https://doi.org/10.1146/annurev-biodatasci-102523-103801

Annual Review of Biomedical Data Science

Harnessing Artificial Intelligence in Multimodal Omics Data Integration: Paving the Path for the Next Frontier in Precision Medicine

Yonghyun Nam,[1] Jaesik Kim,[2,3] Sang-Hyuk Jung,[1] Jakob Woerner,[1] Erica H. Suh,[1] Dong-gi Lee,[1] Manu Shivakumar,[1] Matthew E. Lee,[1] and Dokyoon Kim[1,3]

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA; email: dokyoon.kim@pennmedicine.upenn.edu

[2]Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Keywords

machine learning, multiomics data integration, multimodal data integration, precision medicine, single-cell omics, longitudinal analysis, imaging phenotypes, biobank, risk assessment

Abstract

The integration of multiomics data with detailed phenotypic insights from electronic health records marks a paradigm shift in biomedical research, offering unparalleled holistic views into health and disease pathways. This review delineates the current landscape of multimodal omics data integration, emphasizing its transformative potential in generating a comprehensive understanding of complex biological systems. We explore robust methodologies for data integration, ranging from concatenation-based to transformation-based and network-based strategies, designed to harness the intricate nuances of diverse data types. Our discussion extends from incorporating large-scale population biobanks to dissecting high-dimensional omics layers at the single-cell level. The review underscores the emerging role of

large language models in artificial intelligence, anticipating their influence as a near-future pivot in data integration approaches. Highlighting both achievements and hurdles, we advocate for a concerted effort toward sophisticated integration models, fortifying the foundation for groundbreaking discoveries in precision medicine.

1. INTRODUCTION

The landscape of biological research has undergone a remarkable transformation, courtesy of breakthroughs in high-throughput techniques that have ushered in the era of omics science. Spanning across genomics, epigenomics, transcriptomics, proteomics, and metabolomics, the realm of omics offers a kaleidoscopic view of the constituent elements that make up human biology. These layers, each a unique facet of the broader biological puzzle, have long been the subject of unimodal analyses, yielding critical insights into discrete regulatory mechanisms within our biological systems.

However, the advent of sophisticated artificial intelligence (AI), particularly through deep learning and machine learning algorithms, marks a seminal point in our scientific journey. Traditional machine learning frameworks, such as support vector machines and random forests, laid the groundwork for integrating and interpreting the tapestry of multiomics data. The rise of advanced paradigms—convolutional neural networks, graph neural networks, and recurrent neural networks—has exponentially amplified our capabilities, enabling the decoding of complex patterns and nonlinear relationships sprawling across multiomics data.

This revolution in data analytics puts scientists at the cusp of a new frontier: the comprehensive and holistic exploration of biological processes. By harnessing the singular attributes and synergistic interactions across various omics strata, machine learning serves as the linchpin in our pursuit of a deeper understanding of life's fundamental mechanisms, thereby paving the way for innovations in precision medicine.

Significant strides have been made, transitioning from mono-omics to sophisticated multiomics analyses, spurred by the maturation of data integration techniques and an ever-expanding repository of biological intelligence. Concurrently, the emergence of diverse phenotypic data from electronic health records (EHRs) has added nuanced dimensions that are intimately tied to patient health narratives—ranging from clinical diagnoses and therapeutic histories to vital parameters and advanced imaging findings.

The convergence of multilayered omics data with rich phenotypic breadcrumbs from EHRs signifies an exciting convergence in biomedical research. Particularly, the systematic amalgamation of large-scale multiomics data with nuanced, patient-centric information from EHRs promises to unlock previously cryptic corners of human biology. By shedding light on the subtle dance between molecular constituents and clinical phenomenology, this integrative approach heralds a transformative era in precision medicine.

In this review, we embark on a journey through the dynamic evolution, inherent challenges, innovative methodologies, and untapped potential residing within the realm of multimodal omics data integration. Standing on the brink of this scientific odyssey, we are witnesses to a nascent revolution in integrative biomedical analysis, one that holds the promise to redefine our understanding of health and the therapeutic strategies of tomorrow.

2. OPPORTUNITIES IN MULTIMODAL OMICS DATA ANALYSIS FOR PRECISION MEDICINE

With the precipitous decline in the costs associated with high-throughput sequencing and other massively paralleled biomolecular technologies, a novel window of opportunity has opened up,

offering enhanced accessibility to these advanced tools (1). This progression facilitates their seamless integration into both clinical research and practical applications. In this review, we distinguish between traditional and recent multiomics analysis, emphasizing the emerging opportunities in multimodal omics data integration. Specifically, we highlight how the expanded scale and improved resolution of samples now available are forging new frontiers for multiomics data acquisition, as well as how interdisciplinary studies with multiomics data and machine learning can contribute to precision medicine.

Early multiomics approaches largely center on the analysis of bulk samples, meaning that they use omics data from a collection of numerous cells obtained from a given sample. These methodologies have been instrumental in providing insights into specific tissues or cell populations and in furthering our understanding of dominant pathways and disease pathogenesis. Multiomics analysis using bulk samples has had success in describing a broad overview of biological processes. Over the past three-plus decades, many noteworthy omics-specific projects and consortiums have provided valuable contributions and insights into biological pathways and pathophysiological conditions, as well as high-resolution functional interpretations of specific omics data (**Table 1**).

However, since these findings are derived from the average data across numerous cells, their direct use for precision medicine is challenging; integrative analysis with these techniques often falls short when it comes to deciphering the intricate cellular heterogeneity within samples. Moreover, multiomics data analysis with bulk samples may not capture subtle yet critical differences at the population level, such as variation/diversity in biological pathways across races or genders (31–34). Instead, the remaining gaps in our comprehension of human biology could potentially be bridged with the integrated analysis of emerging multiomics datasets—specifically, multimodal omics data sourced from large-scale biobanks and high-resolution single-cell multiomics data—with comprehensive and sophisticated findings that can help take us a step closer to precision medicine.

In the remainder of this review, we broadly divide recent multiomics data into two categories with respect to high scalability and resolution: The former represents multiomics data derived from large-scale community-based cohorts, and the latter represents single-cell multiomics data. From the perspective of data integration, we describe the scalability of integrated analysis and multiomics data with various modalities that are difficult to access with data consisting of existing random bulk samples. Moreover, we highlight the unique strengths of each category in the context of precision medicine.

Bulk multiomics data: comprehensive omics data from a mixed population of cells, rather than from individual cells

Biomarker: a significant biological indicator discovered through omics data integration analysis, such as a measurable variable or a marker for genes, gene expression, or proteins

2.1. Large-Scale Community-Based Multiomics Data

Large-scale, national, and community-based biobanks have started to incorporate heterogeneous phenotypic data derived from EHRs alongside bulk multiomics data (35) (**Table 2**). Biobank-scaled multiomics data integration is transformative in the sense that it offers a shift from random bulk samples to organized population-level datasets. This shift allows for better representation of individual characteristics, which represent a broader spectrum of the population, incorporating variation due to factors like age, ethnicity, lifestyle, and more (36–39). This improved representation lends itself to a more holistic and real-world-applicable understanding of human biology, particularly new insights into disease risk and manifestation. Notably, there is a significant advantage in integrating omics data with the observable traits or characteristics of individuals, which can illuminate the etiology or biomarkers for specific diseases (40). Integration of multiomics and phenotypic data from biobanks enables the development of molecular-level disease understanding at the population level, because multimodal datasets derived from consistent populations potentially reflect the characteristics of the population. The strength of multiomics approaches lies in

Table 1 Omics-specific projects and consortiums

Omics-specific consortium	Donor status	Year of launch	Reference
Genomics			
Human Genome Project (2, 3)	Healthy	1990	**https://www.genome.gov/human-genome-project**
1000 Genomes Project (2, 4)	Healthy	2007	**https://www.internationalgenome.org/**
gnomAD (5)	Various	2014	**https://gnomad.broadinstitute.org/**
Epigenetics			
ENCODE (6)	Cancer/healthy	2003	**https://www.encodeproject.org/**
Roadmap Epigenomics Project (7)	Healthy	2007	**https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/**
International Human Epigenome Consortium (8, 9) (including Roadmap and ENCODE)	Various	2010	**https://ihec-epigenomes.org/about**
Transcriptomics			
GTEx and eGTEx (10, 11)	Healthy	2011	**https://www.gtexportal.org/home/**
eQTLGen (12)	Various	2018	**https://www.eqtlgen.org/**
FANTOM (13–15)	Healthy	2001	**https://fantom.gsc.riken.jp/5/**
GEUVADIS (16, 17)	Healthy	2010	**https://www.internationalgenome.org/data-portal/data-collection/geuvadis**
GENCODE (18)	Various	2003	**https://www.gencodegenes.org/**
Proteomics			
The Human Proteome Project (19)	Various	2010	**https://hupo.org/human-proteome-project**
CPTAC (20, 21)	Cancer	2011	**https://proteomics.cancer.gov/programs/cptac**
EDRN (22)	Cancer	2000	**https://edrn.nci.nih.gov/**
Fenland study (23)	Healthy	2005	**https://www.omicscience.org/** **https://www.mrc-epid.cam.ac.uk/research/studies/fenland/**
EPIC-Norfolk (24, 25)	Various	1993	**https://www.epic-norfolk.org.uk/**
Metabolomics			
Human Metabolome Database (26)	Healthy	2007	**https://hmdb.ca/**
EPIC-Norfolk (24, 27)	Various	1993	**https://www.epic-norfolk.org.uk/**
Microbiolomics			
Integrative HMP Research Network Consortium (28, 29)	Various	2007 (HMP) 2014 (iHMP)	**https://www.hmpdacc.org/**
Dutch Microbiome Project (30)	Healthy	2015	**https://dutchmicrobiomeproject.molgeniscloud.org/**

Abbreviations: CPTAC, Clinical Proteomic Tumor Analysis Consortium; EDRN, Early Detection Research Network; eGTEx, Enhancing Genotype-Tissue Expression; ENCODE, Encyclopedia of DNA Elements; EPIC, European Prospective Investigation into Cancer; FANTOM, Functional Annotation of the Mammalian Genome; GEUVADIS, Genetic European Variation in Disease; gnomAD, Genome Aggregation Database; GTEx, Genotype-Tissue Expression; HMP, Human Microbiome Project; iHMP, Integrative Human Microbiome Project.

Table 2 Multiomics projects in national and community-based cohorts for biomedical research

Biobank	Project characteristics				Omics data					Reference(s)
	Major ethnicity (region)	Donor status	Cohort size (participants)	Year of launch	Genomics	Epigenetics	Transcriptomics	Proteomics	Metabolomics	
TOPMed Program	Mixed (USA)	Various	~205,000	2014	✓	✓	✓	✓	✓	41, 42
iPOP	Mixed (USA)	Pre-diabetic	100	2012	✓	✓	✓	✓	✓	43, 44
UK Biobank	European (UK)	Various	~500,000	2007	✓			✓	✓	2, 45–47
INTERVAL trial	European (UK)	Healthy	~50,000	2012	✓		✓	✓	✓	48–50
BIOS Consortium	European (Netherlands)	Various	3,841	2014	✓	✓	✓		✓	51–53
Lifelines Biobank	European (Netherlands)	Healthy	167,000	2006	✓	✓	✓		✓	54, 55
Estonian Biobank	European (Estonia)	Healthy	>200,000	2002	✓				✓	56, 57
China Kadoorie Biobank	East Asian (China)	Various	>512,000	2004	✓			✓	✓	58
GNHS	East Asian (China)	Various	>5,000	2008	✓			✓	✓	59
Taiwan Biobank	East Asian (Taiwan)	Healthy	>150,000	2012	✓	✓			✓	60
BioBank Japan	East Asian (Japan)	Various	260,000	2003	✓			✓	✓	61
jMorp	East Asian (Japan)	Healthy	>5,000	2015	✓	✓	✓	✓	✓	62, 63
KoGES	East Asian (Korea)	Various	>245,000	2001	✓				✓	64, 65
KoCAS	East Asian (Korea)	Healthy	>1,500	2005	✓				✓	66, 67
CHAIN Network	African and South Asian	Children with acute illness	3,101	2016	✓			✓	✓	68
SPHS	South Asian (Singapore)	Various	364	2003	✓	✓	✓			69

Abbreviations: BIOS, Biobank-based Integrative Omics Study; CHAIN, Childhood Acute Illness and Nutrition; GNHS, Guangzhou Nutrition and Health Study; iPOP, Integrated Personal Omics Project; jMorp, Japanese Multi-Omics Reference Panel; KoCAS, Korean Children-Adolescent Cohort Study; KoGES, Korean Genome and Epidemiology Study; SPHS, Singapore Population Health Studies; TOPMed, Trans-Omics for Precision Medicine.

their ability to synthesize and interpret these data types collectively, enhancing our understanding of complex biological systems and phenotypes.

Notably, in addition to consortium-based cohorts (**Table 2**), nationwide repositories for biomedical research, including the Database of Genotypes and Phenotypes (dbGaP) and the European Genome-Phenome Archive (EGA), also play a pivotal role in driving the frontier of precision medicine (70, 71). Both dbGaP and EGA collectively provide access to a wide range of multiomics data, including genomic, transcriptomic, proteomic, and metabolomic datasets from diverse populations and studies. These extensive multiomics sources allow us to deeply investigate disease causes and identify biomarkers that are important for understanding complex diseases in diverse populations.

A significant advantage of multiomics approaches is the ability to extract single omics data within the population level. This feature is particularly beneficial for integrative genetics research, allowing for the use of diverse methodologies:

- Epigenome-wide association studies (EWASs) focus on the integration of epigenomic data, such as DNA methylation and histone modifications, with clinical phenotypes. This approach is particularly useful for investigating the impact of epigenetic modifications on disease susceptibility and progression (72). EWASs can uncover epigenetic markers associated with diseases, shedding light on how environmental factors interact with an individual's genetic makeup to influence gene expression patterns.
- Transcription-wide association studies (TWASs) involve the integration of transcriptomics and genome-wide association study (GWAS) data. By leveraging gene expression information from tissues or cell types relevant to a specific trait, derived from sources like the Genotype-Tissue Expression Portal, the PsychENCODE Knowledge Portal, and the CommonMind Consortium, TWASs can identify genes whose expression levels are associated with a trait of interest (11, 73, 74).
- Proteome-wide association studies (PWASs) have emerged to explore the connections between protein profiles and health conditions. PWASs identify associations between changes in protein expression levels and specific health outcomes, offering insights into complex disorders and providing potential biomarkers and treatment targets for precision medicine (75).
- Metabolome-wide association studies (MtWASs) explore the connection between metabolites—small molecules involved in biochemical processes—and traits or diseases. Integrating metabolomic data with genetic information can reveal metabolic pathways and networks that are linked to specific physiological conditions (76, 77).

In addition to integrative analysis within omics data, the use of population-level-derived datasets provides expanded opportunities for integrated analysis with heterogeneous phenotypic data or for discovering associated biological components. For example, the integration of omics data with diverse clinical information, including medical images, has become increasingly feasible. Imaging techniques such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) offer detailed insights into the structure and function of tissues and organs within the body. Common multiomics data analyses often do not support such cross-examinations unless structured within meticulously designed cohorts. However, biobanks present a solution as they collect both imaging phenotypes and omics data from the same individuals, thereby enabling more straightforward combined analysis. Furthermore, integrated analysis of phenotypic and multiomics data facilitates longitudinal integration of multimodal omics data, an emerging approach that combines data collected over extended periods from the same samples (78). This longitudinal perspective can reveal how biological systems evolve over time,

highlighting trends, patterns, and associations that might not be apparent from cross-sectional studies. A longitudinal approach can thus offer a holistic view of the underlying mechanisms driving health, disease, and responses to treatments.

2.2. High-Resolution Multiomics Data Integration Analysis from Single Cells

While biobanks offer scale and better population representation, single-cell omics dives deeper into specific biological and cellular mechanisms, enhancing the resolution to the level of individual cells. That is, instead of understanding a cell population as a homogeneous entity and aggregating data from millions of cells, single-cell analyses recognize and chart the heterogeneity within, providing a detailed, cell-specific landscape that reveals cellular intricacies often overshadowed in bulk analyses. By comprehensively profiling biological molecules—such as DNA (genomics), RNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics)—from individual cells, single-cell multiomics data elucidate the unique molecular signature of each cell (79). Understanding the heterogeneity within cell populations is crucial for understanding complex biological processes, disease mechanisms, and developmental pathways; most of all, it is essential for understanding the comprehensive interplay of molecular processes in health and disease. We introduce single-cell omics technology by highlighting its utilization of various omics and discuss opportunities for integration analysis accordingly. Detailed descriptions of each single-cell omics method are provided in **Supplemental Appendix 1**.

Supplemental Material >

2.2.1. Integration of genomic and transcriptomic data. Genome and transcriptome sequencing (G&T-seq) simultaneously measures genomic DNA and mRNA from single cells, elucidating the genetic makeup, variations, and gene expression profile (80). gDNA-mRNA sequencing (DR-seq) likewise investigates the relationship between the genome and transcriptome in individual cells, enabling an understanding of cellular behavior and functional characteristics (81). Both methods unveil genetic and transcriptomic heterogeneity within tissues, improving precision in the identification of cellular origins of diseases and potential therapeutic targets.

2.2.2. Integration of transcriptomic and proteomic data. Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) combines the specificity of antibody detection with next-generation sequencing by tagging antibodies with unique oligonucleotide barcodes to label proteins in cells (82); messenger RNA (mRNA) from the same cells is sequenced in parallel. This method provides a comprehensive view of cell functions, accounting for differences between protein and mRNA levels. RNA expression and protein sequencing (REAP-seq) is an integrated single-cell sequencing method likewise designed to concurrently profile the transcriptomes and proteomes of individual cells (83) that provides invaluable insights into cellular function and regulation. Both methods combine RNA sequencing with protein-level insights in the same cell.

2.2.3. Integration of epigenetic and transcriptomic data. Transcriptome, epitope, and chromatin accessibility sequencing (TEA-seq) focuses on RNA modifications at the single-cell level, elucidating how these modifications interact with other regulatory layers (84). It is tailored to profile targeted modifications of interest rather than to conduct a global survey. ATAC (assay for transposase-accessible chromatin) with select antigen profiling by sequencing (ASAP-seq) quantifies nascent RNA molecules and so measures the rate of transcription in single cells (85), offering insights into transcriptional dynamics and gene expression regulation. Together, these methods provide granular insights into RNA modifications and transcription rates at the individual cell level, a resolution not attainable with traditional methods.

2.2.4. Integration of multimodal single-cell omics with phenotypic data. Integrating single-cell multiomics with phenotypic data involves associating the detailed molecular profiles of individual cells with their observable characteristics or functions. This comprehensive approach allows for a more holistic understanding of how molecular changes influence cell behavior and the overall organismal phenotype; furthermore, by correlating molecular profiles with phenotypic responses, researchers can better identify therapeutic targets and predict drug responses (86).

Features: measurable attributes used as input for models to make predictions or decisions; also called variables or attributes

High dimensionality: a situation where the number of features in a dataset is very large, often exceeding the number of observations; this can lead to issues such as overfitting and increased computational complexity

Canonical correlation analysis: a statistical method for exploring correlations between two sets of multidimensional variables

Supervised learning: learning system where a model is trained on paired data with input features and output labels to predict outcomes

3. MACHINE LEARNING FOR MULTIMODAL OMICS DATA INTEGRATION ANALYSIS

Many researchers have examined the challenges that arise during data analysis of either single omics or multiomics data and have proposed solutions tailored to each scenario (87, 88). Machine learning plays a critical role in analyzing heterogeneous biological and clinical components independently as well as integrating multimodal omics data by leveraging their underlying relationships. In this section, we briefly introduce representative machine learning approaches for data integration. Subsequently, we provide a more detailed discussion on integrating imaging phenotypes with multiomics and the analysis of longitudinal multiomics data.

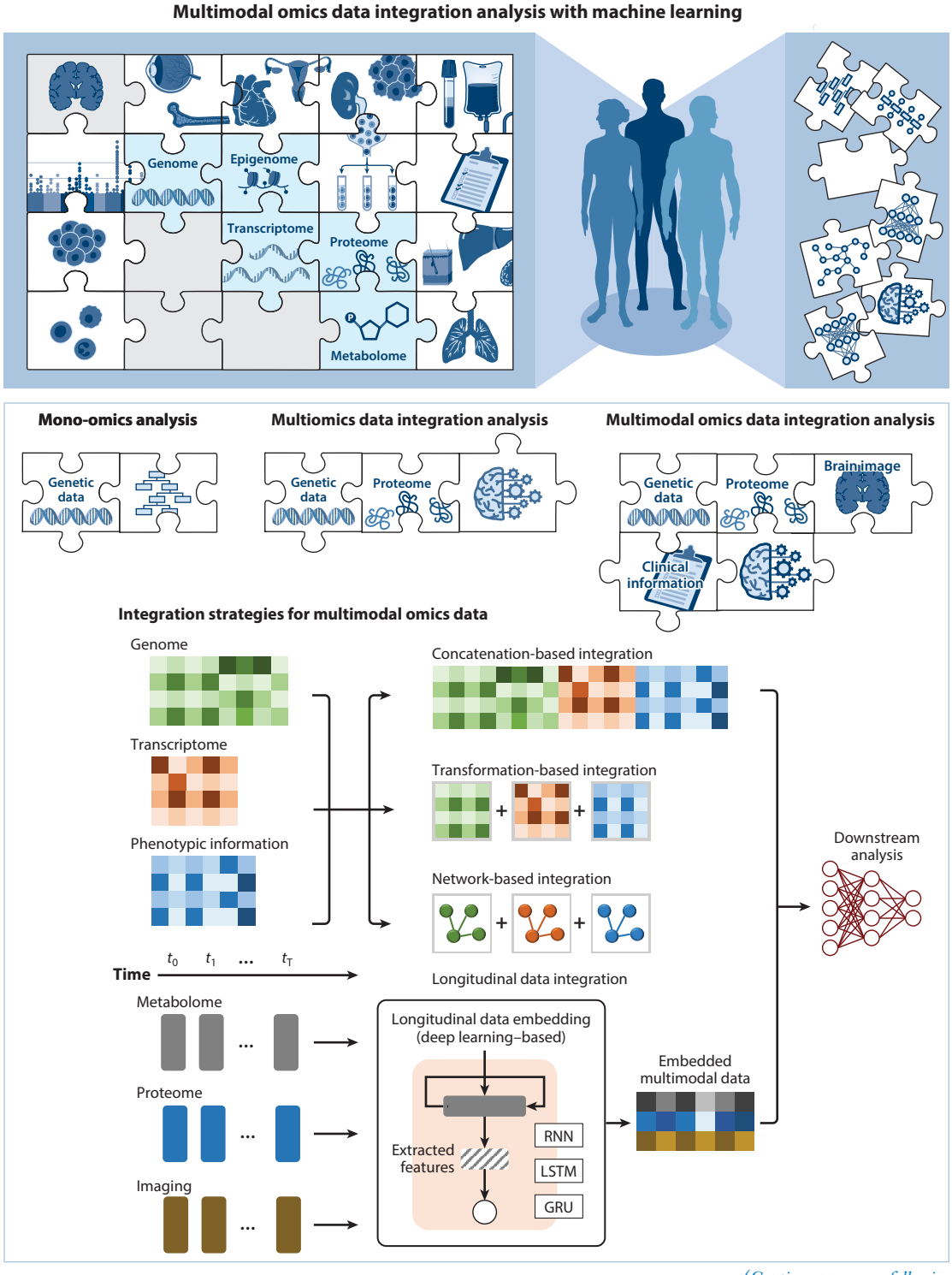
3.1. Integration Strategies for Utilizing Machine Learning with Multimodal Data

Analyzing multiomics data presents significant challenges due to the inherent data heterogeneity that arises from the diverse techniques and platforms used for generating different omics data types. These datasets can typically be visualized as structured matrices, with rows representing samples and columns representing the biological features specific to each omics category. Tasks may involve classification at the sample level, such as predicting cell types or stratifying patients, or identifying biomarkers at the feature level. In either case, machine learning methodologies provide robust solutions for the integration and interpretation of such intricate data.

A major hurdle with multiomics data is their high dimensionality, a common characteristic even for data of a single omics type. This problem is further complicated when integrating data from multiple omics sources, as overlapping samples may be lacking. To combat the issue of high dimensionality, researchers often turn to methods for feature selection, such as filter-based, wrapper-based, or embedded methods. Feature extraction techniques like principal component analysis, canonical correlation analysis, nonnegative matrix factorization, and the use of an autoencoder also come in handy.

Another layer of complexity arises from issues related to data sparsity and the task of interpreting the results. To weave together and make sense of multiomics data, researchers employ a spectrum of methodologies ranging from traditional machine learning to state-of-the-art deep learning, all tailored to data characteristics and the objective of the analysis. Here, we discuss three representative integration strategies: concatenation-based integration, transformation-based integration, and network-based integration (**Figure 1**).

Concatenation-based integration stands out as one of the most straightforward strategies for amalgamating multiomics data. It involves directly combining (or joining) features from different omics datasets by aligning them side by side (89–91). To illustrate, if one possesses two matrices of data, one from genomics and another from transcriptomics, pertaining to the same set of samples, the concatenation approach would position the columns of one matrix adjacent to those of the other, culminating in a broader matrix. With such a concatenated feature matrix, one can use any kind of supervised learning for analysis; moreover, this approach is straightforward to implement and understand because the matrix preserves the original features derived from each omics method. Nonetheless, a concatenated matrix often demands supplementary preprocessing



(Caption appears on following page)

Figure 1 (*Figure appears on preceding page*)

Overview of multimodal omics data integration analysis. The top panel illustrates comprehensive multiomics data sources and advanced deep learning and machine learning algorithms. Each omics layer represents a piece of the human biological puzzle, each with its unique properties. The lower panel depicts broad concepts of integration approaches and integration strategies for multimodal omics data described in this review. The downstream analysis emphasizes the goal of these integrative approaches—deriving actionable insights for precision medicine. Abbreviations: GRU, gated recurrent unit; LSTM, long short-term memory; RNN, recurrent neural network.

steps, such as normalization or batch effect mitigation, as data from different omics sources might operate on different scales or distributions.

Transformation-based integration is a technique used to blend multiomics data by shifting the data from various omics sources into a shared feature space or framework. By mapping data to these unified spaces, it becomes feasible to identify relationships across different omics layers; hence, this method excels at revealing hidden relationships and patterns within and between omics sets (92). However, in contrast to the direct concatenation-based method, transformation-based integration can obfuscate the understanding of the original data, making it difficult to derive direct biological or clinical insights. Several machine learning methods are frequently employed for transformation-based integration. For example, canonical correlation analysis discerns linear combinations of features from two omics datasets that correlate maximally, making it suitable for two-omics data integration. Multiple kernel learning (MKL) combines several kernels (each corresponding to a different omics data type) to optimize a single learning task. Each kernel gauges similarity within its respective omics type, and MKL fuses them into a consolidated kernel space. Nonnegative matrix factorization breaks down data into products of nonnegative matrices, and, when applied to multiomics data, reveals shared patterns or meta-genes. Finally, deep learning techniques such as an autoencoder are trained to jointly represent multiomics data, with the bottleneck layer encapsulating a compressed, integrated representation of the input datasets.

Network-based integration is used to integrate and analyze multiomics data in the context of biological networks (93–95); it leverages the structured relationships between biological entities, such as genes, proteins, or metabolites, to provide context and improve the interpretation of multiomics data. This method essentially visualizes and analyzes data in the form of networks, where nodes represent biological entities and edges represent relationships or interactions between them. Once the multiomics data are converted to an adjacency or similarity matrix, kernel-based algorithms or graph neural networks can be applied. Thus, network-based integration transforms multifaceted omics data into a structured, interconnected framework that facilitates a holistic view of biological systems. The interconnected nature of networks naturally complements the intricate relationships within and between omics datasets, offering an intuitive platform for multiomics data exploration and hypothesis generation.

3.2. Integrative Analysis Through Combining Multiomics and Imaging Phenotypes

Multiple kernel learning (MKL): a machine learning technique that combines multiple kernels (functions defining similarities between data points) from different sources or features

Biobanks that link patient EHRs with omics data have paved the way for developing an unprecedented degree of insight into disease risk and manifestation. As imaging modalities such as CT, MRI, and PET provide detailed insights into the structure and function of tissues and organs, the recent inclusion of medical imaging in EHRs has opened up the potential for finer phenotyping and developing an enhanced understanding of how various structures correlate with disease. Radiomics transforms these routine medical images into quantitative data that can be mined, allowing image features to be analyzed alongside other biological information. However, image features encompass comprehensive characteristics, including pathological conditions, and differ in modality

from other data sources; as such, integrated analysis of imaging with omics data demands significant computational effort. Machine learning is essential for the various stages of this integrated analysis, including data preprocessing, feature selection and extraction, designing integration approaches, model selection, and model evaluation. Notably, deep learning–based algorithms are well suited for handling the complexity of multimodal data and also for identifying intricate patterns and correlations (96). By building descriptive and predictive models, deep learning methods enable researchers to extract valuable insights from the integrated data, contributing to a more comprehensive understanding of complex diseases and enhancing healthcare decision-making.

Radiogenomics, which specifically integrates medical imaging with genetic data, aims to identify disease mechanisms, predict prognosis, and assess treatment responses by investigating the relationships between imaging features and genetic factors; in particular, it can elucidate genomic risk factors associated with a phenotype of interest and the genetic architecture shared between phenotypes and morphologies. This approach has been applied to phenotypes derived from imaging of specific tissues and organs, including structural and functional brain imaging (97–99), cardiac imaging (100), retinal imaging (101, 102), and lung imaging (103). Beyond genomics, further incorporating diverse multimodal omics data into such analyses can provide supplementary insights into the pathobiological pathways and disease-associated risk factors. Such integrated studies have been systematically conducted across a broad spectrum of diseases such as autism spectrum disorder, oral diseases, and lung cancer (103, 104). In addition, a metagenomic study of the gut microbiome has revealed associations between protective bacteria, cognitive traits, and brain regional volumes derived from imaging (105). The integration of microbiome, metabolomics, cytokine measurements, cognitive assessments, and brain imaging data in that study provides further evidence of the greater insights that can be obtained by utilizing more data modalities.

3.3. Longitudinal Multimodal Omics Data Integration

Longitudinal multiomics integration combines data obtained over time from the same samples, shedding light on biological evolution and patterns not seen in cross-sectional studies. It is therefore beneficial in studying disease progression, pinpointing early diagnostic biomarkers, evaluating therapy effects, and understanding interconnected biological mechanisms influencing health and treatment responses. However, longitudinal multiomics integration also presents numerous challenges that demand sophisticated computational and analytical strategies. Specific to longitudinal data, challenges like temporal misalignment, missing data, and irregular sampling intervals require robust methods for data preprocessing and normalization. Interdependency between multiple time points from the same sample, unbalanced datasets, and unexpected outlier events add to the data complexity. Finally, the time element further amplifies the high-dimensional nature of multiomics data, an issue compounded by the typically limited number of samples and high sample variability.

Methods developed for longitudinal multiomics data integration typically have the following components: (*a*) preprocessing, (*b*) modeling, and (*c*) analysis (e.g., predicting a phenotype, identifying omics biomarkers for the progression of a phenotype, extracting clusters of omics features that have similar patterns, etc.). Preprocessing tasks may involve exclusion of subjects with limited time points or low variation, imputation of missing data, normalization, and feature selection to identify the most informative elements for subsequent analysis. Preprocessing and modeling may be performed sequentially or simultaneously, depending on the framework. For example, Bodein et al. (106) developed timeOmics, a longitudinal multiomics integration framework that performs each component sequentially. Preprocessing in this framework includes a fold-change-based filter

to focus on time-sensitive molecules; after preprocessing, a linear mixed model spline framework is used to model the expression of each biological feature while considering interindividual variability. Six unsupervised integration methods with unsupervised learning are then used to cluster multiomics expression profiles and identify key molecular features per cluster. Metwally et al. (107) followed a similar structure with their proposed method, OmicsLonDA, in which they first input preprocessed data, then apply a Gaussian smoothing spline regression model in a semiparametric approach to discern significant time intervals of omics features between study groups.

In addition to statistical methods, recurrent neural networks (RNNs) are a prevalent deep learning method for integrating longitudinal multiomics or multimodal data. Their effectiveness is clear from their success in other areas such as natural language processing, time series prediction, and speech recognition, just to name a few (108, 109). The advantages of using RNN-based models are threefold. First, they process sequences by iterating through individual elements while retaining a memory state that captures the context of prior elements. Second, the same parameters are used across different time steps, promoting compactness and generalization. Third, RNNs are naturally equipped to handle sequences of varying lengths, making them flexible and able to address a variety of tasks (110). However, while vanilla RNNs introduced the concept of processing sequences, they struggle with long sequences due to challenges like vanishing and exploding gradients (111). Modern adaptations, such as long short-term memory networks and gated recurrent units (GRUs), were designed to overcome these issues and are now more prevalent in practice.

When applied to longitudinal multiomics integration, some frameworks using these models combine the preprocessing, modeling, and analysis stages, while others approach them as separate, consecutive steps. Lee et al. (112) used the latter approach by training a unique GRU for each data modality to transform longitudinal data into fixed-size vectors. The respective vectors for each modality were then concatenated and merged into an integrated vector, which was used to predict progression from mild cognitive impairment to Alzheimer's disease (AD) using an l1-regularized logistic regression. The former approach of combining steps is possible because RNNs are able to impute missing values or interpolate between observed time points, creating continuous sequences for further analysis. Nguyen et al. (113) adapted MinimalRNN to simultaneously impute missing data and predict AD progression and found it to perform better than typical interpolation methods such as forward and linear filling and other baseline methods. Jung et al. (114) developed a deepRNN framework that combines imputation, encoding, and analysis to predict clinical status trajectories.

Ultimately, longitudinal multiomics integration has the potential to uncover previously hidden relationships between different biological layers and glean insights that remain concealed in cross-sectional studies. However, the path to such insights is strewn with challenges—from the intricacies of temporal data to the high dimensionality of multiomics data. The advent of deep learning approaches, especially RNNs and their advanced derivatives, offers compelling strategies for managing the complexities inherent to these data. As this field continues to evolve, it will be crucial to refine existing methodologies and develop new tools, ensuring the realization of the full potential of longitudinal multiomics integration for health and disease.

Unsupervised learning: learning system where a model is trained on input features only, focusing on identifying underlying structures or patterns within the data

Long short-term memory: a type of RNN architecture used in deep learning, particularly effective in processing and predicting sequences of data with its ability to remember long-term dependencies

Gated recurrent unit (GRU): a type of RNN designed to adaptively capture dependencies of different timescales in sequence data with a simpler and more efficient architecture

4. DOWNSTREAM MULTIMODAL OMICS DATA INTEGRATION ANALYSES

In this section, we review several studies for downstream applications that leverage the vast resources of large biobanks to uncover valuable insights into complex traits and diseases. These illustrate the scale and potential of multimodal omics data integration.

4.1. Large-Scale Multimodal Omics Data Integration

Large-scale biobanks have paved the way for extensive multimodal omics data integration, which in turn enables a deeper understanding of genetic, epigenetic, and metabolic contributions to human health. We explore the key applications of GWASs, EWASs, TWASs, and MtWASs, which incorporate phenotypic data with genetic, epigenetic, transcriptomic, and metabolomics data, respectively.

TWASs can identify genes whose expression levels are associated with a particular trait. Integrating TWAS results into polygenic risk scores (PRSs) can enhance the predictive power of PRS models, more accurately capturing the functional consequences of genetic variants and hence improving disease risk and other phenotypic outcome predictions. Recent studies on complex neuropsychiatric traits (115), chronic obstructive pulmonary disease (116), and other phenotypes have demonstrated improved prediction performance and enhanced cross-ancestry PRS portability (116, 117). EWASs aid in discovering epigenetic markers related to a disease, providing insights into the interaction between environmental factors and individual genetics that influence gene expression. They also highlight the genes, pathways, and molecular mechanisms associated with common and complex traits like attention-deficit hyperactivity disorder (ADHD) (118), speech sound disorders (119), and coronary artery disease (CAD) (120). In addition, EWASs have been instrumental in identifying specific genetic factors that contribute significantly to diseases known for their high polygenic risk burdens, such as ADHD (121) and autism (122). Integrating metabolomic data with genetic information can further unveil metabolic pathways and networks that are tied to specific physiological states; for instance, Kojouri et al. (76) examined the causal associations between physical activity, body mass index, and metabolites. More broadly, MtWASs can identify metabolites linked to disease risk, giving a more comprehensive view of disease etiology and potential biomarkers, and have been employed for conditions like depression, chronic kidney disease, and neovascular age-related macular degeneration (123, 124). Several studies have further inferred the causality of metabolites in human diseases using medallion randomization (125, 126).

4.2. Multiomics Data Integration for Targeting a Disease of Interest

Multiomics projects and consortia focused on specific diseases are also noteworthy (**Table 3**). Exploration of disease etiology at the molecular level increasingly requires multiomics and computational algorithms. Among them, The Cancer Genome Atlas (TCGA) stands out as a landmark resource providing an extensive overview of genomic data for numerous cancer types. Complementary to TCGA, the cBioPortal for Cancer Genomics provides an interactive platform for accessing and analyzing the vast amounts of data generated by TCGA and other cancer genomics projects. This portal facilitates the exploration of molecular profiles for thousands of cancer samples and supports the integration of diverse omics data for comprehensive cancer research (127–130). For clinical diseases, disease biomarkers play key guiding roles in diagnosis and prognosis, as they have unique advantages in evaluating early, low-level damage and so provide for early warning, prognostic efficacy analysis, and accurate staging and typing. Although many cancer-associated biomarkers have been identified through single omics, multiomics approaches may provide enhanced benefits by uncovering biomarkers that are shared across different cancer types. Multiomics also has many applications in the study of neurodegenerative biomarkers and therapeutic targets.

In addition, recent years have seen increased acknowledgment of the human microbiome's crucial role in health and disease, leading to many microbiome-wide association studies (MWASs) exploring the intricate relationship between the human microbiome and health outcomes; these

Table 3 Multiomics projects and consortiums targeting specific diseases

Disease-specific consortium	Year of launch	Genomics	Epigenetics	Transcriptomics	Proteomics	Metabolomics	Reference
Cancer							
ICGC	2008	✓		✓			https://platform.icgc-argo.org/
TCGA	2011	✓	✓	✓			https://portal.gdc.cancer.gov/
COSMIC Cell Lines	2004	✓	✓	✓			https://cancer.sanger.ac.uk/cell_lines
DepMap	2018	✓	✓	✓	✓		https://depmap.org/portal/
ICGA	2020	TBD	TBD	TBD	TBD		https://www.icga.in/
TARGET	2016	✓	✓	✓			https://www.cancer.gov/ccg/research/genome-sequencing/target
Alzheimer's disease							
AMP-AD	2014	✓	✓	✓	✓	✓	https://adknowledgeportal.synapse.org/Explore/Programs/DetailsPage?Program=AMP-AD
ROSMAP	1994	✓	✓	✓	✓	✓	https://dss.niagads.org/cohorts/religious-orders-study-memory-and-aging-project-rosmap/
ADNI	2003	✓	✓	✓	✓	✓	https://adni.loni.usc.edu/
Parkinson's disease							
AMP-PD (including PPMI)	2018	✓	TBD	✓	✓	TBD	https://www.amp-pd.org/
Depression							
STRADL	2015	✓	✓				https://datashare.ed.ac.uk/handle/10283/2988
DGN	2014	✓		✓			https://www.nimhgenetics.org/
Aging							
Aging Atlas	2020		✓	✓	✓		https://ngdc.cncb.ac.cn/aging/index
COVID-19							
COMBAT Consortium	2019		✓	✓	✓		https://www.combat.ox.ac.uk/
Psychiatric disease							
CommonMind Consortium	2012	✓	✓	✓			http://www.synapse.org/CMC
PsychENCODE	2015	✓		✓			https://psychencode.synapse.org/
Other							
MoTrPAC (mechanisms of exercises)	2019	✓	✓	✓	✓	✓	https://motrpac.org/

References for studies using multiomics from the disease-specific consortium can be found in **Supplemental Appendix 2**. Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; AMP-AD, Accelerating Medicines Partnership Program for Alzheimer's Disease; AMP-PD, Accelerating Medicines Partnership for Parkinson's Disease; COMBAT, COVID-19 Multi-omics Blood Atlas; COSMIC, Catalogue of Somatic Mutations in Cancer; DepMap, Dependency Map; DGN, Depression Genes and Networks; ICGA, Indian Cancer Genome Atlas; ICGC, International Cancer Genome Consortium; MoTrPAC, Molecular Transducers of Physical Activity Consortium; PPMI, Parkinson's Progression Markers Initiative; ROSMAP, Religious Orders Study/Memory and Aging Project; STRADL, Stratifying Resilience and Depression Longitudinally; TARGET, Therapeutically Applicable Research to Generate Effective Treatments; TBD, to be determined; TCGA, The Cancer Genome Atlas.

Supplemental Material >

have revealed potential biomarkers, therapeutic targets, and new pathways (131, 132). Similarly to MtWASs, causal links have been found between microbiome data and diseases like heart failure (133) and COVID-19 severity (134) through Mendelian randomization. These comprehensive repositories shaped by multiomics-driven data will play a pivotal role in unraveling the complex nuances of various diseases, thereby advancing biomedical research toward the realm of precision health.

Transpathology: the integration of molecular imaging and pathology data for in-depth disease mechanism analysis

4.3. Multiomics with Imaging Phenotypes

Medical imaging has vastly improved our understanding of complex diseases, a fact that underscores the importance of transpathology, the combination of molecular imaging and pathology data, in investigating disease mechanisms (135). In transpathology, specific attributes of diseases, such as phenotypic traits, structural abnormalities, and progression patterns, are directly compared to imaging, providing a more comprehensive view of disease-related biological processes. An advantage of this approach is that it provides deeper insights into disease etiology by identifying relationships between molecular imaging and pathology data; disadvantages are that the methods can be computationally intensive and may require prohibitively large datasets to achieve accurate results. Transpathology approaches have been particularly successful in elucidating biological features of neurological diseases, with hundreds of studies connecting brain network activity from imaging studies with the likelihood of developing mental illness at the individual level (136). Recent work has also shown the benefit of integrating imaging data with expression data to learn more about disease. For example, trimodal transcriptomics, genomics, and imaging data were combined in a federated model to understand the relationships between omics levels and how these biologically important factors relate to AD (137). Similarly, Bao et al. (138) recently integrated genomic data, multiple types of expression data, imaging data from 145 brain regions, and AD status using a colocalization framework to distinguish causal AD pathways. Imaging analyses have also been implemented for spatial transcriptomics, enabling a more comprehensive view of molecular processes (139).

4.4. Precision Medicine Approaches with Multiomics Data Integration

Precision medicine represents a tailored approach to healthcare in which medical decisions, treatments, and interventions are specifically adapted to each individual based on their unique characteristics (140). Central to this vision is the integration of multiomics data: joining genomics, transcriptomics, proteomics, metabolomics, and other high-throughput datasets to obtain a holistic view of an individual's biological makeup. Integrating these vast datasets allows clinicians and researchers to identify complex patterns, making it possible to predict susceptibility to specific diseases, understand disease progression, and create personalized treatment strategies. In essence, multiomics integration is a key element that will transform the ambitious vision of precision medicine into a tangible reality, ushering in a new era of personalized and predictable healthcare.

With the incorporation of multiomics data and EHR-derived phenotypic data from large-scale biobanks, numerous studies have been undertaken to stratify or predict patient risk (49). Thompson et al. (141) explored the potential of epigenetic information to enhance phenotype inference in combined biobank–EHR systems by developing a methylation risk scoring model that integrated DNA methylation data and 607 EHR-derived phenotypes from the University of California, Los Angeles health biobank and demonstrated its potential to significantly improve clinical phenotype inferences compared to traditional polygenic risk scores. Talmor-Barkan et al. (142) integrated extensive clinical and multiomics profiling of 199 patients with acute coronary syndrome (ACS) to understand the multifaceted nature of CAD. By integrating serum

metabolomics, gut microbiome data, and data from two major Israeli hospitals, the researchers identified distinct serum and gut microbial signatures in ACS patients compared to controls: The former lacked a previously unidentified bacterial species from the Clostridiaceae family, the absence of which was connected to various circulating metabolites known to increase CAD risk. This finding emphasizes the personalized nature of metabolic deviations in ACS patients and their ties to important clinical factors and cardiovascular outcomes; it also underscores the potential early involvement of metabolic disturbances linked to the microbiome and diet in dysmetabolic phases that predate clinically apparent CAD. Furthermore, using a metabolomics-based model, the authors observed that ACS patients' predicted body mass index exceeded actual measurements, and these predictions correlated with diabetes mellitus and CAD severity. This study highlights the potential of the serum metabolome in helping researchers comprehend the diverse risk factors associated with CAD. In a different study, Parisot et al. (143) utilized graph convolutional networks (GCNs) to introduce a comprehensive framework that integrates both imaging and nonimaging data for brain analysis in large populations. In their approach, the population is represented as a sparse graph where nodes correspond to imaging-based feature vectors and edge weights to phenotypic data, including genetic information. Such network-based integration captures the interplay of individual features and their interactions holistically. They tested this method on two broad datasets: the Autism Brain Imaging Data Exchange (ABIDE) for predicting autism spectrum disorder and the Alzheimer's Disease Neuroimaging Initiative (ADNI) for predicting conversion to AD. They showed a remarkable performance improvement over existing techniques, achieving a classification accuracy of 70.4% for ABIDE and 80.0% for ADNI, highlighting the potential of neuroimaging and phenotypic data integration using GCNs to improve disease prediction. Mathew et al. (144) undertook a comprehensive immune profiling of hospitalized COVID-19 patients using high-dimensional flow cytometry. Their analysis at the single-cell omics level revealed three distinct immunotypes, each associated with differing degrees of disease severity and clinical outcomes, and the longitudinal analysis highlighted stability and fluctuations in patient responses. These findings not only offer a comprehensive map of immune cell responses in COVID-19 but also highlight potential avenues for therapeutic interventions.

Large language model (LLM): state-of-the-art AI model trained on extensive text data, adept at understanding and generating human language for various natural language processing tasks

5. FUTURE DIRECTIONS OF MULTIMODAL OMICS DATA ANALYSIS

In the previous sections, we discussed how the scale of data integration and the utilization of machine learning have revolutionized the landscape of multimodal omics data analysis, with a focus on advancing precision medicine. However, the rapid and large-scale development of state-of-the-art large language models (LLMs), including GPT-4 (145), PaLM2 (146), and Llama 2 (147), has introduced artificial intelligence methods with unprecedented and remarkable performance capabilities. These developments anticipate a new paradigm shift in multimodal omics data analysis. Accordingly, this section explores the trends, opportunities, and challenges that will emerge as the field embraces the potential of LLMs and other cutting-edge technologies to unlock the complexity of multimodal omics data for the benefit of precision medicine and biomedical research.

Supplemental Material >

Multimodal LLMs are designed to accommodate inputs of multiple modalities, extending beyond text-based data (**Table 4**; **Supplemental Table 1**). Notably, ChatGPT has profoundly influenced a number of fields, reshaping our interaction with technology. In addition to the increasing public interest in ChatGPT, as LLMs demonstrate remarkable achievements in the biomedical domain, numerous experimental methodologies applied to multimodal data are emerging, promising to open new avenues of discovery and innovation. In particular, building on established medical LLMs (148, 149), several multimodal medical LLMs have been proposed

Table 4 General-purpose and biomedical-specified large language models

Model(s)	Modalities	Foundation model(s)	Model size(s)	LLM fine-tuning
General-purpose multimodal LLMs				
Flamingo	Language, vision	Chinchilla, NFNets	3B, 9B, 80B	No
LLaVA	Language, vision	Vicuna/LLaMA2, CLIP ViT	7B, 13B	Yes
BLIP2	Language, vision	OPT/Flan-T5, CLIP ViT+Q-former	3B, 7B, 12B	No
MiniGPT4	Language, vision	Vicuna, CLIP ViT+Q-former	7B, 13B	No
InstructBLIP	Language, vision	BLIP2	7B, 13B	No
KOSMOS-1	Language, vision	Magnato, CLIP ViT	1.6B	Yes
PaLM E	Language, vision	PaLM, ViT	12B, 66B, 84B, 562B	Yes
mPLUG-OWL	Language, vision	LLaMA, CLIP ViT	7.2B	Yes
mPLUG- Doc OWL	Language, vision, chart, document	mPLUG-OWL	Unknown	Yes
GPT4	Language, vision, chart, document	NA	Unknown	Unknown
Medical multimodal LLMs				
LLaVA-Med	Language, medical vision (X-ray, MRI, pathology, gross pathology, CT)	LLaVA	7B, 13B	No
MedVInT	Language, medical vision (X-ray, MRI, pathology, gross pathology, CT)	PMC-CLIP, PMC-LLaMA	7B	Yes
Med-Flamingo	Language, medical vision (X-ray, MRI, pathology, gross pathology, CT)	OpenFlamingo	9B	Yes
Visual Med-Alpaca	Language, medical vision (X-ray, MRI, pathology, gross pathology, CT)	NA	7B	No
RadFM	Language, medical vision (X-ray, MRI, pathology, gross pathology, CT)	PMC-LLaMA	14B	Yes
PathAsst	Language, pathology	PLIP, Vicuna	13B	No
PaLM-Med M	Language, medical vision (X-ray, MRI, pathology, gross pathology, CT)	PaLM-Med	12B, 84B, 562B	No

References for each foundation model can be found in **Supplemental Appendix 3**. Abbreviations: CT, computed tomography; LLM, large language model; MRI, magnetic resonance imaging; NA, not applicable.

(150). These models are poised to revolutionize medical research and healthcare by enabling interactive data-driven communication, information extraction, and insights into model results. One key differentiator between LLMs and traditional machine learning is the ability of LLMs to engage in data-driven communication, extract information, and provide insights into the underlying reasons and basis for the model's results. In essence, as new multimodal omics data are generated and the analyzed knowledge is fed into LLMs, they can infer new insights based on previously accumulated data, further advancing our understanding of complex diseases and enabling more informed healthcare decision-making.

Supplemental Material >

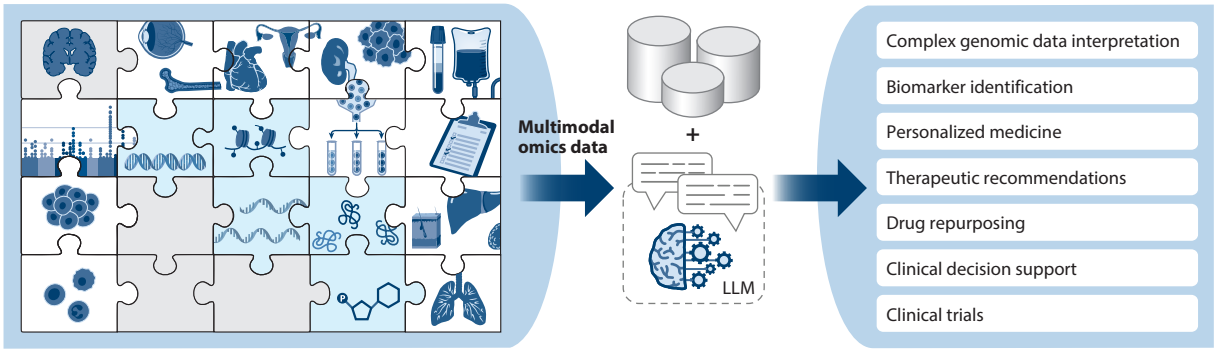


Figure 2

Potential ability of large language models (LLMs) with multimodal omics data. This illustrates how LLMs can be applied to synthesize and interpret multimodal omics data. LLMs with multimodal omics data can lead to improved interpretation of complex genomic data, identification of biomarkers, personalized medicine approaches, therapeutic recommendations, drug repurposing, support for clinical decision-making, and the design of clinical trials.

Supplemental Material >

We briefly describe general-purpose LLMs and biomedical-specific LLMs in **Table 2**. For example, LLaVA (Large Language and Vision Assistant) can pinpoint anomalies in unconventional images (**Supplemental Figure 1**). LLaVA-Med (Large Language and Vision Assistant for BioMedicine), when presented with a radiology image, does not just decode its features but synthesizes medical knowledge to diagnose (**Supplemental Figure 2**). As LLMs prove capable of such multimodal reasoning, we foresee vast potential in tackling complex challenges like multiomics analysis. A conceptual framework referred to as generalist medical AI (151) has been proposed with the aim of flexibly processing diverse medical modalities from imaging to genomics. Subsequently, Acosta et al. (152) underscored the potential of multimodal AI in harnessing a myriad of biomedical datasets, from expansive biobanks to cost-effective sequencing. We likewise anticipate that multimodal medical AI will steer the next era of multiomics analysis (**Figure 2**).

Recent endeavors have demonstrated the efficacy of ChatGPT in tasks like single-cell type prediction, with impressive prediction performance outcomes (153). The impressive accuracy comes from GPT-4's comprehensive training on scientific literature, which has equipped it to identify marker genes and their corresponding cellular associations. There is potential to leverage this capability for multiomics analysis interpretation. While there is a growing interest in employing foundation models, which serve as a base for a wide range of applications, for radiomics, and several multimodal LLMs can stride in this integration (**Table 4**), they are currently lacking in handling molecular-level omics data. For instance, MedGPT has been trained to analyze genomics, yet it still represents a proof-of-concept study with preliminary results. With technological improvements in the foundation model, advances in understanding molecular omics will enable researchers to become more adept at interpreting and integrating vast and diverse multiomics data, thereby overcoming the current limitations of multimodal LLMs. We provide a list of models trained on datasets like transcriptomes, DNA sequences, and protein sequences that may potentially underpin future multimodal LLM development (**Supplemental Table 1**). While forging a multimodal omics-centric LLM is a formidable challenge, we believe that harnessing such diverse modalities will redefine future research paradigms.

Foundation model: model pre-trained on a broad range of data at a large scale, designed to be adaptable to a variety of downstream tasks through further fine-tuning or conditional generation

6. CONCLUSION

As we navigate the intricate landscape of biological systems, the integrative analysis of multimodal omics data emerges as a cornerstone in decoding the complexities inherent in health and disease.

We have witnessed several significant transformations, notably the transition from mono-omics to multiomics analyses, driven by advances in data integration methods and the growing body of biological knowledge. In addition to these high-throughput multiomics techniques at the molecular level, a wealth of heterogeneous phenotypic data has been generated and derived from EHRs, which are closer to the patient's health status than biological information. They include clinical diagnoses, treatment histories, vital signs, and imaging data. The integration of such diverse phenotypic data with multimodal omics data can help uncover a piece of the puzzle in the human biological system through an understanding of the intricate interplay between molecular and clinical aspects of health and disease.

As technology advances over time, new biological information will be generated while integrated analysis methodologies will develop along with it. Due to the high dimensionality and heterogeneity of multiomics data, along with the problem of integrated data analysis, interpretation of diverse and intricate datasets will be a challenge in the future. Translating complex and heterogeneous omics data into meaningful clinical insights remains an arduous task, requiring advances in both computational approaches and biological understanding. Only when both integration and interpretation of multimodal data are successfully achieved will the path be paved for the next frontier in precision medicine.

The current transformative change in data analysis and interpretation is being led by LLMs. We all know that it is important to transfer biological insights obtained through complex multiomics analysis to clinical and medical contexts, but comprehensive analysis is not easily accomplished due to computational complexity and data heterogeneity. Let us imagine an LLM that is trained on all clinical information available from a biobank or hospital. We would be able to obtain actionable knowledge that could be converted to practical clinical use from molecular findings obtained through multiomics data integration analysis, as well as reasoning that serves as a basis for evidence. These innovative systems promise to redefine problem-solving in biomedical contexts and revolutionize personalized healthcare delivery, despite existing obstacles like resource constraints and data accessibility.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants R01 GM138597, R01 AG071470, and R01 HL169458.

LITERATURE CITED

1. Reuter JA, Spacek DV, Snyder MP. 2015. High-throughput sequencing technologies. *Mol. Cell* 58:586–97
2. Karczewski KJ, Snyder MP. 2018. Integrative omics for health and disease. *Nat. Rev. Genet.* 19:299–310
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
4. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–73
5. Karczewski KJ, Francioli LC. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581:434–43
6. Snyder MP, Gingeras TR, Moore JE. 2020. Perspectives on ENCODE. *Nature* 583:693–98

7. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 28:1045–48
8. Stunnenberg HG, Hirst M. 2016. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* 167:1145–49
9. Bae JB. 2013. Perspectives of International Human Epigenome Consortium. *Genom. Inform.* 11:7–14
10. Battle A, Brown CD, Engelhardt BE, Montgomery SB. 2017. Genetic effects on gene expression across human tissues. *Nature* 550:204–13
11. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45:580–85
12. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, et al. 2021. Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* 53:1300–10
13. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, et al. 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* 16:22
14. Abugessaisa I, Ramilowski JA, Lizio M, Severin J, Hasegawa A, et al. 2021. FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Res.* 49:D892–98
15. Alam T, Agrawal S, Severin J, Young RS, Andersson R, et al. 2020. Comparative transcriptomics of primary cells in vertebrates. *Genome Res.* 30:951–61
16. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506–11
17. 't Hoen PAC, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, et al. 2013. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* 31:1015–22
18. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22:1760–74
19. Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, et al. 2011. The Human Proteome Project: current state and future direction. *Mol. Cell Proteom.* 10:M111.009993
20. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, et al. 2015. The CPTAC Data Portal: a resource for cancer proteomics research. *J. Proteome Res.* 14:2707–13
21. Whiteaker JR, Halusa GN, Hoofnagle AN, Sharma V, MacLean B, et al. 2014. CPTAC Assay Portal: a repository of targeted proteomic assays. *Nat. Methods* 11:703–4
22. Tuck MK, Chan DW, Chia D, Godwin AK, Grizzle WE, et al. 2009. Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. *J. Proteome Res.* 8:113–17
23. Pietzner M, Wheeler E, Carrasco-Zanini J, Cortes A, Koprulu M, et al. 2021. Mapping the proteo-genomic convergence of human diseases. *Science* 374:eabj1541
24. Day N, Oakes S, Luben R, Khaw KT, Bingham S, et al. 1999. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br. J. Cancer* 80(Suppl. 1):95–103
25. Koprulu M, Carrasco-Zanini J, Wheeler E, Lockhart S, Kerrison ND, et al. 2023. Proteogenomic links to human metabolic diseases. *Nat. Metab.* 5:516–28
26. Wishart DS, Guo A, Oler E, Wang F, Anjum A, et al. 2022. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* 50:D622–31
27. Carayol M, Leitzmann MF, Ferrari P, Zamora-Ros R, Achaintre D, et al. 2017. Blood metabolic signatures of body mass index: a targeted metabolomics study in the EPIC cohort. *J. Proteome Res.* 16:3137–46
28. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The Human Microbiome Project. *Nature* 449:804–10
29. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, et al. 2019. The Integrative Human Microbiome Project. *Nature* 569:641–48
30. Lopera-Maya EA, Kurilshikov A, van der Graaf A, Hu S, Andreu-Sánchez S, et al. 2022. Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. *Nat. Genet.* 54:143–51

31. Reel PS, Reel S, van Kralingen JC, Langton K, Lang K, et al. 2022. Machine learning for classification of hypertension subtypes using multi-omics: a multi-centre, retrospective, data-driven study. *EBioMedicine* 84:104276
32. Guo L, Zhong MB, Zhang L, Zhang B, Cai D. 2022. Sex differences in Alzheimer's disease: insights from the multiomics landscape. *Biol. Psychiatry* 91:61–71
33. Maitre L, Bustamante M, Hernández-Ferrer C, Thiel D, Lau CE, et al. 2022. Multi-omics signatures of the human early life exposome. *Nat. Commun.* 13:7024
34. Watanabe K, Wilmanski T, Diener C, Earls JC, Zimmer A, et al. 2023. Multiomic signatures of body mass index identify heterogeneous health phenotypes and responses to a lifestyle intervention. *Nat. Med.* 29:996–1008
35. Beesley LJ, Salvatore M, Fritsche LG, Pandit A, Rao A, et al. 2020. The emerging landscape of health research based on biobanks linked to electronic health records: existing resources, statistical challenges, and potential opportunities. *Stat. Med.* 39:773–800
36. Guo LY, Wu AH, Wang YX, Zhang LP, Chai H, Liang XF. 2020. Deep learning-based ovarian cancer subtypes identification using multi-omics data. *BioData Min.* 13:10
37. Qiao J, Wu Y, Zhang S, Xu Y, Zhang J, et al. 2023. Evaluating significance of European-associated index SNPs in the East Asian population for 31 complex phenotypes. *BMC Genom.* 24:324
38. Chatsirisupachai K, Lesluyes T, Paraoan L, Van Loo P, de Magalhães JP. 2021. An integrative analysis of the age-associated multi-omic landscape across cancers. *Nat. Commun.* 12:2345
39. Marabita F, James T, Karhu A, Virtanen H, Kettunen K, et al. 2022. Multiomics and digital monitoring during lifestyle changes reveal independent dimensions of human biology and health. *Cell Syst.* 13:241–55.e7
40. Perakakis N, Yazdani A, Karniadakis GE, Mantzoros C. 2018. Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism* 87:A1–9
41. Jiang MZ, Aguet F, Ardlie K, Chen J, Cornell E, et al. 2023. Canonical correlation analysis for multi-omics: application to cross-cohort analysis. *PLOS Genet.* 19:e1010517
42. Zhao H, Rasheed H, Nøst TH, Cho Y, Liu Y, et al. 2022. Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. *Cell Genom.* 2:100195
43. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, et al. 2012. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148:1293–307
44. Li-Pook-Than J, Snyder M. 2013. iPOP goes the world: integrated personalized omics profiling and the road toward improved health care. *Chem. Biol.* 20:660–66
45. Sanna S, van Zuydam NR, Mahajan A, Kurilshikov A, Vich Vila A, et al. 2019. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* 51:600–5
46. Sveinbjornsson G, Ulfarsson MO, Thorolfsdottir RB, Jonsson BA, Einarsson E, et al. 2022. Multiomics study of nonalcoholic fatty liver disease. *Nat. Genet.* 54:1652–63
47. Ritchie SC, Surendran P, Karthikeyan S, Lambert SA, Bolton T, et al. 2023. Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. *Sci. Data* 10:64
48. Di Angelantonio E, Thompson SG, Kaptoge S, Moore C, Walker M, et al. 2017. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet* 390:2360–71
49. Xu Y, Ritchie SC, Liang Y, Timmers P, Pietzner M, et al. 2023. An atlas of genetic scores to predict multi-omic traits. *Nature* 616:123–31
50. Surendran P, Stewart ID, Au Yeung VPW, Pietzner M, Raffler J, et al. 2022. Rare and common genetic determinants of metabolic individuality and their effects on human health. *Nat. Med.* 28:2321–32
51. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, et al. 2017. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* 49:139–45
52. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, et al. 2017. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* 49:131–38

53. Niehues A, Bizzarri D, Reinders MJT, Slagboom PE, van Gool AJ, et al. 2022. Metabolomic predictors of phenotypic traits can replace and complement measured clinical variables in population-scale expression profiling studies. *BMC Genom.* 23:546
54. Sijtsma A, Rienks J, van der Harst P, Navis G, Rosmalen JGM, Dotinga A. 2022. Cohort profile update: lifelines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* 51:e295–302
55. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, et al. 2016. The effect of host genetics on the gut microbiome. *Nat. Genet.* 48:1407–12
56. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, et al. 2015. Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* 44:1137–47
57. Aasmets O, Krigul KL, Lüll K, Metspalu A, Org E. 2022. Gut metagenome associations with extensive digital health data in a volunteer-based Estonian microbiome cohort. *Nat. Commun.* 13:869
58. Chen Z, Chen J, Collins R, Guo Y, Peto R, et al. 2011. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* 40:1652–66
59. Xu F, Yu EY, Cai X, Yue L, Jing LP, et al. 2023. Genome-wide genotype-serum proteome mapping provides insights into the cross-ancestry differences in cardiometabolic disease susceptibility. *Nat. Commun.* 14:896
60. Feng YA, Chen CY, Chen TT, Kuo PH, Hsu YH, et al. 2022. Taiwan Biobank: a rich biomedical research database of the Taiwanese population. *Cell Genom.* 2:100197
61. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, et al. 2017. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* 27:S2–8
62. Tadaka S, Hishinuma E, Komaki S, Motoike IN, Kawashima J, et al. 2021. jMorp updates in 2020: large enhancement of multi-omics data resources on the general Japanese population. *Nucleic Acids Res.* 49:D536–44
63. Kuriyama S, Yaegashi N, Nagami F, Arai T, Kawaguchi Y, et al. 2016. The Tohoku Medical Megabank Project: design and mission. *J. Epidemiol.* 26:493–511
64. Kim Y, Han BG. 2017. Cohort profile: the Korean genome and epidemiology study (KoGES) consortium. *Int. J. Epidemiol.* 46:e20. Erratum. 2017. *Int. J. Epidemiol.* 46:1350
65. Hahn SJ, Kim S, Choi YS, Lee J, Kang J. 2022. Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: a machine learning analysis of population-based 10-year prospective cohort study. *EBioMedicine* 86:104383
66. Jang HB, Hwang JY, Park JE, Oh JH, Ahn Y, et al. 2014. Intake levels of dietary polyunsaturated fatty acids modify the association between the genetic variation in PCSK5 and HDL cholesterol. *J. Med. Genet.* 51:782–88
67. Lee W, Lee HJ, Jang HB, Kim HJ, Ban HJ, et al. 2018. Asymmetric dimethylarginine (ADMA) is identified as a potential biomarker of insulin resistance in skeletal muscle. *Sci. Rep.* 8:2133
68. Njunge JM, Tickell K, Diallo AH, Sayeem Bin Shahid ASM, Gazi MA, et al. 2022. The Childhood Acute Illness and Nutrition (CHAIN) Network Nested Case-Cohort Study protocol: a multi-omics approach to understanding mortality among children in sub-Saharan Africa and South Asia. *Gates Open. Res.* 6:77
69. Saw WY, Tantoso E, Begum H, Zhou L, Zou R, et al. 2017. Establishing multiple omics baselines for three Southeast Asian populations in the Singapore Integrative Omics Study. *Nat. Commun.* 8:653
70. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, et al. 2014. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* 42:D975–79
71. Freeberg MA, Fromont LA, D'Altri T, Romero AF, Ciges JI, et al. 2022. The European Genome-phenome Archive in 2021. *Nucleic Acids Res.* 50:D980–87
72. Campagna MP, Xavier A, Lechner-Scott J, Maltby V, Scott RJ, et al. 2021. Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clin. Epigenetics* 13:214
73. Huckins LM, Dobbyn A, Ruderfer DM, Hoffman G, Wang W, et al. 2019. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nat. Genet.* 51:659–74
74. Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, et al. 2018. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* 362:eaat8127
75. Brandes N, Linial N, Linial M. 2020. PWAS: proteome-wide association study-linking genes and phenotypes by functional variation in proteins. *Genome Biol.* 21:173

76. Kojouri M, Pinto R, Mustafa R, Huang J, Gao H, et al. 2023. Metabolome-wide association study on physical activity. *Sci. Rep.* 13:2374
77. Wingo TS, Liu Y, Gerasimov ES, Gockley J, Logsdon BA, et al. 2021. Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat. Neurosci.* 24:810–17
78. Vasaikar SV, Savage AK, Gong Q, Swanson E, Talla A, et al. 2023. A comprehensive platform for analyzing longitudinal multi-omics data. *Nat. Commun.* 14:1684
79. Baysoy A, Bai Z, Satija R, Fan R. 2023. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* 24:695–713
80. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, et al. 2015. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* 12:519–22
81. Macaulay IC, Ponting CP, Voet T. 2017. Single-cell multiomics: multiple measurements from single cells. *Trends Genet.* 33:155–68
82. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, et al. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14:865–68
83. Choi JR, Yong KW, Choi JY, Cowie AC. 2020. Single-cell RNA sequencing and its combination with protein and DNA analyses. *Cells* 9:1130
84. Swanson E, Lord C, Reading J, Heubeck AT, Genge PC, et al. 2021. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* 10:e63632
85. Mimitou EP, Lareau CA, Chen KY, Zorzetto-Fernandes AL, Hao Y, et al. 2021. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* 39:1246–58
86. Darwiche R, Struhl K. 2020. Pheno-RNA, a method to associate genes with a specific phenotype, identifies genes linked to cellular transformation. *PNAS* 117:28925–29
87. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. 2021. Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol. Adv.* 49:107739
88. Vahabi N, Michailidis G. 2022. Unsupervised multi-omics data integration methods: a comprehensive review. *Front. Genet.* 13:854752
89. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. 2020. Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* 14:1177932219899051
90. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, et al. 2016. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinform.* 17(Suppl. 2):15
91. Kang M, Ko E, Mersha TB. 2022. A roadmap for multi-omics data integration using deep learning. *Brief. Bioinform.* 23:bbab454
92. Leal LG, David A, Jarvelin MR, Sebert S, Männikkö M, et al. 2019. Identification of disease-associated loci using machine learning for genotype and network data integration. *Bioinformatics* 35:5182–90
93. Wang T, Shao W, Huang Z, Tang H, Zhang J, et al. 2021. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* 12:3445
94. Wang C, Lue W, Kaalia R, Kumar P, Rajapakse JC. 2022. Network-based integration of multi-omics data for clinical outcome prediction in neuroblastoma. *Sci. Rep.* 12:15425
95. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, et al. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11:333–37
96. Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestøl K, et al. 2020. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* 395:350–60
97. Zhao B, Li T, Yang Y, Wang X, Luo T, et al. 2021. Common genetic variation influencing human white matter microstructure. *Science* 372:eabf3736
98. Zhao B, Luo T, Li T, Li Y, Zhang J, et al. 2019. Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat. Genet.* 51:1637–44
99. Zhao B, Li T, Smith SM, Xiong D, Wang X, et al. 2022. Common variants contribute to intrinsic human brain functional networks. *Nat. Genet.* 54:508–17
100. Zhao B, Li T, Fan Z, Yang Y, Shu J, et al. 2023. Heart-brain connections: phenotypic and genetic insights from magnetic resonance images. *Science* 380:abn6598

101. Zhao B, Li Y, Fan Z, Wu Z, Shu J, et al. 2023. Eye-brain connections revealed by multimodal retinal and brain imaging genetics in the UK Biobank. medRxiv 2023.02.16.23286035. **https://doi.org/10.1101/2023.02.16.23286035**
102. Alipanahi B, Hormozdiari F, Behsaz B, Cosentino J, McCaw ZR, et al. 2021. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *Am. J. Hum. Genet.* 108:1217–30
103. Li Y, Wu X, Yang P, Jiang G, Luo Y. 2022. Machine learning for lung cancer diagnosis, treatment, and prognosis. *Genom. Proteom. Bioinform.* 20:850–66
104. Xu M, Calhoun V, Jiang R, Yan W, Sui J. 2021. Brain imaging-based machine learning in autism spectrum disorder: methods and applications. *J. Neurosci. Methods* 361:109271
105. Liang X, Fu Y, Cao WT, Wang Z, Zhang K, et al. 2022. Gut microbiome, cognitive function and brain structure: a multi-omics integration analysis. *Transl. Neurodegener.* 11:49
106. Bodein A, Scott-Boyer MP, Perin O, Lê Cao K-A, Droit A. 2022. timeOmics: an R package for longitudinal multi-omics data integration. *Bioinformatics* 38:577–79
107. Metwally AA, Zhang T, Wu S, Kellogg R, Zhou W, et al. 2022. Robust identification of temporal biomarkers in longitudinal omics studies. *Bioinformatics* 38:3802–11
108. Ang JS, Ng KW, Chua FF. 2020. *Modeling time series data with deep learning: a review, analysis, evaluation and future trend*. Paper presented at the 8th International Conference on Information Technology and Multimedia (ICIMU), Selangor, Malaysia. **https://ieeexplore.ieee.org/document/9243546**
109. Choi K, Yi J, Park C, Yoon S. 2021. Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. *IEEE Access* 9:120043–65
110. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436–44
111. Bengio Y, Simard P, Frasconi P. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5:157–66
112. Lee G, Nho K, Kang B, Sohn KA, Kim D. 2019. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci. Rep.* 9:1952
113. Nguyen M, He T, An L, Alexander DC, Feng J, Yeo BTT. 2020. Predicting Alzheimer's disease progression using deep recurrent neural networks. *Neuroimage* 222:117203
114. Jung W, Jun E, Suk HI. 2021. Deep recurrent model for individualized prediction of Alzheimer's disease progression. *Neuroimage* 237:118143
115. Zhao B, Shan Y, Yang Y, Yu Z, Li T, et al. 2021. Transcriptome-wide association analysis of brain structures yields insights into pleiotropy with complex neuropsychiatric traits. *Nat. Commun.* 12:2878
116. Hu X, Qiao D, Kim W, Moll M, Balte PP, et al. 2022. Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program. *Am. J. Hum. Genet.* 109:857–70
117. Liang Y, Pividori M, Manichaikul A, Palmer AA, Cox NJ, et al. 2022. Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries. *Genome Biol.* 23:23
118. Mattheisen M, Grove J, Als TD, Martin J, Voloudakis G, et al. 2022. Identification of shared and differentiating genetic architecture for autism spectrum disorder, attention-deficit hyperactivity disorder and case subgroups. *Nat. Genet.* 54:1470–78
119. Benchek P, Igo RP Jr., Voss-Hoynes H, Wren Y, Miller G, et al. 2021. Association between genes regulating neural pathways for quantitative traits of speech and language disorders. *NPJ Genom. Med.* 6:64
120. Li L, Chen Z, von Scheidt M, Li S, Steiner A, et al. 2022. Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes. *Basic Res. Cardiol.* 117:6
121. Mooney MA, Ryabinin P, Wilmot B, Bhatt P, Mill J, Nigg JT. 2020. Large epigenome-wide association study of childhood ADHD identifies peripheral DNA methylation associated with disease and polygenic risk burden. *Transl. Psychiatry* 10:8
122. Hesam-Shariati S, Overs BJ, Roberts G, Toma C, Watkeys OJ, et al. 2022. Epigenetic signatures relating to disease-associated genotypic burden in familial risk of bipolar disorder. *Transl. Psychiatry* 12:310
123. Sekula P, Goek ON, Quaye L, Barrios C, Levey AS, et al. 2016. A metabolome-wide association study of kidney function and disease in the general population. *J. Am. Soc. Nephrol.* 27:1175–88

124. Osborn MP, Park Y, Parks MB, Burgess LG, Uppal K, et al. 2013. Metabolome-wide association study of neovascular age-related macular degeneration. *PLOS ONE* 8:e72737
125. Dehghan A, Pinto RC, Karaman I, Huang J, Durainayagam BR, et al. 2022. Metabolome-wide association study on ABCA7 indicates a role of ceramide metabolism in Alzheimer's disease. *PNAS* 119:e2206083119
126. Ge A, Sun Y, Kiker T, Zhou Y, Ye K. 2023. A metabolome-wide Mendelian randomization study prioritizes potential causal circulating metabolites for multiple sclerosis. *J. Neuroimmunol.* 379:578105
127. Vasaikar SV, Straub P, Wang J, Zhang B. 2018. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46:D956–63
128. Khadirnaikar S, Shukla S, Prasanna SRM. 2023. Machine learning based combination of multi-omics data for subgroup identification in non-small cell lung cancer. *Sci. Rep.* 13:4636
129. Malik V, Kalakoti Y, Sundar D. 2021. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genom.* 22:214
130. Dimitrakopoulos C, Hindupur SK, Häfliger L, Behr J, Montazeri H, et al. 2018. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* 34:2441–48
131. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, et al. 2016. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* 535:94–103
132. Perez-Garcia J, Espuela-Ortiz A, Hernández-Pérez JM, González-Pérez R, Poza-Guedes P, et al. 2023. Human genetics influences microbiome composition involved in asthma exacerbations despite inhaled corticosteroid treatment. *J. Allergy Clin. Immunol.* 152:799–806.e6
133. Dai H, Hou T, Wang Q, Hou Y, Wang T, et al. 2023. Causal relationships between the gut microbiome, blood lipids, and heart failure: a Mendelian randomization analysis. *Eur. J. Prev. Cardiol.* 30:1274–82
134. Li Z, Zhu G, Lei X, Tang L, Kong G, et al. 2023. Genetic support of the causal association between gut microbiome and COVID-19: a bidirectional Mendelian randomization study. *Front. Immunol.* 14:1217615
135. Tian M, He X, Jin C, He X, Wu S, et al. 2021. Transpathology: molecular imaging-based pathology. *Eur. J. Nucl. Med. Mol. Imaging* 48:2338–50
136. Rashid B, Calhoun V. 2020. Towards a brain-based predictome of mental illness. *Hum. Brain Mapp.* 41:3468–535
137. Wu J, Chen Y, Wang P, Caselli RJ, Thompson PM, et al. 2021. Integrating transcriptomics, genomics, and imaging in Alzheimer's disease: a federated model. *Front. Radiol.* 1:777030
138. Bao J, Wen J, Wen Z, Yang S, Cui Y, et al. 2023. Brain-wide genome-wide colocalization study for integrating genetics, transcriptomics and brain morphometry in Alzheimer's disease. *Neuroimage* 280:120346
139. Bergenstråhle J, Larsson L, Lundeberg J. 2020. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genom.* 21:482
140. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, et al. 2021. Precision medicine, AI, and the future of personalized health care. *Clin. Transl. Sci.* 14:86–93
141. Thompson M, Hill BL, Rakocz N, Chiang JN, Geschwind D, et al. 2022. Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. *NPJ Genom. Med.* 7:50
142. Talmor-Barkan Y, Bar N, Shaul AA, Shahaf N, Godneva A, et al. 2022. Metabolomic and microbiome profiling reveals personalized risk factors for coronary artery disease. *Nat. Med.* 28:295–302
143. Parisot S, Ktena SI, Ferrante E, Lee M, Guerrero R, et al. 2018. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med. Image Anal.* 48:117–30
144. Mathew D, Giles JR, Baxter AE, Oldridge DA, Greenplate AR, et al. 2020. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science* 369:eabc8511
145. OpenAI. 2023. GPT-4 technical report. arXiv:2303.08774 [cs.CL]
146. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, et al. 2023. PaLM 2 technical report. arXiv:2305.10403 [cs.CL]
147. Touvron H, Martin L, Stone KR, Albert P, Almahairi A, et al. 2023. Llama 2: open foundation and fine-tuned chat models. arXiv:2307.09288 [cs.CL]

148. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. 2023. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus* 15:e40895
149. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, et al. 2023. Large language models encode clinical knowledge. *Nature* 620:172–80
150. Li C, Wong C, Zhang S, Usuyama N, Liu H, et al. 2023. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. arXiv:2306.00890 [cs.CV]
151. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, et al. 2023. Foundation models for generalist medical artificial intelligence. *Nature* 616:259–65
152. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. 2022. Multimodal biomedical AI. *Nat. Med.* 28:1773–84
153. Hou W, Ji Z. 2023. Reference-free and cost-effective automated cell type annotation with GPT-4 in single-cell RNA-seq analysis. bioRxiv 2023.04.16.537094. **https://doi.org/10.1101/2023.04.16.537094**