harutyunyan, N.; Stewart, C.; Cobrinik, D.; Borchert, M. Aberrant Gene Expression yet Undiminished Retinal Ganglion Cell Genesis in iPSC-Derived Models of Optic Nerve Hypoplasia. *Ophthalmic Genet.* **2024**, *45*, 1–15. [CrossRef]
53. Cao, W.; Luo, C.; Fan, Z.; Lei, M.; Cheng, X.; Shi, Z.; Mao, F.; Xu, Q.; Fu, Z.; Zhang, Q. Analysis of Potential Biomarkers and Immune Infiltration in Autism Based on Bioinformatics Analysis. *Medicine* **2023**, *102*, e33340. [CrossRef]
54. Ali Moussa, H.Y.; Shin, K.C.; de la Fuente, A.; Bensmail, I.; Abdesselem, H.B.; Ponraj, J.; Mansour, S.; Al-Shaban, F.A.; Stanton, L.W.; Abdulla, S.A.; et al. Proteomics Analysis of Extracellular Vesicles for Biomarkers of Autism Spectrum Disorder. *Front. Mol. Biosci.* **2024**, *11*, 1467398. [CrossRef]
55. Zhao, H.; Chen, P.; Gao, X.; Huang, Z.; Yang, P.; Shen, H. Spatiotemporal Proteomic and Transcriptomic Landscape of DAT+ Dopaminergic Neurons Development and Function. *iScience* **2025**, *28*, 112115. [CrossRef]
56. Lê Cao, K.-A.; Boitard, S.; Besse, P. Sparse PLS Discriminant Analysis: Biologically Relevant Feature Selection and Graphical Displays for Multiclass Problems. *BMC Bioinform.* **2011**, *12*, 253. [CrossRef]
57. Witten, D.M.; Tibshirani, R.; Hastie, T. A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis. *Biostatistics* **2009**, *10*, 515–534. [CrossRef] [PubMed]

58. Jang, W.E.; Park, J.H.; Park, G.; Bang, G.; Na, C.H.; Kim, J.Y.; Kim, K.-Y.; Kim, K.P.; Shin, C.Y.; An, J.-Y.; et al. Cntnap2-Dependent Molecular Networks in Autism Spectrum Disorder Revealed through an Integrative Multi-Omics Analysis. *Mol. Psychiatry* **2023**, *28*, 810–821. [CrossRef] [PubMed]
59. Meng, Y.; Jia, J.; Ding, Y.; Wang, P.; Wang, Z.; Zhang, R.; He, Z.; Wang, Z.; Zhang, H.; Feng, L.; et al. Characterizing Immune and Metabolic Profiles in Autism Spectrum Disorder through Combined Transcriptomics-Metabonomics Analysis. *J. Psychiatr. Res.* **2025**, *190*, 92–101. [CrossRef] [PubMed]
60. Sokolov, A.; Carlin, D.E.; Paull, E.O.; Baertsch, R.; Stuart, J.M. Pathway-Based Genomics Prediction Using Generalized Elastic Net. *PLoS Comput. Biol.* **2016**, *12*, e1004790. [CrossRef]
61. Qureshi, F.; Adams, J.B.; Audhya, T.; Hahn, J. Multivariate Analysis of Metabolomic and Nutritional Profiles among Children with Autism Spectrum Disorder. *J. Pers. Med.* **2022**, *12*, 923. [CrossRef]
62. Remori, V.; Airoldi, M.; Alberio, T.; Fasano, M.; Azzi, L. Prediction of Oral Cancer Biomarkers by Salivary Proteomics Data. *Int. J. Mol. Sci.* **2024**, *25*, 11120. [CrossRef]
63. Zhang, J.; Ji, G.; Gao, X.; Guan, J. Single-Nucleus Gene and Gene Set Expression-Based Similarity Network Fusion Identifies Autism Molecular Subtypes. *BMC Bioinform.* **2023**, *24*, 142. [CrossRef]
64. Tang, X.; Feng, C.; Zhao, Y.; Zhang, H.; Gao, Y.; Cao, X.; Hong, Q.; Lin, J.; Zhuang, H.; Feng, Y.; et al. A Study of Genetic Heterogeneity in Autism Spectrum Disorders Based on Plasma Proteomic and Metabolomic Analysis: Multiomics Study of Autism Heterogeneity. *MedComm* **2023**, *4*, e380. [CrossRef]
65. Singh, A.; Shannon, C.P.; Gautier, B.; Rohart, F.; Vacher, M.; Tebbutt, S.J.; Lê Cao, K.-A. DIABLO: An Integrative Approach for Identifying Key Molecular Drivers from Multi-Omics Assays. *Bioinformatics* **2019**, *35*, 3055–3062. [CrossRef]
66. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat. Methods* **2014**, *11*, 333–337. [CrossRef] [PubMed]
67. Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J.C.; Buettner, F.; Huber, W.; Stegle, O. Multi-Omics Factor Analysis—A Framework for Unsupervised Integration of Multi-omics Data Sets. *Mol. Syst. Biol.* **2018**, *14*, e8124. [CrossRef] [PubMed]
68. Liu, W.; Li, L.; Xia, X.; Zhou, X.; Du, Y.; Yin, Z.; Wang, J. Integration of Urine Proteomic and Metabolomic Profiling Reveals Novel Insights Into Neuroinflammation in Autism Spectrum Disorder. *Front. Psychiatry* **2022**, *13*, 780747. [CrossRef] [PubMed]
69. Bougeard, S.; Dray, S. Supervised Multiblock Analysis in R with the Ade4 Package. *J. Stat. Softw.* **2018**, *86*, 1–17. [CrossRef]
70. Hubers, N.; Hagenbeek, F.A.; Pool, R.; Déjean, S.; Harms, A.C.; Roetman, P.J.; van Beijsterveldt, C.E.M.; Fanos, V.; Ehli, E.A.; Vermeiren, R.R.J.M.; et al. Integrative Multi-Omics Analysis of Genomic, Epigenomic, and Metabolomics Data Leads to New Insights for Attention-Deficit/Hyperactivity Disorder. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2024**, *195*, e32955. [CrossRef]
71. Huang, S.; Chaudhary, K.; Garmire, L.X. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front. Genet.* **2017**, *8*, 84. [CrossRef]
72. Osama, A.; Anwar, A.M.; Ezzeldin, S.; Ahmed, E.A.; Mahgoub, S.; Ibrahim, O.; Ibrahim, S.A.; Abdelhamid, I.A.; Bakry, U.; Diab, A.A.; et al. Integrative Multi-Omics Analysis of Autism Spectrum Disorder Reveals Unique Microbial Macromolecules Interactions. *J. Adv. Res.* **2025**, *S2090-1232*, 00055-4. [CrossRef]
73. Slobodyanyuk, M.; Bahcheli, A.T.; Klein, Z.P.; Bayati, M.; Strug, L.J.; Reimand, J. Directional Integration and Pathway Enrichment Analysis for Multi-Omics Data. *Nat. Commun.* **2024**, *15*, 5690. [CrossRef]
74. Liufu, C.; Luo, L.; Pang, T.; Zheng, H.; Yang, L.; Lu, L.; Chang, S. Integration of Multi-Omics Summary Data Reveals the Role of N6-Methyladenosine in Neuropsychiatric Disorders. *Mol. Psychiatry* **2024**, *29*, 3141–3150. [CrossRef]
75. Nour-Eldine, W.; Ltaief, S.M.; Ouararhni, K.; Abdul Manaph, N.P.; de la Fuente, A.; Bensmail, I.; Abdesselem, H.B.; Al-Shammari, A.R. A Multi-Omics Approach Reveals Dysregulated TNF-Related Signaling Pathways in Circulating NK and T Cell Subsets of Young Children with Autism. *Genes Immun.* **2025**, 1–13. [CrossRef] [PubMed]
76. Greenwood, C.J.; Youssef, G.J.; Letcher, P.; Macdonald, J.A.; Hagg, L.J.; Sanson, A.; Mcintosh, J.; Hutchinson, D.M.; Toumbourou, J.W.; Fuller-Tyszkiewicz, M.; et al. A Comparison of Penalised Regression Methods for Informing the Selection of Predictive Markers. *PLoS ONE* **2020**, *15*, e0242730. [CrossRef] [PubMed]
77. Torshizi, A.D.; Duan, J.; Wang, K. Cell-Type-Specific Proteogenomic Signal Diffusion for Integrating Multi-Omics Data Predicts Novel Schizophrenia Risk Genes. *Patters* **2020**, *1*, 100091. [CrossRef] [PubMed]
78. Yuwattana, W.; Saeliw, T.; van Erp, M.L.; Poolcharoen, C.; Kanlayaprasit, S.; Trairatvorakul, P.; Chonchaiya, W.; Hu, V.W.; Sarachana, T. Machine Learning of Clinical Phenotypes Facilitates Autism Screening and Identifies Novel Subgroups with Distinct Transcriptomic Profiles. *Sci. Rep.* **2025**, *15*, 11712. [CrossRef]
79. Zhu, C.; Preissl, S.; Ren, B. Single-Cell Multimodal Omics: The Power of Many. *Nat. Methods* **2020**, *17*, 11–14. [CrossRef]
80. Huynh, L.; Hormozdiari, F. Combinatorial Approach for Complex Disorder Prediction: Case Study of Neurodevelopmental Disorders. *Genetics* **2018**, *210*, 1483–1495. [CrossRef]

81. Litman, A.; Sauerwald, N.; Green Snyder, L.; Foss-Feig, J.; Park, C.Y.; Hao, Y.; Dinstein, I.; Theesfeld, C.L.; Troyanskaya, O.G. Decomposition of Phenotypic Heterogeneity in Autism Reveals Underlying Genetic Programs. *Nat. Genet.* **2025**, *57*, 1611–1619. [CrossRef]
82. Wu, Y.; Li, W.; Tan, B.; Luo, S. Identification of Novel SHANK2 Variants in Two Chinese Families via Exome and RNA Sequencing. *Front. Neurosci.* **2023**, *17*, 1275421. [CrossRef]
83. Lee, S.-M.; Koo, B.; Carré, C.; Fischer, A.; He, C.; Kumar, A.; Liu, K.; Meyer, K.D.; Ming, G.; Peng, J.; et al. Exploring the Brain Epitranscriptome: Perspectives from the NSAS Summit. *Front. Neurosci.* **2023**, *17*, 1291446. [CrossRef]
84. Deng, S.; Tan, S.; Guo, C.; Liu, Y.; Li, X. Impaired Effective Functional Connectivity in the Social Preference of Children with Autism Spectrum Disorder. *Front. Neurosci.* **2024**, *18*, 1391191. [CrossRef] [PubMed]
85. Mongad, D.; Subramanian, I.; Krishanpal, A. Deriving Comprehensive Literature Trends on Multi-Omics Analysis Studies in Autism Spectrum Disorder Using Literature Mining Pipeline. *Front. Neurosci.* **2024**, *18*, 1400412. [CrossRef] [PubMed]
86. Pascual-Alonso, A.; Xiol, C.; Smirnov, D.; Kopajtich, R.; Prokisch, H.; Armstrong, J. Identification of Molecular Signatures and Pathways Involved in Rett Syndrome Using a Multi-Omics Approach. *Hum. Genom.* **2023**, *17*, 85. [CrossRef] [PubMed]
87. Marx, V. Method of the Year: Spatially Resolved Transcriptomics. *Nat. Methods* **2021**, *18*, 9–14. [CrossRef]
88. Hu, T.; Allam, M.; Cai, S.; Henderson, W.; Yueh, B.; Garipcan, A.; Ievlev, A.V.; Afkarian, M.; Beyaz, S.; Coskun, A.F. Single-Cell Spatial Metabolomics with Cell-Type Specific Protein Profiling for Tissue Systems Biology. *Nat. Commun.* **2023**, *14*, 8260. [CrossRef]
89. Ha, D.; Kong, J.; Kim, D.; Lee, K.; Lee, J.; Park, M.; Ahn, H.; Oh, Y.; Kim, S. Development of Bioinformatics and Multi-Omics Analyses in Organoids. *BMB Rep.* **2023**, *56*, 43–48. [CrossRef]
90. Drakulic, D.; Djurovic, S.; Syed, Y.A.; Trattaro, S.; Caporale, N.; Falk, A.; Ofir, R.; Heine, V.M.; Chawner, S.J.R.A.; Rodriguez-Moreno, A.; et al. Copy Number Variants (CNVs): A Powerful Tool for iPSC-Based Modelling of ASD. *Mol. Autism* **2020**, *11*, 42. [CrossRef]
91. Sabitha, K.R.; Shetty, A.K.; Upadhya, D. Patient-Derived iPSC Modeling of Rare Neurodevelopmental Disorders: Molecular Pathophysiology and Prospective Therapies. *Neurosci. Biobehav. Rev.* **2021**, *121*, 201–219. [CrossRef]
92. Remori, V.; Bondi, H.; Airoldi, M.; Pavinato, L.; Borini, G.; Carli, D.; Brusco, A.; Fasano, M. A Systems Biology Approach for Prioritizing ASD Genes in Large or Noisy Datasets. *Int. J. Mol. Sci.* **2025**, *26*, 2078. [CrossRef]
93. De Domenico, M. More Is Different in Real-World Multilayer Networks. *Nat. Phys.* **2023**, *19*, 1247–1262. [CrossRef]