Briefings in Bioinformatics, 2026, 27, bbag276
https://doi.org/10.1093/bib/bbag276
Published: 4 June 2026
Review

OXFORD

Assessing current capabilities for incorporating lipidomics in multiomics data integration

Dylan H. Ross[1], Raghav Jain[1], Hyeyoon Kim[1], Javier E. Flores[1], Soumaydeep Sarkar[1], Chaevien S. Clendinen[2], Jennifer E. Kyle[1,*], Tao Liu[1,*], Sara J.C. Gosline[1,*]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, 902 Batelle Boulevard, Richland, WA 99354, United States
[2]Environmental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory, 902 Batelle Boulevard, Richland, WA 99354, United States
[3]Department of Biomedical Engineering, Oregon Health and Sciences University, 3303 SW Bond Ave, Portland, OR 97239, United States

*Corresponding authors. Sara J.C. Gosline, Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, United States. E-mail: sara.gosline@pnnl.gov; Tao Liu, Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, United States. E-mail: tao.liu@pnnl.gov; Jennifer E. Kyle, Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, United States. E-mail: jennifer.kyle@pnnl.gov

The research team is a group of postdoctoral researchers and senior scientists in the Biological Sciences Division at Pacific Northwest National Laboratory.

Abstract

Comprehensive analyses of multiple biological components including nucleic acids, proteins, metabolites, and lipids (i.e. "multiomics") provide unique insights into complex biological processes. Combining insights from these components through multiomics data integration enhances the depth and nuance of biological understanding available from these measurements. Among methods that integrate data across different technologies (e.g. mass spectrometry, sequencing), those that link components based on biological prior knowledge—pathway analyses—represent the most direct way of translating molecular-level observations into meaningful biological insights. However, significant barriers exist that prevent full utilization of metabolomics and especially lipidomics data in pathway integration. Challenges include the fast turnover and complex interactions of small molecules compared to biological macromolecules, low metabolite annotation rates, isomerism among lipids, and a lack of lipid representation in existing pathway knowledge bases. While these issues stem from a variety of causes, improvements to multiomics pathway integration including better incorporation of lipids into pathway knowledge bases and increased adoption of artificial intelligence approaches can greatly enhance the utility of small molecule -omics data, particularly for lipidomics. Here, we describe the current landscape of multiomics integration tools with an emphasis on support for metabolomics and lipidomics data, we highlight their capabilities and gaps with tangible examples using real multiomics data, and provide our perspective on how these approaches can be improved to better support generation of useful biological insights from complex multiomics data.

Keywords multi-omics integration, lipidomics, pathway analysis, artificial intelligence

Introduction

Diverse biotechnological advances have enabled the measurement of detailed cellular activity and discovery of numerous biological and biomedical findings. DNA sequencing platforms [1, 2] and analysis tools have rapidly accelerated the measurement of complete genomes, leading to the identification of specific cancer-driving mutations [3, 4] and genetic indicators of disease [5, 6]. Likewise, the measurement of complete transcriptomes has enabled the use of RNA as a proxy to protein-level expression, providing deeper insights across human disease (e.g. cancer [7, 8], Alzheimer's disease [9, 10]) to identify systematic alterations in protein signaling across diseases and tissues. Comprehensive mass spectrometric (MS) analyses have provided orthologous measurements of biomolecules including proteins, post-translational modifications such as phosphorylation sites, metabolites, and lipids, giving rise to additional atlases of cell-wide activity: the proteome [11], metabolome [12], and lipidome [13], respectively, which have provided numerous biological advances in cancer and metabolic disorders.

While each -ome on its own can produce useful findings, these individual results represent an incomplete picture of the biological system. Combining measurements of different components from a sample (i.e. multiomics) promises to provide a more complete and nuanced picture of biological processes and their interplay within a larger system (Fig. 1) [14]. Successful application of multiomics relies on advancements in computational tools [15–17] that integrate this multimodal data to identify new and/or complementary features of interest or patterns of expression among related samples, enabling cell-wide modeling of specific perturbations (Fig. 2). A subset of tools

Received: December 28, 2025. **Revised:** March 12, 2026. **Accepted:** May 4, 2026

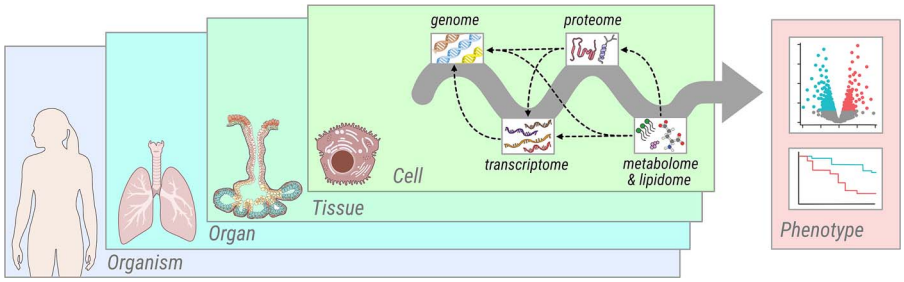

Figure 1 The multiomics data landscape. The solid line indicates the canonical transfer of biological information from genetic/non-genetic sources to phenotype. Dotted lines indicate the known interactions and feedback mechanisms between different biomolecules within a system running counter to the canonical direction.

of particular interest is those that focus on pathway-based integration: mapping molecules to annotated biological pathways based on prior knowledge to devise mechanistic hypotheses from -omics level data [18]. To date, however, these pathway-based tools have limited or no capability to fully capture and integrate small molecule data from lipidomics and metabolomics measurements [19–22]. This is largely due to gaps in existing data resources: pathway enrichment tools rely on knowledge of which molecules physically interact with each other to carry out a biological function; this data, with few exceptions (cite KEGG), is absent for small molecules.

Another major challenge of pathway analysis of small molecules lies in their inherent nature: lipids and metabolites are chemically diverse and lack a genetically encoded blueprint for identification, such as a known amino acid or nucleic acid sequence. This structural complexity results in largely incomplete database to support feature annotation [23] leading to a significant amount of metabolite "dark matter" (molecules not yet mapped) [24]. There is also the added difficulty that some metabolites fall on multiple pathways or have not yet been mapped to pathways (as is especially the case for lipids [25]). Despite these challenges, small biomolecule -omics provides a direct window into the dynamic processes of metabolism, including signaling and cellular regulation, that cannot be fully captured by proteogenomic data alone [26]. This noticeable gap in small molecule pathway integration comes at a time when the role of these molecules in disease has been highlighted in numerous studies [27–31].

In this review, we highlight the challenges of incorporating metabolomics and lipidomics data for multi-omic pathway integration. We (i) review of the current landscape of strategies and tools for integrating small biomolecule -omics data; (ii) describe the gaps and limitations among existing methods, chief among which is in the incorporation of small molecule data with other -omes using biological context using pathways; (iii) introduce a novel classification scheme for integration tools oriented around practical analysis objectives while grounding our discussion using tangible multi-omics data analyses examples, and lastly; (iv) share our perspective on future solutions to improve algorithms supporting biological interpretation, including the promise of artificial intelligence (AI) in this domain.

The landscape of multiomics data integration tools is diverse and supports a variety of analysis objectives

The general purpose of multiomics data integration is to combine observations of different biological components from a sample in order to build a more complete picture of its overall biological state. To achieve the end goal of improved biological understanding, we classify multiomics integration tools as supporting either sample-focused or feature-focused analysis objectives (Fig. 2). This categorization diverges from previous reviews [17, 32–35] that have classified multiomics integration methods based on the mechanics of the approach or algorithmic details (e.g. concatenation-based, transformation-based, and model-based). In this scheme, sample-focused methods demonstrate capabilities primarily aligned with sample-level analyses, such as fitting models predictive of a particular sample property (e.g. disease status) or identifying latent groupings of samples that might be tied to domain-relevant classifications (e.g. cancer subtypes). Broadly, these methods answer the question: "how are my samples different from one another?". Conversely, feature-focused methods are better suited for characterizing inter-feature relationships, such as through network or pathway analyses, answering the question: "what processes are changing in my samples beyond expected noise?". Both sample- and feature-focused methods can be matrixed further into labeled and unlabeled approaches, subcategories distinguished by whether samples are assigned labels reflecting a trait of interest (e.g. treatment or control) and those labels are taken into consideration during data analysis [36, 37]. While these categorizations are not mutually exclusive of one another, they serve as a useful guide for understanding the particular types of problems a tool is best suited to solve.

Sample-focused integration predicts biological states

The main objective of sample-focused integration methods is to determine multiomic features that can explain and predict differences between biological states that are associated with an external trait of interest (labeled data) or for grouping samples based on shared characteristics (unlabeled data). In either case, construction of these signatures tends to prioritize sparsity and predictive power over interpretability. These methods are applied when the objective is to robustly predict an external trait of interest or to assign a sample to a group or subtype using a small number of important features.

DIABLO [38, 39], JACA [40], DeepIMV [41], and CrossAttOmics [42] (Fig. 2, upper left) are computational methods designed to integrate labeled multi-omics datasets to uncover meaningful biological insights while supporting biomarker discovery and predictive modeling. All four approaches share the overarching goal of identifying cross-omics relationships, enabling the integration of diverse biological layers to reveal correlated features associated with a trait of interest. Though each employs distinct algorithmic frameworks—ranging from generalized dimension reduction to deep learning architectures—their

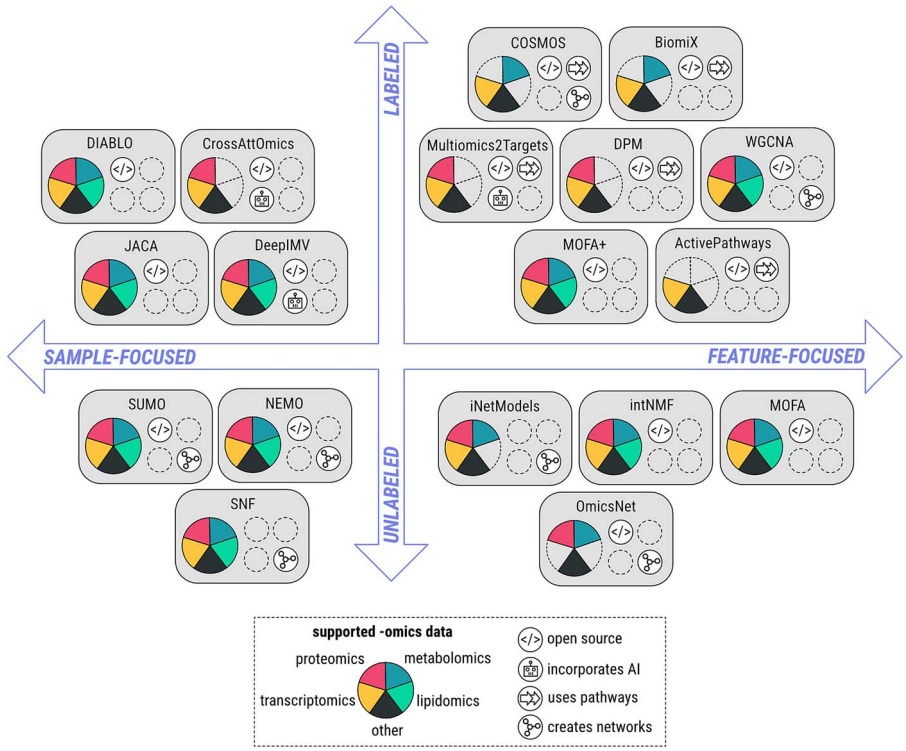

Figure 2 Categorization of selected multiomics data integration tools based on analysis objectives. The position of the tools indicates the primary analysis objectives that they support, according to a 2D categorization scheme that distinguishes sample- versus feature-focused analyses in addition to those that deal with labeled versus unlabeled data. The depiction of each tool reflects the -omics data types that they support, and additional highlighted characteristics of interest including whether the software is open source or incorporates AI in some capacity.

differences primarily lie in implementation focus: DIABLO emphasizes flexibility and interpretability for exploratory analyses, JACA provides a statistically unified framework for robust classification and association modeling, DeepIMV leverages nonlinear modeling for accommodating missing data and handling large-scale omics studies, and CrossAttOmics uses a cross-attention mechanism to capture interactions between components with known regulatory links. These methods offer a practical means of predicting traits of interest from multiomics data, and the compact multi-component signatures that form the basis of these predictions can additionally provide some insights into the systems that underly the predicted biological states.

Unlabeled multi-omics data integration methods focus on characterizing sample-level associations using trends in the high-dimensional data without relying on external labels or predefined traits. Similarity Network Fusion (SNF) [43], SUMO [44], and NEMO [45] exemplify this approach (Fig. 2, lower left), each aiming to uncover shared patterns and clustering structures across omics datasets. While all three methods integrate sample–sample similarity matrices to reveal cross-omics relationships, they differ in how they combine information. SNF iteratively reinforces consistent signals using a nonlinear combination method [46] to construct a single similarity network that can be clustered. SUMO leverages symmetric matrix factorization to learn shared low-dimensional representations for robust cross-omics clustering. Meanwhile, NEMO emphasizes local neighborhood relationships, using relative similarity scores and simple averaging for integration. Together, these methods offer the ability to perform exploratory analysis of multiomics data at the sample level without relying on labeled data, revealing groups of related samples and the features that drive those groupings.

Feature-focused integration contextualizes changes across components

Feature-focused multi-omics integration methods aim to place the molecular measurements in a biological context across samples by identifying shared patterns and sources of variation among features. This process may be guided by external traits or known sample groups (labeled data), or it can be used to find new relationships between samples (unlabeled data). These feature-focused analyses tend to prioritize contextualizing important features over sparsity or outright predictive power, and are therefore best suited to analysis objectives including mechanistic interpretation, hypothesis generation, or expansion of biological knowledge.

Feature-focused analyses are typically applied to labeled data, where the objective is to identify the biological processes that are altered with respect to some trait of interest. Integration tools in this category (Fig. 2, upper right) frequently leverage inter-feature associations (known relationships between molecules in the form of a network) and/or prior knowledge (list of molecules belonging to pathways) in order to broaden the biological context around important features, supporting deeper biological interpretation and hypothesis generation. Weighted Gene Correlation Network Analysis (WGCNA) [47] constructs feature-level networks based on pairwise correlations where the pairwise relationship between features is determined by correlation. As a result, WGCNA identifies highly similar groups of features, summarizing them with “eigengenes” to model their associations across datasets. Multiomics2Targets [48] packages together existing tools for performing enrichment analyses on phosphoproteomics (measurement of active phosphorylation site on proteins), proteomics, and transcriptomics data from the same

samples, and uses a large language model (LLM) to summarize and interpret analysis results into an output report, with a focus on application to studying human cancers. ActivePathways [49] combines statistical significance from multiple omics datasets to identify and prioritize enriched pathways, represented by lists of features, using a ranked hypergeometric test, highlighting pathways found uniquely through data integration for visualization and further exploration. Directional P-value Merging (DPM) [50] integrates statistical evidence across omics datasets by combining *P*-values while accounting for the expected direction of molecular changes, highlighting functionally coherent entities like genes or pathways. COSMOS [51] incorporates biological knowledge into network-level reasoning to causally connect deregulated elements, including transcription factors, metabolites, and kinases, within prior knowledge networks (e.g. known associations between molecules such as physical interactions), enabling contextualization of sample-specific changes. MOFA [52] and MOFA+ [53] generalize principal components analysis into a multi-omics framework, uncovering latent factors that capture shared and unique sources of variation across datasets, supporting applications such as trait associations (labeled data, MOFA+) and clustering (unlabeled data, MOFA). BiomiX [54] is a tool that integrates single-omics and multi-omics data using MOFA to capture sources of variation, optimize factor selection, perform pathway analysis, annotate factors with PubMed-assisted literature searches, and provide user-friendly interactive data visualization and flexible pipeline settings. This category is particularly rich with tools because understanding feature-level associations is a central analysis objective across many multiomics studies. The diversity of tools within this category reflects the broad range of solutions that have been developed to tackle a common problem, with each tool prioritizing different aspects of integration and interpretation to meet specific scientific needs and goals.

Feature-focused methods can also be applied to unlabeled data, where the general analysis objective is to derive cross-omics profiles of features that are characteristic of different groups of samples without knowing anything about the samples *a priori*. OmicsNet 2.0 [55] integrates multiomics data to create and refine biologically relevant networks, offering advanced visual analytics in 2D/3D, support for diverse omics types like SNPs, and enhanced reproducibility through incorporation of other R-based tools and collaborative sharing features. iNetModels [56] is a web-based platform that integrates tissue- and cancer-specific gene co-expression networks (GCNs) with personalized multi-omics biological networks (MOBNs), providing tools for interactive visualization, identifying functional relationships, uncovering biomarkers, and validating clinical findings. intNMF [57] is an integrative clustering method based on nonnegative matrix factorization (NMF) that classifies samples into distinct subtypes by integrating multiple molecular data types without relying on distributional assumptions, offering robust and flexible subtype discovery with user-defined weights. These methods are particularly useful for exploration of complex multiomics datasets, without the need for predefined sample labels or associated traits.

Example multiomic analysis highlights sample- and feature-focused insights and limitations

To compare the abilities of these tools to analyze and contextualize multiomics measurements that contain lipids and metabolites together with other molecules, we analyzed the same data using both a sample-focused tool, DIABLO, and a feature-focused tool, COSMOS. The data we used was comprised of transcriptomic, proteomic, phosphoproteomic, lipidomic, and metabolomic data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) acute myeloid leukemia (AML) study [58]. This study analyzed peripheral blood and/or bone marrow samples from 84 patients representing 2 classifications of AML maturation states [59], committed (n = 41) or primitive (n = 43), which have been shown to drive prognosis in most patients and exhibit a strong set of transcriptional markers that can be used to assign maturation state to unseen patient samples. In this study, we expanded upon our initial work by showing how multiomic integration tools can either (i) better classify samples by maturation state (using DIABLO as an exemplar of sample-based approaches) or (ii) identify features and pathways that determine maturation state (using COSMOS as our feature-based approach).

In DIABLO analysis, the expected strength of associations between different omics types is a user configurable model parameter, via a so-called "design matrix." For our analysis, we prepared a design matrix to connect the individual -omics data (transcripts, proteins, phosphosites, metabolites, and lipids) according to their observed covariances, as determined by an initial partial least squares (PLS) regression analysis. We performed automated parameter tuning with cross validation to determine the optimal number of components to use in the final DIABLO analysis and how many features to retain from each -omics type. Ultimately, a final model was trained using a single component consisting of 5–15 features from each -omics data type, which was capable of predicting committed versus primitive tumor subtype with an area under the receiver operating characteristic curve (ROC AUC) of >0.95. The impressive classification performance of the model using only a tiny subset (<0.09%) of the original multi-omics features is a testament to the power of sample-focused analyses to distill complex datasets down to a highly informative subset that is useful for robustly predicting biological states. A secondary benefit of this analysis is the feature-level insights, which can provide some clues as to the broader biological context underlying the predicted biological states. Figure 3A depicts the selected multiomic features (columns) and their normalized abundances from the samples (rows) colored according to primitive or committed tumor classification. From these features, we see a highly correlated abundance profile across the mRNA, protein, metabolite, and phosphosite features selected, where these features are up-regulated in the committed samples, while the lipids selected are anti-correlated and down-regulated in the same patients (Fig. 3A). This pattern is further exemplified by the correlation pattern across features depicted in the circos plot in Fig. 3B, where mRNA, proteins, and phosphosites are highly correlated with each other and lipids are anti-correlated. Interestingly, we fail to identify any significant correlation between the metabolomic features and other types of features, reinforcing their scattered abundance profiles observed in the heatmap. The sparsity of this composite signature makes it difficult to discern any higher-level insights into the biological systems that are altered between tumor classifications, requiring a feature-by-feature search of the results. In doing so, one can find evidence of disease relevance in diacylglycerophosphocholines (PCs) containing arachidonic acid (FA 20:4) suggesting an alteration in the outer membrane fluidity and/or an increase in fatty acid precursors with immune-related functions, and ITGB2, which has been found as a transcript that correlates with patient outcome in AML [60].

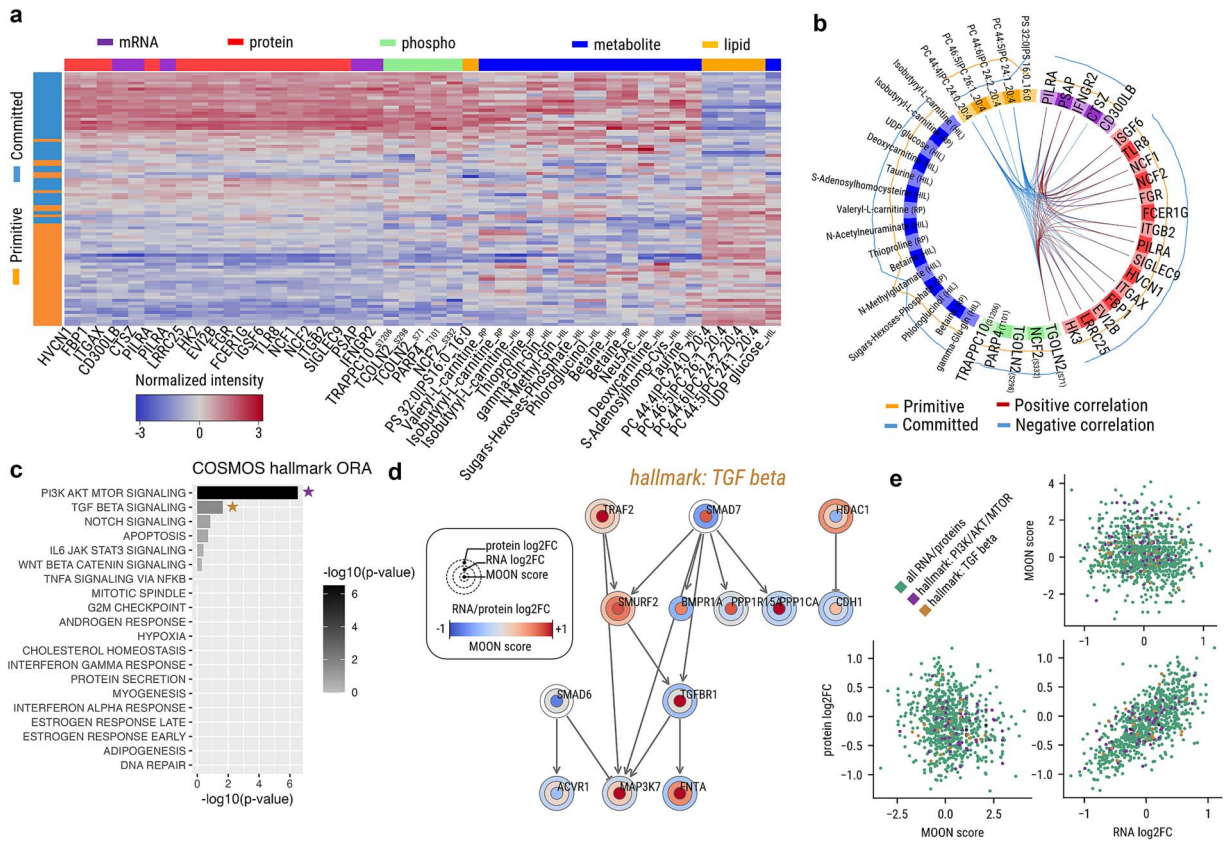

Figure 3 Results from DIABLO and COSMOS analysis of example multiomics dataset. (a) Heatmap showing normalized abundances of features selected for training the DIABLO model. The features (columns) are colored according to the -omics dataset they come from and the samples (rows) are colored according to their corresponding tumor classification (primitive or committed). (b) Circos plot showing correlation between features selected for training DIABLO model. Each group of features is colored according to the -omics dataset they come from. The lines along the outer perimeter indicate average relative abundance in primitive and committed tumor classification groups, and interior lines indicate features with a high degree of positive or negative correlation. (c) Results from hallmark ORA performed using genes included in the MOON analysis result network. The top two identified hallmark pathways with *P*-values <.05 are highlighted with stars. (d) Visualization of subsets of the MOON analysis result network from the highlighted pathway: TGF beta signaling. Round nodes represent proteins, which may have corresponding transcriptomics/proteomics log2FC values and an associated MOON score, which is mapped to the centermost circle. Log2FCs from transcriptomics and proteomics datasets, where available, are mapped to outer concentric rings. All log2FC and MOON scores are mapped to the same color scale as indicated by inset legend. Graph edges depict activation and inhibition relationships between nodes as specified in the legend. (e) Scatter plots comparing transcriptomic/proteomic log2FC versus MOON score from complete filtered PKN. Colors indicate subset of features, as described by legend inset.

To compare these results to a more feature-focused tool, we analyzed transcriptomic, proteomic, and metabolomic data from the same AML cohort using COSMOS (Fig. 3C–E). COSMOS does not support lipidomics data so they were not included in this analysis. We employed the metafootprint (MOON) function in COSMOS that infers transcription factor activities based on transcript abundances, maps these inferred activities onto a prior knowledge network (PKN) along with transcriptomic and metabolomic inputs, and then filters and scores based on coherence between information in the PKN, inferred activities, and observed abundances. Ultimately, the MOON analysis produces a filtered subset of the PKN, which is represented by node attributes and signed interactions. Among these node attributes is a MOON score, which reflects consistency between the sign of interactions in the PKN, inferred activities, and observed abundances both up- and downstream of a given node with the sign reflecting regulation in the up (+) or down (−) direction. We used this filtered subset of the PKN to perform overrepresentation analysis (ORA) for hallmark pathway genes [61] against a human proteome background with BioMart [62], which highlighted two significantly enriched pathways: PI3K AKT MTOR signaling and TGF beta signaling (Fig. 3C). The PI3K/MTOR pathway activity confirms the findings from the original study and is well known to be aberrantly activated in AML [63]. TGF beta signaling also has promising biological relevance as it has been shown to be associated with greater degrees of AML progression [64]. However, when we examine the feature-specific results (Fig. 3D), we see that the MOON scores do not always agree well with the observed protein/transcript abundances—sometimes disagreeing entirely—drawing into question whether the MOON analysis may be obfuscating, rather than summarizing, what is happening with pathways at the individual protein level. Indeed, bulk comparison of MOON scores against RNA or protein Log2 fold-changes showed no discernable correlations despite the apparent correlation between RNA and protein fold-changes (Fig. 3E). While most models assume that there is some noise in the high-throughput measurements, we expect systematic results to generally align with the -omics readouts. We additionally note that there are no metabolites present in the highlighted pathway; therefore, the metabolomics data were underutilized in this example analysis. This demonstration produced high-level insights into biological processes that may be altered between primitive and committed AML, but understanding

how the system changes between these states at a more granular level would require further analyses.

By comparing examples of the sample- and feature-focused categories of multiomics integration tools, we can see the distinct types of insights they provide. DIABLO identifies the minimum features needed to confidently classify samples as "Primitive" or "Committed" using each multiomics data type. While these features offer hints about higher-order processes altered between sample groups, their sparsity limits the ability to interpret biological significance solely from this analysis. In contrast, the feature-focused COSMOS analysis prioritized biological interpretation by contextualizing interactions between features within a framework of prior knowledge, facilitating hypothesis generation. This reliance on prior knowledge, however, also creates the potential for overlooking important features that are not yet represented in the knowledge base. We also found that the MOON analysis has the potential to obfuscate the overall up- or down-regulation of a pathway of interest. Together, these examples highlight the complementary strengths and inherent limitations of sample- and feature-focused multiomics integration tools, underscoring the importance of aligning tool selection with data analysis objectives.

Pathway analysis and networks apply statistical rigor to biological hypotheses

An important capability among -omics integration tools is aligning features to known biological networks or pathways, collectively termed functional analyses. These functional analyses represent a critical link between statistical observations made at the level of the omics dataset and meaningful biological hypotheses based on mechanistic understanding of the biological system [18, 22, 65]. As such, there are a wealth of databases available for mapping unknown features such as KEGG [66], The Gene Ontology (GO) [67], the molecular signatures database (MSIGDB) [61], LIPID MAPS [68], SwissLipids [69], and the human metabolome database (HMDB) [70], as well as tools that can leverage these databases to identify pathways such as Gene Set Enrichment Analysis (GSEA) [71], enrichR [72], WebGestalt [73], and leapR [74]. Each of these databases highlight different biological activity—for example, KEGG is uniquely able to integrate metabolites with protein/gene-based markers, while the GO maps genes to subcellular compartments (among other things). Each tool also provides unique analysis such as rank-based enrichment (GSEA) or pairwise correlation analysis (leapR). Traditionally, proteogenomics datasets benefit from extensive prior knowledge databases, with defined IDs and known feature characteristics. As a result, current pathway analysis tools have been successful at helping build meaningful connections between features, facilitating new hypothesis generation. However, biological hypotheses are highly dependent upon the database used for the enrichment analysis, and the lack of a single database that can map all molecules (e.g. genes, proteins, metabolites, and especially lipids) is a large gap for multiomics analysis.

Artificial intelligence is gaining traction in multiomics data integration

It is impossible to discuss multiomic integration without discussing how AI has impacted the field. Like most fields, integration of multiomic measurements has benefited from the recent breakthroughs in the usability of AI-based algorithms [75, 76]. One notable use of AI-based approaches is the integration of variational autoencoders as a method of reducing the dimensionality of complex multiomic data [77–79]. These tools provide alternatives to linear approaches such as principal component analysis or nonnegative matrix factorization. Additional AI advances have come from the incorporation of LLMs [80], which have been used to augment multiomics integration using clinical data [81] or information extracted from literature [82]. Lastly, foundation models built from libraries of single-cell or spatially resolved omics data enabled improved annotation of functional relationships between gene-based measurements [83, 84]. However, there are very limited applications of these approaches to small molecules other than basic chemical property prediction by tools such as ChemBERTa [85]. While newer AI-driven techniques show promise in expanding capabilities in multiomics data integration, their application is hindered by challenges such as limited explainability, reproducibility, and the absence of clear ground truth in multiomics data, making robust and reliable solutions difficult to achieve at present [86, 87]. There are also limitations with respect to the analysis objectives that are supported by AI-based tools, with most current examples performing predictions of external traits or assignment of subtypes (i.e. sample-focused analyses). Lastly, since AI/ML tools rely on the same pathway databases as other tools, they are also unable to use multiomic data to its full potential due to the lack of databases that accurately link all molecules, particularly lipids.

Paucity of lipidomics support is a trend across all pathway tools

Across the examples we showed above, as well as the summary depicted in Fig. 2, there is a noticeable gap in the coverage for small molecular measurement for feature-focused analyses. While sample-focused tools such as DIABLO are able to identify metabolites and lipids of interest (Fig. 3A and B), the heavy reliance of feature-focused methods such as COSMOS on prior knowledge for connecting components from different -omics data in a meaningful way, together with the known difficulties in connecting lipids with other biological components makes it incredibly challenging to put these molecules in biological context [25]. This lack of support for integrating lipidomic data with other -omics data types is self-reinforcing, as this type of integration is an important mechanism for expanding and refining our understanding of how lipids interact with other components in biological systems, which in turn increases our ability to interpret and use lipidomic data both alone and in a multiomic context.

Summary and perspective

Metabolomic and lipidomic data integration can improve biological understanding

There is a significant gap in existing knowledge bases that support multiomic pathway analyses, significantly limiting the degree to which lipids can be meaningfully connected with other components in a coherent biological context. This gap is attributable to a variety of factors including the high degree of isomerism among lipid species measured in lipidomics, and a lack of clean mapping for complex lipid species onto existing cellular pathways. Ultimately, it is our view that functional integration of lipidomics data represents an area of profound opportunity within the broader field of multiomics for enhancing the depth and complexity of insights available from these studies.

To further explore how the paucity of lipidomics data integration capabilities at the feature level can limit the ability to provide biological context to multiomic data, we performed an additional feature-focused analysis using the same AML cohort data from above. Specifically, we calculated the fold change between primitive and committed AML tumor classifications across all transcripts, proteins, metabolites, and lipids and then manually mapped the features to the "glycerolipids and glycerophospholipids" pathway from WikiPathways [46] (https://www.wikipathways.org/pathways/WP4722) (Fig. 4A). This pathway provides a high-level view of the synthesis and interconversion of major lipid classes and is among the most complete of the few existing reference pathways that meaningfully incorporates lipids with metabolites and proteins. The results are difficult to interpret: data from different modalities (e.g. transcriptomics versus proteomics in protein nodes) or measurement conditions (e.g. HILIC versus RP in metabolite nodes) are not always concurrent. Individual lipid species are typically mapped to larger LIPID MAPS subclasses [88], where there can be tens to hundreds of individual lipid species that can map to a given node (Fig. 4B–H), making meaningful interpretation even more difficult. Nevertheless, looking at the individual lipids within these groups, an interesting pattern does emerge: phosphatidylcholines (PCs, Fig. 4H), phosphatidylethanolamines (PEs, Fig. 4E), and phosphatidylglycerols (PGs, Fig. 4F) containing arachidonic acid (FA 20:4) are increased among primitive samples, whereas triacylglycerols (TGs, Fig. 4B), PEs (Fig. 4E), and lysophosphatidylethanolamines (LPEs, Fig. 4G) that contain polyunsaturated acids (PUFAs, e.g. FA 22:6) are increased among committed samples. While lipid-pathway annotation is necessary for lipidomic integration, this example shows that it is not sufficient: additional tools to interpret the different lipid subspecies and visualize them automatically (this figure was custom generated) are also lacking.

The promise of network models, artificial intelligence, and pathway data curation

The challenges of multiomic data integration with lipidomics data stem from multiple factors including analytical capabilities, informatics pipelines for metabolite and lipid annotation from raw MS data, and the underlying libraries/database, statistics, and algorithms. While improving each of these aspects will improve overall small molecule interpretation, the multiomic tools that incorporate small molecules (Fig. 2) are currently insufficient. Therefore, an improvement in multiomics data integration tools to meaningfully incorporate lipids is the most impactful way to enable biological findings from these data.

One approach to improving the integration of multiomic data lies in biological networks [16, 89, 90]. Networks (or graphs) are a natural way to structure complex data made up of interactions (edges) between components (nodes). Heterogenous networks, or networks that contain multiple different node and/or edge types, are particularly well-suited for representing complex multiomics data without the information loss that inherently occurs when multiomics data is coerced into a uniform representation. Structuring multiomics data within a heterogenous network better supports the interpretation of multiomics data, both qualitatively through visualization methods and quantitatively through statistical methods and network-specific algorithms.

AI is another approach that could improve the integration of small molecule data. One major constraint we found in our example analysis above was the expectation of linear relationships between groups of functional similarly molecules. For example, enrichment statistics (Fig. 3B) generally seek out groups of molecules in a pathway that are differentially expressed in the same direction and with similar scales, implicating the pathway in the biological phenomenon measured. Deep learning algorithms, due to their founding in artificial neural networks, are easily able to identify nonlinear and indirect relationships between molecules [91]. Nonlinear modeling approaches could help capture the nuances of small molecule data in a combinatorial nature, such as the isomeric potential of metabolite measurements, and account for the various aspects of lipid structure, such as fatty acid composition, degree of unsaturation, or class-specific characteristics. Through its capacity to uncover hidden patterns that are difficult to detect using traditional methods, AI could provide insights into metabolite and lipid biology that significantly enhance functional analyses.

Lastly, both network integration and AI improvements will suffer the same fate as previous methods without improved molecular annotation and expanded knowledge bases. For example, most network modeling algorithms (e.g. Tuncbag *et al.* [92]) rely on often hand-curated relationship data between molecules, such as protein–protein interaction networks [93–95] or metabolic pathway information [66, 70]. AI, also, requires large training datasets upon which they can identify relationships in the data. As such, careful integration of lipidomic data to existing pathway models would dramatically help these tools move forward.

Clinical and translational implications of improved lipidomics data integration

Recent studies demonstrate the tangible clinical impact of multiomics integration across diverse disease contexts. A proteogenomic survey of 242 ovarian tumors identified a panel of 1082 proteins that was able to predict response in patients high-grade serous ovarian cancer [96]. In a prospective precision oncology trial, the Tumor Profiler (TuPro) project applied nine independent -omics technologies to melanoma patient samples, generating comprehensive molecular profiles that informed treatment recommendations and achieved a 38% objective response rate in difficult-to-treat patients receiving individualized, multibiomarker-driven therapies over the standard of care [97].

Beyond direct treatment guidance, multiomics approaches have proven valuable for biomarker discovery and mechanistic insight, which could then inform drug development. Multiomics profiling of tumor-associated macrophages in esophageal squamous cell carcinoma uncovered a cholesterol-mediated immunosuppressive mechanism, identifying PLD3 expression as both a prognostic biomarker and potential therapeutic target [98]. Similarly, integrative analysis in HER2+ breast cancer revealed that the CNS-enriched metabolite N-acetylaspartate enables tumor immune evasion, suggesting its synthetic enzyme NAT8L as a target to enhance immunotherapy efficacy [99]. In Alzheimer's disease, multiomics integration identified lysophosphatidylethanolamine (LPE) species, particularly LPE 22:6, as strongly associated with disease-relevant protein modules, positioning these lipids as both potential therapeutic targets and candidates for dietary supplementation strategies [100]. Notably, several of these clinically impactful insights either implicate specific lipid species directly or reveal mechanistic roles for lipid metabolism more broadly, even in cases where comprehensive lipidomics was not a primary focus of the investigation. This pattern suggests that lipid biology

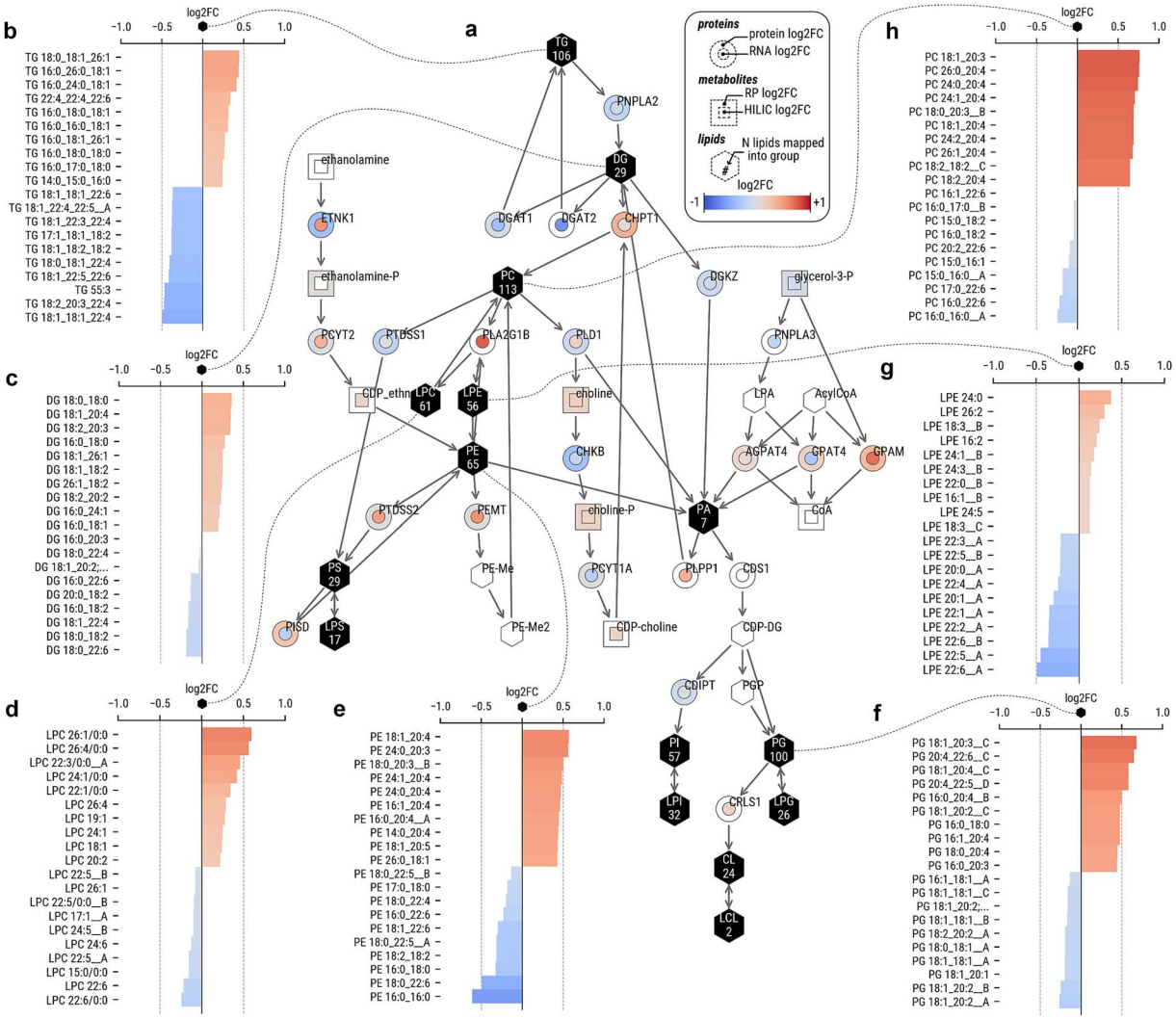

Figure 4 Manual multiomic mapping of protein, transcript, metabolite, and lipid measurements to a single pathway illustrates inherent complexities of small molecule pathway analysis. Log_2 fold-change (log2FC) values calculated between the "Primitive" and "Committed" AML tumor classifications across transcriptomics, proteomics, metabolomics, and lipidomics measurements mapped to the "glycerolipids and glycerophospholipids" pathway from WikiPathways (https://www.wikipathways.org/pathways/WP4722). (a) Network visualization of pathway depicts round nodes as proteins, which may have corresponding transcript (outer ring) and/or protein (inner ring) log2FC values, square nodes represent metabolite log2FC values with corresponding HILIC measurement (inner circle), and/or RP measurement (outer ring), hexagon nodes represent lipid class measurements with counts of individual lipid species that map to each node included when available, and edges between all nodes represent directional relationships in the metabolic pathway (e.g. substrate to enzyme or enzyme to product). Log2FCs for individual lipid species representing the top 20 up- and down-regulated members of a lipid class are plotted around the margins of the network visualization for selected lipid classes: (b) triacylglycerols, TGs; (c) diacylglycerols, DGs; (d) lysophosphatidylcholines, LPCs; (e) phosphatidylethanolamines, PEs; (f) phosphatidylglycerols, PGs; (g) lysophosphatidylethanolamines, LPEs; and (h) phosphatidylcholines, PCs.

may be an underexplored contributor to disease mechanisms across diverse pathologies. As integration tools improve their capacity to systematically incorporate lipidomics data alongside other -omics layers, we anticipate that the landscape of clinically actionable biomarkers and therapeutic targets will expand substantially.

Conclusions

In this work, we highlight the broad diversity of tools available for multiomics data integration and examine systematic gaps in tools that handle small molecules, particularly lipids. We identify numerous reasons for these gaps and propose potential solutions through graphical models, AI data curation, and improved pathway mapping. We conclude that better incorporation of lipids into pathway analyses is an essential next step in multiomics data integration tool development.

Key Points

- Multiomics analysis tools produce complex and complementary insights into biological processes.
- Multiomics data integration, especially pathway integration, is critical for translating complex multiomics data into meaningful biological insights.

- Different strategies for multiomics data integration exist that provide insights at the sample- or feature-level, and are best suited to different analysis objectives.
- Integration of small molecule -omics data faces unique barriers including isomerism among lipids and the lack of lipid representation in existing pathways.
- Promising solutions to these barriers include new data models for centralizing multimodal data, curation of reference pathways that include lipids, and application of artificial intelligence for predicting and scoring pathway dysregulation.

Acknowledgements

This work was supported by grants U24CA271012 and U24CA210955 (to T.L.) from the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC). Portions of the MS-based multiomics work described herein were performed at the Environmental Molecular Sciences Laboratory (grid.436923.9), a US Department of Energy (DOE) National Scientific User Facility located at the Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is a multiprogram national laboratory operated by the Battelle Memorial Institute for the DOE under contract DE-AC05-76RL01830.

Author contributions

Dylan H. Ross (Conceptualization, Formal analysis, Visualization, Supervision, Writing—original draft, Writing—review & editing), Raghav Jain (Data curation, Writing—original draft, Writing—review & editing), Hyeyoon Kim (Formal analysis, Visualization, Writing—original draft, Writing—review & editing), Javier E. Flores (Writing—original draft, Writing—review & editing), Soumaydeep Sarkar (Visualization, Writing—original draft, Writing—review & editing), Chaevien S. Clendinen (Writing—original draft, Writing—review & editing), Jennifer E. Kyle (Conceptualization, Supervision, Writing—original draft, Writing—review & editing), Tao Liu (Conceptualization, Funding acquisition, Writing—review & editing), and Sara J.C. Gosline (Conceptualization, Supervision, Writing—original draft, Writing—review & editing)

Conflict of interest

The authors declare no competing interests.

Funding

None declared.

Data availability

Harmonized genomic, transcriptomic, proteomics, metabolomics, and lipidomics data files generated for the CPTAC AML cohort can be accessed via Genomic Data Commons (GDC) at: https://portal.gdc.cancer.gov.

References

1. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J Appl Genet* 2011;**52**:413–35. https://doi.org/10.1007/s13353-011-0057-x
2. McCombie WR, McPherson JD, Mardis ER. Next-generation sequencing technologies. *Cold Spring Harb Perspect Med* 2019;**9**. https://doi.org/10.1101/cshperspect.a036798
3. Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci* 2018;**109**:513–22. https://doi.org/10.1111/cas.13505
4. Xiao C, Chen Z, Chen W *et al.* Personalized genome assembly for accurate cancer somatic mutation discovery using tumor-normal paired reference samples. *Genome Biol* 2022;**23**:237. https://doi.org/10.1186/s13059-022-02803-x
5. Tilemis FN, Marinakis NM, Veltra D *et al.* Germline CNV detection through whole-exome sequencing (WES) data analysis enhances resolution of rare genetic diseases. *Genes (Basel)* 2023;**14**:1490. https://doi.org/10.3390/genes14071490
6. Schobers G, Derks R, den Ouden A *et al.* Genome sequencing as a generic diagnostic strategy for rare disease. *Genome Med* 2024;**16**:32. https://doi.org/10.1186/s13073-024-01301-y
7. Cancer Genome Atlas Research, N, Weinstein JN, Collisson EA *et al.* The cancer genome atlas Pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20. https://doi.org/10.1038/ng.2764
8. Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* 2020;**578**:82–93. https://doi.org/10.1038/s41586-020-1969-6
9. Aguzzoli Heberle B, Fox KL, Lobraico Libermann L *et al.* Systematic review and meta-analysis of bulk RNAseq studies in human Alzheimer's disease brain tissue. *Alzheimers Dement* 2025;**21**:e70025. https://doi.org/10.1002/alz.70025
10. Neff RA, Wang M, Vatansever S *et al.* Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Sci Adv* 2021;**7**. https://doi.org/10.1126/sciadv.abb5398
11. Mani DR, Krug K, Zhang B *et al.* Cancer proteogenomics: current impact and future prospects. *Nat Rev Cancer* 2022;**22**:298–313. https://doi.org/10.1038/s41568-022-00446-5
12. Lin C, Tian Q, Guo S *et al.* Metabolomics for clinical biomarker discovery and therapeutic target identification. *Molecules* 2024;**29**:2198. https://doi.org/10.3390/molecules29102198
13. Wu JH, Zhang WP, Ouyang Z. Marching toward human lipidome project - advancement of structural lipidomics. *Trac-Trend Anal Chem* 2024;**177**. https://doi.org/10.1016/j.trac.2024.117765
14. Chen C, Wang J, Pan D *et al.* Applications of multi-omics analysis in human diseases. *MedComm (2020)* 2023;**4**:e315. https://doi.org/10.1002/mco2.315
15. Subramanian I, Verma S, Kumar S *et al.* Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 2020;**14**:117793221989905. https://doi.org/10.1177/1177932219899051
16. Zitnik M, Li MM, Wells A *et al.* Current and future directions in network biology. *Bioinform Adv* 2024;**4**:vbae099. https://doi.org/10.1093/bioadv/vbae099
17. Ritchie MD, Holzinger ER, Li R *et al.* Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* 2015;**16**:85–97. https://doi.org/10.1038/nrg3868
18. Zhao K, Rhee SY. Interpreting omics data with pathway enrichment analysis. *Trends Genet* 2023;**39**:308–19. https://doi.org/10.1016/j.tig.2023.01.003
19. Lee KS, Su X, Huan T. Metabolites are not genes - avoiding the misuse of pathway analysis in metabolomics. *Nat Metab* 2025;**7**:858–61. https://doi.org/10.1038/s42255-025-01283-0
20. Piwowar M, Jurkowski W. ONION: functional approach for integration of Lipidomics and Transcriptomics data. *PLoS One* 2015;**10**:e0128854. https://doi.org/10.1371/journal.pone.0128854

21. Lam SM, Wang Z, Li B *et al.* High-coverage lipidomics for functional lipid and pathway analyses. *Anal Chim Acta* 2021;**1147**:199–210. https://doi.org/10.1016/j.aca.2020.11.024
22. Maghsoudi Z, Nguyen H, Tavakkoli A *et al.* A comprehensive survey of the approaches for pathway analysis using multi-omics data integration. *Brief Bioinform* 2022;**23**:bbac435. https://doi.org/10.1093/bib/bbac435
23. Novoa-Del-Toro EM, Witting M. Navigating common pitfalls in metabolite identification and metabolomics bioinformatics. *Metabolomics* 2024;**20**:103. https://doi.org/10.1007/s11306-024-02167-2
24. da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A* 2015;**112**:12549–50. https://doi.org/10.1073/pnas.1516878112
25. Chappel JR, Kirkwood-Donelson KI, Reif DM *et al.* From big data to big insights: statistical and bioinformatic approaches for exploring the lipidome. *Anal Bioanal Chem* 2024;**416**:2189–202. https://doi.org/10.1007/s00216-023-04991-2
26. Garcia-Segura ME, Durainayagam BR, Liggi S *et al.* Pathway-based integration of multi-omics data reveals lipidomics alterations validated in an Alzheimer's disease mouse model and risk loci carriers. *J Neurochem* 2023;**164**:57–76. https://doi.org/10.1111/jnc.15719
27. Soundararajan R, Maurin MM, Rodriguez-Silva J *et al.* Integration of lipidomics with targeted, single cell, and spatial transcriptomics defines an unresolved pro-inflammatory state in colon cancer. *Gut* 2025;**74**:586–602. https://doi.org/10.1136/gutjnl-2024-332535
28. Pattaroni C, Begka C, Cardwell B *et al.* Multi-omics integration reveals a nonlinear signature that precedes progression of lung fibrosis. *Clin Transl Immunology* 2024;**13**:e1485. https://doi.org/10.1002/cti2.1485
29. Ancel P, Martin JC, Doukbi E *et al.* Untargeted multiomics approach coupling Lipidomics and metabolomics profiling reveals new insights in diabetic retinopathy. *Int J Mol Sci* 2023;**24**:12053. https://doi.org/10.3390/ijms241512053
30. Cadby G, Giles C, Melton PE *et al.* Comprehensive genetic analysis of the human lipidome identifies loci associated with lipid homeostasis with links to coronary artery disease. *Nat Commun* 2022;**13**:3124. https://doi.org/10.1038/s41467-022-30875-7
31. MoTr PACSG, Lead A, MoTr PACSG. Temporal dynamics of the multi-omic response to endurance exercise training. *Nature* 2024;**629**:174–83. https://doi.org/10.1038/s41586-023-06877-w
32. Bersanelli M, Mosca E, Remondini D *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016;**17**:15. https://doi.org/10.1186/s12859-015-0857-9
33. Nicora G, Vitali F, Dagliati A *et al.* Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. *Front Oncol* 2020;**10**:1030. https://doi.org/10.3389/fonc.2020.01030
34. Picard M, Scott-Boyer MP, Bodein A *et al.* Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* 2021;**19**:3735–46. https://doi.org/10.1016/j.csbj.2021.06.030
35. Baiao AR, Cai Z, Poulos RC *et al.* A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. *Brief Bioinform* 2025;**26**:bbaf355. https://doi.org/10.1093/bib/bbaf355
36. Vahabi N, Michailidis G. Unsupervised multi-omics data integration methods: a comprehensive review. *Front Genet* 2022;**13**:854752. https://doi.org/10.3389/fgene.2022.854752
37. Novoloaca A, Broc C, Beloeil L *et al.* Comparative analysis of integrative classification methods for multi-omics data. *Brief Bioinform* 2024;**25**:bbae331. https://doi.org/10.1093/bib/bbae331
38. Singh A, Shannon CP, Gautier B *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019;**35**:3055–62. https://doi.org/10.1093/bioinformatics/bty1054
39. Rohart F, Gautier B, Singh A *et al.* mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* 2017;**13**:e1005752. https://doi.org/10.1371/journal.pcbi.1005752
40. Zhang Y, Gaynanova I. Joint association and classification analysis of multi-view data. *Biometrics* 2022;**78**:1614–25. https://doi.org/10.1111/biom.13536
41. Lee C, Van der Schaar M. A Variational Information Bottleneck Approach to Multi-Omics Data Integration. *International Conference on Artificial Intelligence and Statistics*. PMLR. 1513–21.
42. Beaude A, Auge F, Zehraoui F *et al.* CrossAttOmics: multiomics data integration with cross-attention. *Bioinformatics* 2025;**41**:btaf302. https://doi.org/10.1093/bioinformatics/btaf302
43. Narayana JK, Mac Aogain M, Ali N *et al.* Similarity network fusion for the integration of multi-omics and microbiomes in respiratory disease. *Eur Respir J* 2021;**58**:2101016. https://doi.org/10.1183/13993003.01016-2021
44. Sienkiewicz K, Chen J, Chatrath A *et al.* Detecting molecular subtypes from multi-omics datasets using SUMO. *Cell Rep Methods* 2022;**2**:100152. https://doi.org/10.1016/j.crmeth.2021.100152
45. Rappoport N, Shamir R. NEMO:cancer subtyping by integration of partial multi-omic data. *Bioinformatics* 2019;**35**:3348–56. https://doi.org/10.1093/bioinformatics/btz058
46. Agrawal A, Balcı H, Hanspers K *et al.* WikiPathways 2024: next generation pathway database. *Nucleic Acids Res* 2024;**52**:D679–89. https://doi.org/10.1093/nar/gkad960
47. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559. https://doi.org/10.1186/1471-2105-9-559
48. Deng EZ, Marino GB, Clarke DJB *et al.* Multiomics2Targets identifies targets from cancer cohorts profiled with transcriptomics, proteomics, and phosphoproteomics. *Cell Rep Methods* 2024;**4**:100839. https://doi.org/10.1016/j.crmeth.2024.100839
49. Paczkowska M, Barenboim J, Sintupisut N *et al.* Integrative pathway enrichment analysis of multivariate omics data. *Nat Commun* 2020;**11**:735. https://doi.org/10.1038/s41467-019-13983-9
50. Slobodyanyuk M, Bahcheli AT, Klein ZP *et al.* Directional integration and pathway enrichment analysis for multi-omics data. *Nat Commun* 2024;**15**:5690. https://doi.org/10.1038/s41467-024-49986-4
51. Dugourd A, Kuppe C, Sciacovelli M *et al.* Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol Syst Biol* 2021;**17**:e9730. https://doi.org/10.15252/msb.20209730
52. Argelaguet R, Velten B, Arnol D *et al.* Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;**14**:e8124. https://doi.org/10.15252/msb.20178124
53. Argelaguet R, Arnol D, Bredikhin D *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:111. https://doi.org/10.1186/s13059-020-02015-1
54. Iperi C, Fernández-Ochoa Á, Barturen G *et al.* BiomiX, a user-friendly bioinformatic tool for democratized analysis and integration of multiomics data. *BMC Bioinformatics* 2025;**26**:8. https://doi.org/10.1186/s12859-024-06022-y

55. Zhou G, Pang Z, Lu Y *et al.* OmicsNet 2.0: a web-based platform for multi-omics integration and network visual analytics. *Nucleic Acids Res* 2022;**50**:W527–33. https://doi.org/10.1093/nar/gkac376
56. Arif M, Zhang C, Li X *et al.* iNetModels 2.0: an interactive visualization and database of multi-omics data. *Nucleic Acids Res* 2021;**49**:W271–6. https://doi.org/10.1093/nar/gkab254
57. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS One* 2017;**12**:e0176278. https://doi.org/10.1371/journal.pone.0176278
58. Chu SA, Hsiao Y, Wang C, Kyle J *et al.* Integrated proteogenomic and metabolomic profiling of acute myeloid leukemias to identify molecular subtypes and associated therapy targets. *Nature Cancer* https://doi.org/10.1038/s43018-026-01175-6 (in press).
59. Cheng WY, Li JF, Zhu YM *et al.* Transcriptome-based molecular subtypes and differentiation hierarchies improve the classification framework of acute myeloid leukemia. *Proc Natl Acad Sci U S A* 2022;**119**:e2211429119. https://doi.org/10.1073/pnas.2211429119
60. Wei J, Huang XJ, Huang Y *et al.* Key immune-related gene ITGB2 as a prognostic signature for acute myeloid leukemia. *Ann Transl Med* 2021;**9**:1386. https://doi.org/10.21037/atm-21-3641
61. Liberzon A, Birger C, Thorvaldsdóttir H *et al.* The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst* 2015;**1**:417–25. https://doi.org/10.1016/j.cels.2015.12.004
62. Smedley D, Haider S, Ballester B *et al.* BioMart–biological queries made easy. *BMC Genomics* 2009;**10**:22. https://doi.org/10.1186/1471-2164-10-22
63. Nepstad I, Hatfield KJ, Gronningsaeter IS *et al.* The PI3K-Akt-mTOR Signaling pathway in human acute myeloid Leukemia (AML) cells. *Int J Mol Sci* 2020;**21**:2907. https://doi.org/10.3390/ijms21082907
64. Liang X, Zhou J, Li C *et al.* The roles and mechanisms of TGFB1 in acute myeloid leukemia chemoresistance. *Cell Signal* 2024;**116**:111027. https://doi.org/10.1016/j.cellsig.2023.111027
65. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data. *Front Genet* 2019;**10**:535. https://doi.org/10.3389/fgene.2019.00535
66. Kanehisa M, Furumichi M, Sato Y *et al.* KEGG: biological systems database as a model of the real world. *Nucleic Acids Res* 2025;**53**:D672–7. https://doi.org/10.1093/nar/gkae909
67. Gene Ontology, C, Aleksander SA, Balhoff J *et al.* The gene ontology knowledgebase in 2023. *Genetics* 2023;**224**:iyad031. https://doi.org/10.1093/genetics/iyad031
68. Conroy MJ, Andrews RM, Andrews S *et al.* LIPID MAPS: update to databases and tools for the lipidomics community. *Nucleic Acids Res* 2024;**52**:D1677–82. https://doi.org/10.1093/nar/gkad896
69. Aimo L, Liechti R, Hyka-Nouspikel N *et al.* The SwissLipids knowledgebase for lipid biology. *Bioinformatics* 2015;**31**:2860–6. https://doi.org/10.1093/bioinformatics/btv285
70. Wishart DS, Guo AC, Oler E *et al.* HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res* 2022;**50**:D622–31. https://doi.org/10.1093/nar/gkab1062
71. Subramanian A, Tamayo P, Mootha VK *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**:15545–50. https://doi.org/10.1073/pnas.0506580102
72. Xie Z, Bailey A, Kuleshov MV *et al.* Gene set knowledge discovery with Enrichr. *Curr Protoc* 2021;**1**:e90. https://doi.org/10.1002/cpz1.90
73. Elizarraras JM, Liao Y, Shi Z *et al.* WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Res* 2024;**52**:W415–21. https://doi.org/10.1093/nar/gkae456
74. Danna V, Mitchell H, Anderson L *et al.* leapR: an R package for Multiomic pathway analysis. *J Proteome Res* 2021;**20**:2116–21. https://doi.org/10.1021/acs.jproteome.0c00963
75. Wu Y, Xie L. AI-driven multi-omics integration for multi-scale predictive modeling of genotype-environment-phenotype relationships. *Comput Struct Biotechnol J* 2025;**27**:265–77. https://doi.org/10.1016/j.csbj.2024.12.030
76. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform* 2022;**23**:bbab454. https://doi.org/10.1093/bib/bbab454
77. Hira MT, Razzaque MA, Angione C *et al.* Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Sci Rep* 2021;**11**:6265. https://doi.org/10.1038/s41598-021-85285-4
78. Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biol* 2021;**22**:351. https://doi.org/10.1186/s13059-021-02556-z
79. Feldner-Busztin D, Firbas Nisantzis P, Edmunds SJ *et al.* Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics* 2023;**39**:btad021. https://doi.org/10.1093/bioinformatics/btad021
80. Lin A, Ye J, Qi C *et al.* Bridging artificial intelligence and biological sciences: a comprehensive review of large language models in bioinformatics. *Brief Bioinform* 2025;**26**:bbaf357. https://doi.org/10.1093/bib/bbaf357
81. Ye X, Shi T, Huang D *et al.* Multi-omics clustering by integrating clinical features from large language model. *Methods* 2025;**239**:64–71. https://doi.org/10.1016/j.ymeth.2025.03.017
82. Ge Y, Zhang F, Liu Y *et al.* Leveraging large language models for redundancy-aware pathway analysis and deep biological interpretation. *bioRxiv*, 2025.2008.2023.671949. 2025. https://doi.org/10.1101/2025.08.23.671949
83. Alsaedi S, Gao X, Gojobori T. Beyond digital twins: the role of foundation models in enhancing the interpretability of multiomics modalities in precision medicine. *FEBS Open Bio* 2025;**15**:1192–208. https://doi.org/10.1002/2211-5463.70003
84. Cui H, Tejada-Lapuerta A, Brbić M *et al.* Towards multimodal foundation models in molecular cell biology. *Nature* 2025;**640**:623–33. https://doi.org/10.1038/s41586-025-08710-y
85. Singh R, Barsainyan AA, Irfan R *et al.* ChemBERTa-3: An Open Source Training Framework for Chemical Foundation Models. *ChemRxiv* 2025. https://doi.org/10.26434/chemrxiv-2025-4glrl-v2
86. Ciobanu-Caraus O, Aicher A, Kernbach JM *et al.* A critical moment in machine learning in medicine: on reproducible and interpretable learning. *Acta Neurochir* 2024;**166**:14. https://doi.org/10.1007/s00701-024-05892-8
87. Chi J, Shu J, Li M *et al.* Artificial intelligence in metabolomics: a current review. *Trends Analyt Chem* 2024;**178**:117852. https://doi.org/10.1016/j.trac.2024.117852
88. Liebisch G, Fahy E, Aoki J *et al.* Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures. *J Lipid Res* 2020;**61**:1539–55. https://doi.org/10.1194/jlr.S120001025
89. Garrido-Rodriguez M, Zirngibl K, Ivanova O *et al.* Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. *Mol Syst Biol* 2022;**18**:e11036. https://doi.org/10.15252/msb.202211036
90. Somers J, Fenner M, Kong G *et al.* A framework for considering prior information in network-based approaches to omics

data analysis. *Proteomics* 2023;**23**:e2200402. https://doi.org/10.1002/pmic.202200402
91. Almeida JS. Predictive non-linear modeling of complex data by artificial neural networks. *Curr Opin Biotechnol* 2002;**13**:72–6. https://doi.org/10.1016/s0958-1669(02)00288-4
92. Tuncbag N, Gosline SJC, Kedaigle A *et al.* Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLoS Comput Biol* 2016;**12**:e1004879. https://doi.org/10.1371/journal.pcbi.1004879
93. Szklarczyk D, Nastou K, Koutrouli M *et al.* The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res* 2025;**53**:D730–7. https://doi.org/10.1093/nar/gkae1113
94. Schmidt T, Samaras P, Frejno M *et al.* ProteomicsDB. *Nucleic Acids Res* 2018;**46**:D1271–81. https://doi.org/10.1093/nar/gkx1029
95. Han K, Park B, Kim H *et al.* HPID: the human protein interaction database. *Bioinformatics* 2004;**20**:2466–70. https://doi.org/10.1093/bioinformatics/bth253
96. Chowdhury S, Kennedy JJ, Ivey RG *et al.* Proteogenomic analysis of chemo-refractory high-grade serous ovarian cancer. *Cell* 2024;**187**:1016. https://doi.org/10.1016/j.cell.2024.01.016
97. Miglino N, Toussaint NC, Ring A *et al.* Feasibility of multiomics tumor profiling for guiding treatment of melanoma. *Nat Med* 2025;**31**:2430–41. https://doi.org/10.1038/s41591-025-03715-6
98. Tan L, Zhou H, Zhang B *et al.* Multiomics identifies a cholesterol-TFEB-PLD3-TLR9 axis driving immunosuppressive tumor-associated macrophage polarization in esophageal squamous cell carcinoma. *Proc Natl Acad Sci U S A* 2026;**123**:e2520427123. https://doi.org/10.1073/pnas.2520427123
99. Li Y, Huang M, Wang M *et al.* Tumor cells impair immunological synapse formation via central nervous system-enriched metabolite. *Cancer Cell* 2024;**42**:985–1002.e18. https://doi.org/10.1016/j.ccell.2024.05.006
100. Chen CY, Maner-Smith K, Khadka M *et al.* Integrative brain omics approach highlights sn-1 lysophosphatidylethanolamine in Alzheimer's dementia. *Nat Commun* 2025;**16**:9627. https://doi.org/10.1038/s41467-025-64328-8

Briefings in Bioinformatics, 2026, **27**, bbag276
https://doi.org/10.1093/bib/bbag276
Review