Deep Science Publishing
https://doi.org/10.70593/978-93-49307-76-6_5

DeepScience
Open Access Books

# Chapter 5: Real-time integration of multi-omics data: Leveraging big data pipelines for holistic health insights

## 5.1 Introduction

With rapid technological advancements, unprecedented data availability, and intrinsic interactivity among different levels of molecular data, the real-time integration of multi-omics data has now promised big potentials for a better understanding of individual health profiles and subsequently an early provision of precise and cost-efficient preventative and care interventions. Enhanced systematic and computational methods coupled with ultra-fast big data pipelines provide the essential power for meaningful health insights, ultimately translating into improved healthcare outcomes. In the next few years, integrated multi-omics data and enhanced big data pipelines are expected to be a significant and essential driving force of holistic insights into individual human health and wellbeing. Real-time integration of multi-omics data has recently provided a paradigm to dissect human diseases. A diverse range of health traits can be characterized more thoroughly through multi-omics data integration, generating a comprehensive profile of health status that facilitates the early detection of potential ailments.

Between 2013 and 2018, the global volume of collected data increased from half to up to 33 zettabytes. Individuals are wearing devices that monitor their health-generated data in real-time, and AI algorithms are being utilized to provide diagnoses and to suggest treatments. These advances derive primarily from the advent of big data pipelines, along with the growing ubiquity of sensors and devices. As a result, real-time health insights are now more accessible than ever before. The goal of multi-omics data is to provide a comprehensive holistic view of the different levels of molecules jointly involved in the regulation of a biological system. Even more, an integrated data-analysis is hard to be performed in real-time, owing to the characteristic high dimensionality of multi-omics data. However, even in the absence of integration, each single level of omics data is

informative on its own, providing a partial view of health status. In many conditions, the molecular changes that occur at a particular level of the omics cascade are followed by downstream modifications. As a consequence, the monitoring of intermediate or affected molecular levels with respect to the pathological start might offer better prevention or treatment opportunities. Thus, a device equipped with a fast omics data readout should be able to run in real-time analytical methods to convert the data into insights, anticipating the possible occurrence of the disease and suggesting possible treatments.

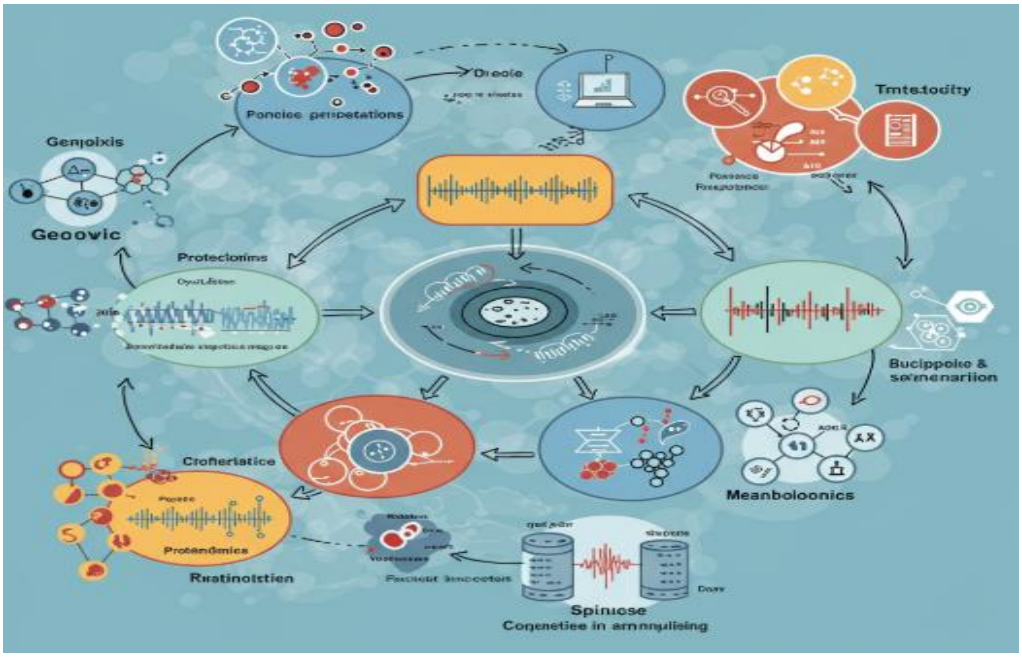


**Fig 5.1:** Real-Time Integration of Multi-Omics Data

## 5.2. Understanding Multi-Omics

Imagine a world where big data pipelines leverage advanced algorithmic methods and novel sequence-based technologies for the real-time and comprehensive analyses of an individual's whole genome, epigenome, transcriptome, proteome, and metabolome. In this world, clinical doctors interpret these analyses on a browser-based platform that seamlessly reads and stores all necessary information with powerful visualizations provided automatically. With a simple blood test, millions of measurements are taken within minutes and stored in a cloud infrastructure for an automated pipeline analysis. With these results, millions of possible disparities from diseased population-specific data are queried and a live report provides a few prioritized, easy-to-understand graphics. These graphics highlight the disparities and their association with disease symptoms, as well as recommendations for preventative steps based on the trivial pharmacological

impact of small metabolite deficiency compensations. An appointment call is placed to the next available health consultant who has instant access to this report, and these results drive a shared decision-making conversation that aims to discover deeper sub-clinical concerns. This imaginative reality, lying only slightly beyond the current forefront of the available technology, is enabled by recently developed and fully open-source software infrastructure for the real-time integration and interpretation of multi-omics, an approach characterized by the holistic and dynamical study of multiple data layers and across different types of omics using high-throughput analytical technologies. The field of genomic health research has long been driving significant advancements in collection, storage, analysis, and interpretation techniques of genetic information. Genomic data is the foundation of biological data and provides a Window to the most static representation of biological functions: the biomolecules encoded by the genome, mainly the proteins and RNA. However, only understanding the genome is limiting because it is a more generic blueprint describing the potentiality of a biological entity. To fully unveil the components and dynamics of biological systems (including healthy and diseased states) one typically requires so-called readouts from additional, interlinked layers that can either be produced by the genome or have a reverse, regulatory impact on it. The latter is the motivating theme of this topic, discussing the field of multi-omics through a comprehensive understanding of the term, its relevance to larger health outcomes as opposed to genetic ones, an overview of the diversity of omics types understanding biological systems, and recent advances in genomic technologies and informatics currently reshaping the collection and integration methodology of such data.

### 5.2.1. Definition and Scope

The term multi-omics refers to the comprehensive analysis of an individual's multiple omics data layers. A variety of distinct omics disciplines have emerged in the study of complex biological systems. These include, but are not limited to, gene- or transcriptomics, protein- or proteomics, and metabolite- or metabolomics. Each individual set of data provides a view of intricate biological processes but lacks context outside of its own domain. Although the majority of health analytic research has sought to leverage one particular omics discipline at a time, these various data layers are tightly linked and exert a larger influence on health outcomes when integrated. Due to technological advances, different layers feed into one another; proteins are encoded by genes, and metabolites are products of proteins, etc. Different omics disciplines are thus interdependent, and health insights are much more accurate and informative when based on this integration. Additionally, newly-developed omics approaches, such as the study of epitranscriptomics or glycomics, can be quite powerful, providing more comprehensive insight into the human health component. It is necessary, therefore, to put a sufficient focus on the emerging methodologies that facilitate the integration and

joint analysis of these various data types, which currently span a multitude of standard and cutting-edge technologies (Kaulwar, 2023; Koppolu, 2022).

This work is limited to the collection of multi-omics data, paying inquiring attention to methods of integrative and post-integrative analyses. It aims to fully grasp the scope and complexity of the topic as a means of sharing the findings with a wider community of health researchers and practitioners. The separate sections define multi-omics and take a closer look at its position within the broader health research landscape. A concrete definition of multi-omics is a nascent field, and it follows that the convergence of different multi-omics methodologies must be classified as well.

### 5.2.2. Types of Omics Data

Health research technologies have come a long way over the years. New research fields will contribute to a better understanding of what happens in our bodies and enable many illnesses to be diagnosed and treated at a much earlier stage. Some of these technologies have already been around for a few years but, with the advent of big data processing frameworks they can be leveraged for the analysis of big data in health research. This not only refers to each "omics" - genomics, transcriptomics, proteomics, metabolomics, etc. but also to multi-omics integration, for instance.

The genomics discipline examines the structure, function, and variations of genes, providing a foundation for the understanding of genetic influences on health. Traditionally, a genome refers to an individual's DNA sequence, which contains genetic information in the form of the combination of the four nucleotides - adenine (A), cytosine (C), guanine (G), and thymine (T). However, the significance of a single genetic variation is often limited to particular gene products. This variation across genotypes, cells, tissues, individuals, and life stages can contribute to the complex phenotypic variability in populations. Hence, the structural and functional annotations of the genome are vital. The discipline that studies the genome organization and annotation, and its interaction with environmental factors is genomics.

Transcriptomics is the comprehensive study of RNA entities transcribed from the genes of an organism. RNA entities are intermediary molecules in the flow of the genetic information between the DNA sequences and the amino acid sequences. They form functionally diverse classes of molecules, and are abundantly produced in the human cell. During the process of transcription a segment of nuclear DNA is copied into RNA molecules. There are several types of RNA molecules, including messenger RNAs (mRNAs), transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), as well as other non-coding RNA molecules, serving important transportation and regulatory functions. The discipline that studies all RNA entities, their levels, structure and function, following

differential protein encoding over time and under different conditions, is termed transcriptomics.

## 5.3. Real-Time Integration of Multi-Omics Data

In the age of big data, the world is generating information on an unprecedented scale. This includes health data, with trends pointing towards the generation of x-bytes of genomics, proteomics and other health-related data. But simply generating data is not enough – it must be accessible in real-time and in a convenient, immediately usable format for maximum effectiveness. Real-time accessibility of data and the novel information given is the quintessence of big data in applications resulting from or concerning dealing with data. The rationale for this is obvious: there is an advantage to having all possible information at the time of making decisions. For example, in the clinical case, data-driven clinical decision support systems can only be built on the basis of all clinical and biological data accessible at a given time. In the case of predictive modeling, the ability to use streamed data comes from the desire to make non-intrusive and early interventions. It is recognized that having easier access to data can improve existing pipelines and in some noisy cases even replace expert analytics. Finally, in the more specific case of health monitoring, more efficient and potential life-saving analytical procedures can be established if abnormalities detected by algorithms are given proper interpretations or even seen in the context of other available data.

### 5.3.1. Current Trends in Health Data

The paper gives an outline of state-of-the-art in the evolution of health data landscapes and the increased complexity and number of data sources characterizing their present form. With the proliferation of electronic health records (EHRs), confidential research initiatives, and big data applications, the near future availability to researchers of the enormous amount of sensitive, de-identified health and healthcare data en masse, opens unprecedented opportunities for the implementation of cutting-edge methods and algorithms to shift the paradigm of healthcare delivery and medical research. It provides an overview of the shift towards patient-generated health data (PGHD), including, among the others, the recent diffusion of wearable devices and the early initiatives launched by the major pharmaceutical companies. The result of the simultaneous increase in health data awareness in the general public and the rapid developments in the field of '-omics' technologies, is the new data-rich landscape of health and healthcare in the beginning of the twenty-first century. In the framework of the envisioned paradigm shift, data accumulation unveils new challenges concerning the underlying methods and technologies for data collection, storage, transfer, quality control, analysis, and display.

For instance, here is the emerging evidence on the effectiveness of digitally delivered interventions for numerous conditions including asthma, type 2 diabetes, smoking cessation, ischemic heart disease and depression. Although a full development of big data applications is still pending, currently these have become a buzz-term in economy and research paradigms, and this is poignantly true for clinical and healthcare settings. Technologies based on advanced analytics – in the form of machine learning (ML) and artificial intelligence (AI) – will play an important role in handling and drawing insights from these large data sets. Importantly, such methods have been shown to significantly improve the predictive performances of the models. Similarly, it allows real-time monitoring of health data to generate suggestions and acts as a facilitator for action-based interventions and will subsequently improve the patient outcome. Technological advancements can significantly accelerate this evolution.

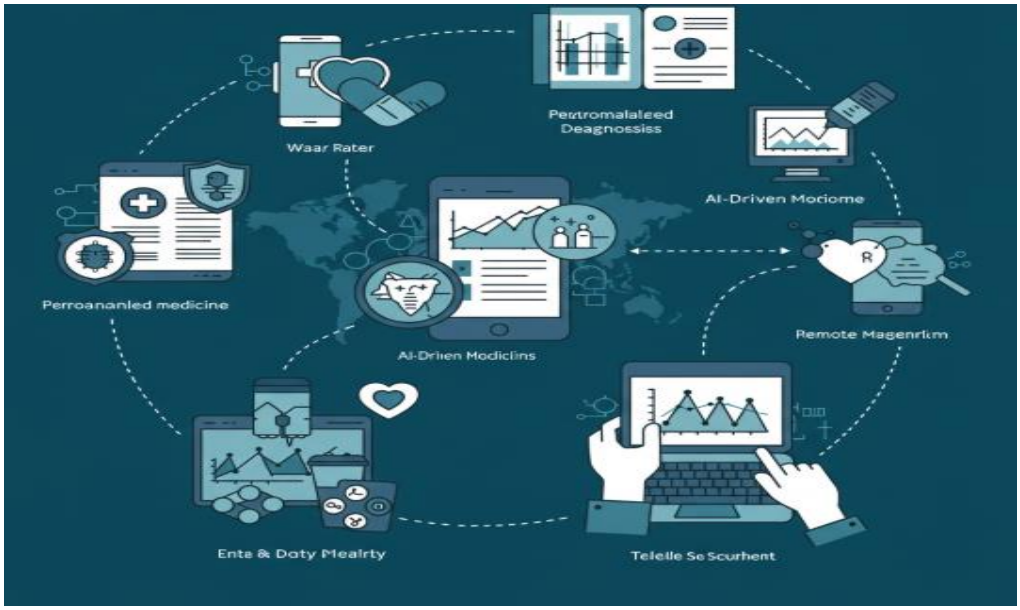


**Fig 5.2:** Current Trends in Health Data

### 5.3.2. Challenges in Real-Time Integration

Introduction of multi-omics approaches adds new opportunities to investigate health status holistically, as reflected by individual molecular profiles such as genomics, epigenomics, proteomics, transcriptomics, and metabolomics. However, integration of multi-omics types and data formats as well as connection of clinical information, lifestyle data, or environmental data requires advanced analytic technology, as data acquisition, generation, handling, and interpretation is complex, time sensitive, and resource intensive (Singireddy, J., 2024; Singireddy, S., 2024).

Recent technological developments such as big data IT arrangements, fast and robust data processing, and analysis methods in combination with increasing computer resources provide new opportunities to analyze such large datasets. Novel approaches to analyze patterns in large-scale data in the context of holistic research are discussed. Real-time data integration technologies in health research and care are addressed. The partnership of academia with healthcare facilities and industry data research units has increased, as reflected in an increase in the number of national activities and international alliances in the field. It is expected that robust ethical, legal, and social frameworks and standards will become established for smooth data integration and high-level analytics providing balance of safety and data integrity between security agencies, services, technology providers, and research institutions.

Health research and clinical practice require effective infrastructure and appropriate methodology for the real-time analysis and interpretation of big and complex multi-variety datasets. Modern hospitals are customized for professional work, but the rigid slow processes and the red tape of traditional governance do not allow the quick implementation of rapidly changing technologies including new IT solutions, fast communication, and large file safe data sharing. On the other hand, a lot of sensitive data are produced by variables that can be measured in the care process; this data is usually stored in proprietary electronic clinical databases, separated from the outside world. Thus, a huge amount of data are not usable or are underused. The utilisation of administrative databases is complicated by the various structures, different data order, data collection processes, and meanings across patient groups or locations.

### 5.4. Big Data Pipelines in Health Research

Health research is at a historical pinnacle because of the ability to collect and store orders of magnitude more data than in decades past. This explosion of information goes hand in hand with an onslaught of new analytic techniques, meaning that the right data, when processed appropriately, can yield life-saving information never before accessible. Taking full advantage of this wealth of data starts with efficient tools for data management. In the age of cloud computing, huge numbers of processors can be scaled up in seconds, and cheap, near-infinite storage is available to anyone with a web browser. But how does one effectively use these tools to store, analyze, and ultimately provide insight on massive multi-omics datasets? Big data technologies have emerged to meet this challenge. Fueled by the aforementioned advances in computing power and storage, big data technologies have revolutionized data management through data storage and computational processing solutions capable of managing datasets exceeding the limits of traditional data processing systems.

Structured data processing frameworks represent the backbone of big data technologies and are essential for the effective integration and analysis of large datasets. By defining the format of macroscopic data analysis, these frameworks are a core component of big data technologies. They allow immense datasets to be dissected and analyzed in very specific ways, for defining patterns and obtaining insights altogether impossible with manual investigation. The transformation from raw data into, ostensibly more valuable, processed data, involves data processing frameworks as a necessary middle man. Data processing frameworks define the macroscopic interaction with the data, as well as the means and method of analysis. Only by defining how to ingest, store, and transform the data via the structured framework, can the data truly be analyzed, and insights derived. This section serves to describe, at a high level, the components of a data processing framework, with the intention of providing a big picture of the chain of events allowing data to seamlessly flow from the point of data collection to the derivation of insights.

Big data pipelines represent a means of orchestrating the sequence of events by which data is transformed into consumable insights. To better convey the utility of big data pipelines, and facilitate a high-level BPOLDMA-based understanding, the chain of events defining data processing frameworks are described within the context of a big data pipeline. The necessity of big data pipelines is accentuated by the comprehensive coordination and implementation of data ingestion, storage, processing and analysis solutions. These aligned backends are vital to ensure any dataset collected has a clear path to analysis, thus conferring an immense advantage for the proper, and efficient, interrogation of extensive multi-omics datasets. Conceptual depth is added by detailing commonly used data pipeline solutions, historically rooted in the bioinformatics realm, giving an idea of the adaptations necessary for biomedical applications. Additionally, an argument is made detailing the potential benefits from employing big data pipelines for health research as promoting improved data accuracy and the capacity for efficient scaling are overarching goals. The practical utility and importance of big data pipelines for conducting holistic analyses in medical applications are clearly demonstrated by detailing one “pseudo-real” case study, showcasing the algorithmic advantages, as compared to common biomedical paradigms, gleaned from the implementation of big data pipelines.

### 5.4.1. Overview of Big Data Technologies

Over the past 20 years, the size of genetic information has grown from megabytes to gigabytes, and with the advent of high-throughput technology, a terabyte of data can be easily produced in a single measurement. It is expected that a single genome/exome BWA alignment will yield gigabytes of data. This brings on two challenges. The first is the need to store these huge datasets on an infrastructure that can be managed. The

second challenge relates to a sustainable scientific elaboration of these data through skilled and trained personnel. The four characteristics of big data are volume, velocity, variety and value. Volume refers to the extremely large amounts of data with sizes ranging from a few terabytes to thousands of petabytes; so large that it is not possible to manage or process them by using the traditional tools. Velocity is the fast rate of growth of big data. Values and challenges related to big data were reviewed, and six data mining tasks for big data were highlighted: text mining, sentiment analysis, recommendation, classification, clustering, and outlier detection. Variety relates to the different forms of big data, including time series, images, text, videos, voice, and social data that are used in many big data applications. The goal of prospective health research is to understand the underpinning mechanisms leading to disease and health improvement. The technologies of Big Data architecture are Data capture, Curation, Storage, Search, Sharing, Transfer, Analysis, and Visualization. Database systems are designed to simplify the process of storing, retrieving and managing structured information. Different types of database-management systems have been developed over the years, which differ in the representations they use for the different forms of data and the operations they can efficiently support. Cloud computing allows for on-request architecture, which enables users to pay based on their use. The goal of cloud computing is moving data to large data centers where it can be easily collected, and the useful resources derived accordingly. Big Data analytics are the newest, yet the most promising in terms of handling health data, generating new insights and creating a new way for health discovery. Thus the examples and the context focus on the data capture, storage and analysis of health big data. For scientific elaboration multi-classification tools are used with architectures of scoring optimizers of the best performing run, which allows the non-expert user to execute these processes in a user-friendly manner.

### 5.4.2. Components of a Big Data Pipeline

The earlier tremendous progress in biotechnologies has made large-scale and comprehensive molecular data, so-called big molecular data. Big molecular data can describe and measure the experimentally observable molecular mechanisms or characteristics of a biological system in multiple omics levels, such as genomics, transcriptomics, proteomics, and metabolomics. Big molecular data has stretched the interdisciplinary frontier and revolutionized biological research. There are a few grand challenges in biological research that could be addressed through analysis of big molecular data, which in turn triggers and fuels numerous computational methods and tools developments. The most outstanding grand challenge expects to The data pipeline comprises so many components that have been classified into four categories. A grand challenge in systems biology is to predict the system-level response of a biological system to perturbation.

The first stage of the big data pipeline regarding big molecular data is data collection, including but not limited to modeling of sampling processes, measurement and detection principles, and sample pre-processes. Big molecular data is generated via various high-throughput measurement platforms and devices, by which quantitative data of multiple molecules in a biological system is acquired in a high-throughput, large-scale, global, and comprehensive manner. The biological system undergoing the measurement process should be properly sampled at given conditions and contexts. The sampling process and conditions of the biological system should be documented for further analysis. The sampling device or method and its characteristic is modeled to facilitate data collection procedure in a computational way. Data collected through the measurement platform and device might have inherent noise, deviation, and missing values due to various protocols and detection principles. Pre-process (clean and process) the raw measurement data is commonly conducted prior to the downstream data analysis. The quality of the big molecular data should be assessed after sampling and measurement. It is non-trivial to decide the quality assessment and thresholding criterion of big molecular data. The raw measurement data might be too big to be stored in memory or disk. It is necessary to develop a file format and storage for large scale big molecular data.

### 5.5. Methodologies for Integrating Multi-Omics Data

Most recent technological advancements have enabled high-throughput omics data generation from various biological samples. Researchers have generated vast amounts of data regarding nucleic acids, proteins, methylome, and metabolites, among others. Data heterogeneity and the large scale of omics data bring about challenges both in achieving data scalability. Numerous methodologies have been developed for the analysis of each type of omics data individually, while only a few methods have been produced for a simultaneous analysis of two or more types of omics data. However, such analysis is required to leverage integration of multi-omics data generation of new insights and knowledge. This section aims to delineate, for researchers and technologists, the methodologies that may be employed for the processing of multi-omics data through suggesting big data pipelines. A narrative of these methodologies will be followed by a description of the collaboration with other research and technology organizations and a call to action for more systematic study. Since the origination of PCR, many methodologies have been developed for the analysis of each type of omics data individually. With the use of these methodologies, researchers have vastly increased knowledge about diseases and phenotypic traits through the analysis of genomics, transcriptomics, epigenomics, and other types of omics data. However, part of the distinctive function and connection mechanism of biomolecules in cells might only be understood through a simultaneous analysis of several types of omics data. A few methodologies have been developed for such a simultaneous analysis of two or more

types of omics data, but challenges remain in the availability and usability of such methodologies.

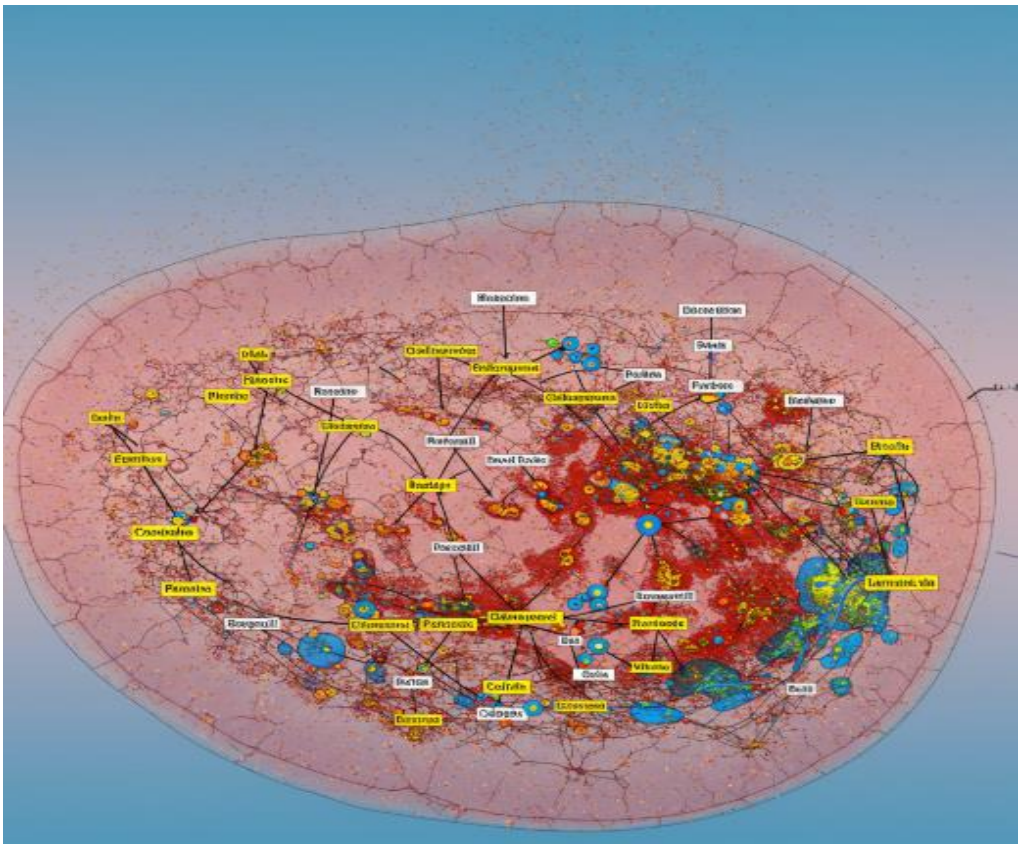

**Fig :** Spatial transformation of multi-omics data

### 5.5.1. Data Preprocessing Techniques

The first-step advanced statistical analysis is data preprocessing, to ensure that the data inputted is of optimal quality. Preprocessing is an intermediate or initial step in which raw data are transformed into standardized formats to ensure quality and consistency before initiating an analysis. Even though omics data can provide insights into complex biological phenomena, each omics layer has biases in its measurements. Multi-omics data, specifically, differ in biases and discrepancies across data layers. Due to these biases and discrepancies, it is necessary to perform common preprocessing techniques such as normalization, imputation, and transformation, using statistical algorithms to correct for biases and ensure that the datasets are comparable across the different omics layers. In addition, missing values and outliers often occur in omics data and directly impact the results of integration analysis. Therefore, thorough data preprocessing is

essential to remove or reduce these issues. The treatment of these missing values and outliers is a crucial part in preparing the data for further analysis (integrating, clustering, classifying). This protocol suggests best practices for handling missing data and outliers in omics data to ensure efficient and accurate analysis, as well as reproducibility and comparability of results across different datasets. The proposed preprocessing steps have been performed using omics data analysis tools, wherein R is a powerful statistical computing programming language for the analysis of omics data that is widely used these days due to the presence of numerous packages for handling and analyzing omics data. In addition to that, tools may also be cited that have GUI form to facilitate the omics preprocessing procedure. This non-exhaustive list aims to provide a starting point for leveraging big data pipelines in omics data research.

### 5.5.2. Statistical Methods for Integration

Multi-omics data sets include, without limitation, genomics, transcriptomics, epigenomics, proteomics, and metabolomics, and other regulated entities. Substantial growth is foreseen in the holistic study of biological systems in order to promote health. However, accounting for the highly multi-dimensional and large data is difficult. This article reviews central needs and issues with such research and identifies advances that urgently occur. To simplify exploration and investigation of complex and big data in general and multi-omics large data in particular, it is recommended that the study and use of big data pipelines should be leveraged efficiently.

Statistical methods are the cornerstone of multi-omics data integration. Strategies for the integration of several forms of example data are shown, and the factors are discussed in picking the appropriate methodology while accounting for the study context. The integration of multi-omics data is a difficult problem that has contributed to a growing body of literature. Various forms of techniques can be used to sum up and derive information from the dataset while also reflecting relationships and signals. Principal component analysis (PCA) is a basic strategy to achieve this, and network-based strategies continue to increase in popularity. Cluster-based and other unsupervised procedure types will also be used to identify common functions and data formats within the input or to group the output results. Together, those approaches are either conceptualised or computationally modelled, and a comprehensive range of methodologies have been created and tested.

Applied workflows differ in complexity in multi-omics data integration despite the many techniques that exist and considering the data blankness, privacy policy, and goals of the research as well as the different aspects of the multi-omics data being researched.

Different datasets contribute various formats of data: single omics wide, single omics dense, numerous omics wide, and numerous omics dense. Each of these forms of data therefore keeps correspondence to specific variations of the model preparation and completion. Sample datasets are not appropriate for all network-based strategies, and the model distribution has to be taken into consideration in using certain modelling strategies. Nevertheless, both of these integration strategies are pliable to many various types of studies and data. Further, the results of these methodologies must be rigorously tested and validated.

### 5.6. Applications of Integrated Multi-Omics Data

Integrated analysis of multi-omics has the transformative potential for health research with the influx of high-dimensional omics datasets, recent advancements in computational tools and frameworks, and the emergence of big data pipelines tailored to processing such data. A focus on some of the successes and implications of current approaches prevails from this perspective: precise prediction medicine; disease prediction models; and case studies drawn from more recent literature. However, a broader overview of application in the literature is also provided that spans population-scale assessments of linked multi-omics features; wearable multiple sensor data integration; and holistic patient monitoring, analysis, and behavior profiling for public healthcare management. These applications are shown to have far-reaching implications for a recontextualized 'omics' era and advocate for a pivot towards data-driven public health policy.

Precision medicine is the most pervasive modern application for integrated omics. Precision medicine aims to tailor treatments to individual patients based on their biology as manifested through omics. By discovering the biological markers that relate to the clinical response of a patient, treatment can be targeted to those patient populations most likely to have a beneficial clinical effect, efficacy, whilst minimizing adverse effects, safety. In this sense, clinical decisions ultimately become data-driven, based on the biological profile of each patient's disease. This simultaneous relationship of multiple variables is most effectively captured in a multi-omics framework. Multi-omics uses various types of genomic data that fit within biological applications that, when taken together, give information about the biological state of a patient: here, mutations, gene expression, and proteomes of tumors are integrated, and inturn, correlated with a patient's clinical outcomes, thereby enhancing the personalization of therapy. Early success in this application has shown that multi-omics data driven approaches outperform other types of classifiers, which use only individual omics modalities. Application-ready big data pipelines that integrate machine learning tools with multi-omics data preprocessing steps are also open source and increasingly accessible, and

successful applications of this type of method in the literature. For ease of illustration and wider impact, intensive focus is directed on three case-studies and the growing repertoire of disease indicating models, with an acknowledgment to the broader success and nuance of other applications. A more current perspective on the successful use of integrated multi-omics to characterize the emergent properties of complex diseases is water.

### 5.6.1. Precision Medicine

Precision medicine, also known as personalized medicine, is an approach that aims to provide the most suitable healthcare solution to a patient, considering his/her individual genetic, environmental, and lifestyle factors. It focuses on tailoring a therapeutic and preventive strategy that best suits the unique features characterizing a person. Given the large amount of information characterizing the biological makeup of an individual, the successful clinical application of a tailored therapeutic solution requires a deep and comprehensive characterization of the biological profile. The profile consists of multiple levels of information: genetic variants, together with the epigenetic modifications regulating gene expression, control the possible structural traits of an organism and its potential responses; the gene expression and pathway regulation; the metabolism profile and its regulation of environmental and diet interaction; disease-related proteins and markers; and, at the level of the environment, microbiota and other multi-omics related factors. The success of precision medicine mostly derives from linking this complex profile with the real clinical outcome of interest. The integration and interpretation of such a large and diverse amount of information proved to be a challenging task, tackled in recent years by leveraging some advances in the big data analytics field, with special focus on machine learning methodologies. Many successful stories have been witnessed in recent years in improving the patients' responses to a particular treatment.

But the full introduction of precision medicine in health care practices still faces several open challenges. The main problems currently limiting the full exploitation of omics data for personalized health are related to the privacy of data, the standardization of clinical omics procedures, and the reimbursement of omics technologies. Since a large variety of -omics technologies is available, each focusing on different specific biological traits, the big data analytics field has applied a variety of approaches to effectively model the extracted information. Big data analytics are widely based on network theory, which is traditionally used to represent and analyze complex systems. The general idea behind most of these models is the representation of biologic structures and processes characterizing the omics data as network structures (biological networks) modeling the interactions between the entities related to the multi-omics data of interest. A first set of methods on the market reuses the generated network structures to interpret the extracted

information, serving as biological knowledge bases. Alternatively, other approaches characterize the resulting networks underlying the -omics profiles. Several case studies highlighted in the section may show a number of successful applications of precision oncology that exploit the integration with multi-omics data. A separate paragraph finally discusses the future perspectives in the field, outlining how the technological advances and novel big data models may further leverage the application of the integrated multi-omics strategies in precision health.

### 5.6.2. Disease Prediction Models

Early detection and preventative strategies can significantly reduce the burden of various diseases on individuals and health systems. To develop reliable strategies, risk factors and symptoms associated with particular diseases need to be identified, so that individuals at risk can take appropriate action. A variety of studies have investigated factors associated with different diseases using clinical data, genome-wide association studies, and animal models, among other approaches. However, predicting disease likelihood remains challenging because many influential factors are unknown, complex, or stochastic, and the predictive accuracy of models developed so far is limited. Multi-omics data is now widely available, reflecting a variety of biological layers that are correlated with each other. The strength and nature of these correlations are often altered in pathways or networks upon perturbation of biological systems. By leveraging these correlations, disease prediction models can be informed by multiple layers of data or predictions obtained from other models. These correlations can be at the peak of gene expression from an animal model and be captured by knowledge-guided feature selection before fitting the model. Since these correlations are not captured directly by the model, they provide efficient communication and improve the understanding of biological systems. Both machine learning and statistical techniques can be employed to construct disease prediction models that can consider the complicated interaction between different omics and clinical/symptomatic data. These techniques can be formalized to generate estimates, expectations, or inferences in a more rigorous process, better managing the challenge of heterogeneous and large datasets well for high-dimensional and noisy data, which is typical for bulk omics datasets.

## 5.5. Conclusion

Modern healthcare practices are evolving. The era of ‘one-size-fits-all’ is being replaced with precision medicine where each patient will receive optimized healthcare based on their genetic, environmental, and lifestyle factors. A plethora of health determinants have been discovered that contribute to individual health, such as air quality, social-economic

status, diet, bacteria, viruses, land-use, genomics etc. With current technological advancements, it is now possible to monitor such health determinants in real-time.

The integration of multi-omics data in modern health research is discussed. Highlights from the new health research methodology to generate holistic health insights are shared. Furthermore, the immediate health determinants and the expected real-time monitoring are discussed. Healthcare as-a-service is being transformed to personalized medicine. However, current health research practitioners are not capable of deciphering this wave due to various factors. A new evolution based on the integration of modern technologies in omics data will be discussed, aiming to ease the day-to-day practices of health research and health professionals. Similarly, continuous advancements and usage of various technology standards from sensor manufacturing companies will also be shared. In order to translate the holistic insights to actionable health results, big data pipelines will be discussed. Throughout this journey, challenges faced since they are evolving with technology and the ways to reduce or avoid them are also shared. Finally, potential future directions of this new evolution are shared, ultimately the goal of this awakening discourse is to pass a wave to the health research community in order to adapt and evolve with the technological advancements. The eon of precision medicine is upon us – the time to get ready is now!

### 5.7.1. Future Trends

Understanding of the human body has come a long way since Hippocrates first described his influence on health or disease as the notion of the ‘four humours’ working together to achieve ‘humoral balance’ over 2,000 years ago. Although other external factors were considered, at the core of everything were four balanced elements, or humours (blood, phlegm, yellow bile, and black bile). Imagine a world where the philosophy is replaced by a physics approach on analysis or observation of a noisy recording of the noisy body, exploiting the important elements such as measurement, signal processing and networks, as if the body is just another form of signal recording devices. Separately, they may not deliver full insights for the whole disease mechanism; but appreciate the fasciculating interactions among those elements that supply holistic and enriched knowledge. ‘Multi-omics’ as a concept consists of multiple data types including clinical, environmental, and omics features. Acceleration of technologies along with reduction in price in the last 10 years has also led to a dramatic increase in multi-omics data availability. At the top level, population-level ‘big data’ can be analyzed at the same time as utilizing the big data for individual participants. Naturally a critical part is the biology: the way things are working together, the networks and the biomolecular pattern, how the metabolism affect the way a person is presenting (or the other way around), and so on; there is a need to be the bridge between biologists and data scientists, uni-discipline research seems

likely fell short to impact real real-world science. In near future, something called passive omics can be found hanging above hippocratic heads for the multi-omics reading. When the data is analyzed in real-time by considering patient felt in a machine learning model, they never thought that any inherent potential in the aura of a care provider could push predictive modeling in holistic health, ultimately converting the obsolete intervention. Plus, new omics technologies have paved the ways in deciphering the non-invasive humours to better understand the health status of the body; and making the analysis is just a matter of sensing needs to surface the whole sickness humours. However, interpretation by onlookers belonging to current/parallel visions may differ; some may understand the early potential before its maturity like that big bang in the 40s, while others keep uttering words of doubt and indifference. Nonetheless, this big "multi-omics" bang is happening, and its wave is propagating across every aspect of health. Ideally, this needs to be handled by a leveraged real-time distributed big data pipeline that may guarantee like what is achieved in dark matter research scenarios that multi-omics data becomes a common entity or routine measurement in precision health.

## References


Kaulwar, P. K. (2023). Tax Optimization and Compliance in Global Business Operations: Analyzing the Challenges and Opportunities of International Taxation Policies and Transfer Pricing. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 150-181.

Koppolu, H. K. R. (2022). Advancing Customer Experience Personalization with AI-Driven Data Engineering: Leveraging Deep Learning for Real-Time Customer Interaction. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3736.

Singireddy, J. (2024). Deep Learning Architectures for Automated Fraud Detection in Payroll and Financial Management Services: Towards Safer Small Business Transactions. Journal of Artificial Intelligence and Big Data Disciplines, 1(1).

Sneha Singireddy. (2024). Leveraging Artificial Intelligence and Agentic AI Models for Personalized Risk Assessment and Policy Customization in the Modern Insurance Industry: A Case Study on Customer-Centric Service Innovations. Journal of Computational Analysis and Applications (JoCAAA), 33(08), 2532–2545.