American Journal of Biology and Natural Sciences
https://biojournals.us/index.php/AJBNS
ISSN: 2997-7185

Integrative Multi-Omics and AI-Driven Biomarker Discovery for Early Diagnosis of Complex Diseases

Lamia Amin Ghaly Rakhis
Basra University Faculty of Science Department of Biology

Haneen Saad jabbar fadhil
Department of biology, College of Science, Al Qadisiyah University, Al-Qadisiyah

Faten Shaker Hariod Khadir
University of Karbala College of Science Department of Biology

Mustafa riadh abed mohammed
Samarra University, College of Applied Sciences, Department of Pathological Analysis

Sura Rahim Abdul Zahra Jawad
University of Babylon College of Science, Department of Biology

Received: 2024, 15, Mar
Accepted: 2025, 21, Apr
Published: 2025, 28, May

Annotation: The advent of massive data-gathering technologies offers unprecedented opportunities for exploratory, hypothesis-generating research. Due to the complexity of biological systems, such data represent an incredibly intricate combination of biological, technical, biological-laboratory, data-integration, and analytical noise. Consequently, to glean conclusions that can genuinely advance knowledge, the first step is to apply validated data-agnostic and data-driven clustering- and dimensionality-reducing algorithms to reveal the key biological variables contributing most, and then study their interaction and interdependence.

This article presents a general multi-omics framework that integrates gene expression, methylation, expression protein mass spectrometry, and copy number alteration, along with clinical follow-up,

patient information, key pathways, and gene-gene networks of involvement, and encompasses unsupervised algorithms operating to reveal features most informative of ovarian cancer (OC).

Such an integrated framework, which can incorporate other “omics” data as they become available, offers multiple opportunities, ranging from supervised and non-supervised feature generation of a multi-omics type to integration of different types. It opens up unexplored avenues for the extraction of any type of biological knowledge from any type of data, irrespective of its discipline, bioprocess involved, or its dimensionality be that empirical time-series data, Boolean data, or others.

The focus presented here is strictly on experimental biological data relevant to comprehension of a dynamic biological system and its discoveries of networkome, pathway redundancies, driver(s) under combinations/chains of events, or other outputs informative of such a dynamical biological system. Emphasis is put on data exploration methodology, priori requirements, types, and advantages of different T.sensor-linked experimental data, observables, and the phase space explored, and on how to prepare the data in a compatible way for ensuing analysis and hypotheses generation. Key aspects in terms of generality of application to biological knowledge discovery from any type of experimental data are also discussed.

1. Introduction

Cancer is a diverse and complex disease caused by the accumulation of multiple alterations in genome at various levels such as DNA, RNA, protein modifications, and epigenetic changes. Multiple biological studies are needed to unravel the underlying complex mechanisms of tumor development and progression before unveiling preventive strategies or superior therapeutic approaches. Such omics investigations reinforce the need to search for clues across platforms or modalities, which are commonly referred to as multi-omics data. Over the past decades, many breast cancer omics studies using diverse platforms have generated an avalanche of publicly accessible data from a wide variety of resources. These data provide an extraordinary opportunity to study complex diseases integratively, and are expected to deliver new biological insights unattainable by single-omics studies. Complex diseases such as cancer are rarely caused by a single alteration in a single gene. Elucidating the underlying complex molecular basis of these diseases requires the combined information across multiple levels of molecular characterizations using genetic, epigenetic, transcriptomic, post-transcriptional, proteomic, metabolomic, and multi-level integration approaches.

Within the past two decades, great advances have been achieved in high-throughput technologies to nimble large-scale molecular level characterizations of tissues and cells. These profiled omics data provide molecular information on various levels of biological hierarchies, and are expected to reveal the molecular mechanisms from the bird's-eye view. Consequently, increasing efforts have been devoted to adopt integrative multi-omics data analysis approaches or techniques to fulfill this expectation. The omics data under investigation are described, a general bioinformatics framework developed for integrative multi-omics data analysis, and the expected impact of this framework on breast cancer research are discussed on multiple respects. The use of artificial intelligence (AI) based systems biology approaches to analyze integrated multi-omics data is summarized. The general bioinformatics framework consists of three major parts, namely, pre-processing or data cleaning, analysis, and post-analysis. First, multi-omics data cleaning essentially includes genomic data transformation, filtering of noise, and imputation of missing values. After data cleaning, biological data integration is performed to identify different types of biologically relevant associations among molecular entities across platforms or modalities. On the basis of identified relevant associations, various models are designed to analyze the integrated multi-omics data, which is followed by post-analysis of the results of model optimization [1].

2. Understanding Complex Diseases

Complex diseases such as cancer are difficult to diagnose, and machine learning approaches are increasingly used to analyze multi-omics data for early diagnosis of diseases. In the past decade, knowledge of cancer-related dysregulated processes and aberrations has become increasingly comprehensive, and diverse data have been acquired by high-throughput systems biology platforms. Cancer biology is highly complex and dynamic, with multiple pathogenic processes and various forms of molecular aberrations including DNA mutation, copy number variation, and dysregulation of transcription and protein abundance. Dysregulation of normal genes occurs across all catalogue levels via different aberrant mechanisms, and most of the complex diseases do not arise as a consequence of failure of a single molecular entity. Rather, they can be viewed as system-level disturbances of biological constellations formed by interdependent molecules, pathways and processes [1]. Undoubtedly, the incorporation of knowledge from different "omes" is necessary to gain broad insights into the molecular paths and aberrant processes leading to disease, and thus to the identification of targets for intervention. However, using multi-omics data in a "horizontal version" is challenging.

Multi-omics studies comprise multiple data types, different natures of data, and different platforms of systems biology. Each type of omics data is sufficiently heterogeneous, and the scale and dimensionality of high-throughput biological data have exponentially increased. For cancer, a data repository known as The Cancer Genome Atlas has been developed since 2006 [2]. Data-driven inference of cancer-related aberrant mechanisms and processes requires the coordination and incorporation of a large number of high-throughput cancer data across several omics and non-omics levels. Here, a bioinformatics-oriented conceptual framework to support public efforts as well as individual research work towards the integration of large multi-omics data is presented. A data acquisition platform is developed to enable "vertical" integration of multi-omics datasets from different resources (types of disease, organ sites, data level) with detailed mapping of data provenance.

2.1. Definition and Characteristics

Multi-omics, a types of omics that integrate two or multiple levels of omics datasets that consist of high-throughput and large-scale information and analytics for particular biosystems [1]. Typically, a dataset consisted of one type of omics dataset, e.g. genomic data, transcriptomic data, proteomic data or metabolomic data was referred as single-omics dataset. In contrast, multi-omics dataset or data is composed of omics that contain different (typically two) types of omics datasets. For example, one type of multi-omics dataset is composed of transcriptomic expression data and genomic mutation data. Various omics-based analyses of biosystems data have been extensively studied to gain novel biological insights, particularly DNA/protein/gene mutations, or variants impacting aberrant molecular functions and properties driving dysregulation of cellular functions, preclinical models and drug responsiveness, or therapeutic/patient stratification of cancers. Recently, the integrated analysis of multimodal biomedical data by data-mining and bioinformatics tools has gained increasing popularity and importance to provide a system-level view to understand complex biological systems. Cancer is a complex disease involving the dysregulation of genes via multiple mechanisms. It is unlikely that cancers will be fully explained by solely one type of data (or omic level, denoted as single-omics). Conceiving the hypothesis that by combining different omics, the discovery of novel bio-molecular associations with cancer-related phenotypes will be increased, joint analysis methods for integrative modeling are developed. By investigating functional relations among genes associated with the same disease condition, the knowledge for developing more accurate disease-relevant prediction models will be further gained [2].

2.2. Current Challenges in Diagnosis

The development of effective biomarkers for the early diagnosis of complex diseases is one of the biggest challenges faced by researchers in the field of healthcare and medicine. Advances in several domains, including the introduction of next-generation sequencing techniques in genomics, the development of advanced imaging techniques in imaging 'omics' (e.g., medical imaging), and the development of techniques for molecular pathology and proteomics, have increased the availability of different types of 'omics' data for biomedical studies. Novel multi-modal and multi-'omics' data integration strategies are, therefore, expected to play a key role in patient clinical management in the coming years. It has helped increase the accuracy of diagnosis and the efficacy of therapy [3]. Machine learning (ML) and artificial intelligence (AI)-based approaches have also been developed to assist multi-'omics' data integration in biomedical research. Generally, the main aim of this systematic review is to advocate for the adoption of multi-'omics' data integration along with AI-based approaches. In addition to a description of the current state-of-the-art models and methods, challenges and future research directions related to this are also presented in this section. As an advanced approach to integration, analysis, and prediction of multi-omics data integration, the use of deep learning (DL) models and its variants is also discussed [4].

Despite the promise of the technological efforts and advancements, a number of challenges still remain in the development of effective biomarkers for the early diagnosis of complex diseases. To help researchers better understand the situation of the field, this section presents a detailed view about the current challenges in the development of effective biomarkers for the early diagnosis of complex diseases. It has been organized under four main categories (general challenges, omics data challenges, cross-species challenges, and machine learning challenges), which consist of diverse challenges. The discussion of how these challenges could be addressed is presented in the subsequent section (Section 3). This section also presents questions that could assist researchers in identifying promising research topics related to the current challenges of the field.

3. Overview of Multi-Omics Approaches

In recent years, there has been an increasing recognition of the complex interactions among biological systems. Novel technologies like next-generation sequencing, transcriptomics, and high-throughput proteomics enable a simultaneous measurement of various classes of molecular species in an ensemble of samples, allowing the picture of biological processes at a systems level to be interpreted. Nevertheless, each layer of analysis produces abundant, complex, and diverse data, hindering their effective extraction of biological knowledge, which cannot be approached with any traditional analysis by an isolated domain. A growing appreciation of the combined information of diverse data sources laid the foundation for the analysis at system-wide levels [5]. Multi-Omics data integration emerged as a fresh paradigm to explore the new insights into normal physiology, pathology, and treatment of distinct diseases, well beyond mere correlations and phenotyping in the single-omics era. This field of research is becoming an important focus of investigations.

Recent technological advances yield high-throughput platforms for the efficient accumulation of immense amounts of information across multiple omics levels. Multi-omic refers to the various –omics approaches deployed to observe the molecules of interest that showcase the systems biology of organisms – from genes to the bioinformatics underlying their interactions. The heterogeneity, complexity, and largeness related to multi-omic data accompanied by various bioinformatics analyses, corrections, and modeling ensure a rapid mining of large-scale and novel insights from biology, ecology, agriculture, and medicine. Although researchers work with assorted modalities of analyses, such as statistical approaches, machine learning, systems approaches, global bioinformatics databases, and network-based conceptualization of integrative biology, this paper provides an overview of key aspects and recent advancements for an accessible reference to this swiftly evolving field [2]. The sources of multi-level omics data integration, the corresponding bioinformatics workflow, a summary of key concepts and methods for data analyses, and potential biomedical applications are outlined and illustrated with several real-life examples. Finally, challenges, recent advances, and perspectives on future developments and applications of multi-omics data integration are discussed.

3.1. Genomics

More than 50 years ago, the Human Genome Project (HGP) revolutionized biology and medicine by developing new techniques for large-scale sequencing and analyzing genomes. The past few decades have seen an explosion in rapidly decreasing sequencing costs, with the rapid proliferation of massive genomic sequence datasets made in public archives. The combination of sequence data with functional genomics and other related data, otherwise known as multi-omics, represents the next frontiers to understand the genome. The genomic era poses monumental opportunities for the field of systems biology to revolutionize its approaches to patient-based precision medicine, as well as for high-quality Artificial Intelligence (AI) systems [2].

The dynamics of multifactorial diseases systems at the molecular, cellular, organismic, and societal levels are understood to be better represented by the networks of their biomolecular perturbations than by the mechanistically simpler linear mechanisms discussed above. Multi-omics technologies are key to providing a broad view of the biomolecular activity of a disease system. Nevertheless, they only uncover some of the associated biomedical systems while leaving many others unknown. These unidentified biomedical systems may hold key pieces of information in understanding the complex nature of diseases. Unsupervised AI approaches have been developed to mine unseen networks expressed in multi-omics. For instance, fuzzy clustering techniques have been applied to decipher subtypes of human gliomas. Multi-omics data, along with patient-specific clinical traits, have been demonstrated to improve patient stratification and better prediction of clinical outcomes. Integrative networks are formally learned from the multi-omics data, patient-specific clinical traits, and discrete or continuous network properties, which are used to train interpretable and predictive support vector machines classifiers.

3.2. Transcriptomics

The transcriptome is defined as the total set of RNA transcripts or gene products in a cell, which refers to the set of messenger RNA (mRNA) molecules in the cell [6]. Transcriptional control of gene expression is an important physiological process and plays a crucial role in regulating cell fate. In this section, multi-level information on the transcriptome, such as differential expression, alternative splicing, and large non-coding RNA, is presented. Furthermore, multi-omics biomarker identification methods of transcriptomics data, such as ontology-based and Venn diagram-based analyses, and the databases containing transcriptome data used by biologists are also introduced.

The transcriptome contains the full range of mRNA molecules expressed by an organism. The transcriptome's three major components include: coding RNAs (protein-coding mRNAs), non-coding RNAs (ncRNAs), and rRNAs, in which genes and mRNAs are 4-7 kb long and composed of 3 exons on average; the number of alternative spliced isoforms per gene is 12,396 and the number of protein domains is 5,652. Understanding the transcriptome is critical to understanding the biology of an organism and has important implications for modern biomedicine. Currently, the transcriptome provides us with an appropriate time window in which various cellular states can be selectively captured. With the advances in sequencing and bioinformatics skills, understanding and capturing the temporal, spatial, and cell-type-dependent landscape of the genome has become feasible. A snapshot of the levels of each transcript in the transcriptome at a specific time point in a typical organism is called transcriptome profiling. Currently, several transcriptome profiling platforms, including conventional bulk mRNA sequencing, planar and spherical microarray technology, and no-lyse sequencing are widely used.

However, traditional transcriptome sequencing faces difficulties in routine applications in clinical and biological studies, including, a lack of a standardized library construction protocol, an ill-defined bioinformatics platform, and a price disadvantage. To resolve these limitations, a new sequencing platform, LUNA Sequencing, was developed, which includes a reverse transcription-free, constant-timing, constant-volume, and constant-temperature real-time amplification mechanism, and multiple channel Macshyny (M) chips to enhance multiplexing of up to 10,000 channels, with obtained raw sequences of 450 nt. This high accuracy of LUNA seq-based mRNA sequencing, which is superior to bulk mRNA-Seq, planar mRNA transflective phase-array microarray, or nth-geometry multiplex detection microchip technology, was applied to obtain transcriptome profiles in a solid sample, highlighting its robustness in a-few-cell analysis.

3.3. Proteomics

Proteomics provides a powerful platform through the identification of proteins as complex disease-associated features to be used for prediction, diagnosis and now drug development, especially in conjunction with genomic and metabolomic features. However, the slow production and expensive analysis of large plasma proteomes drove continuous efforts towards simpler feature sets. The potential of machine learning to harness the biology intrinsic to a vast number of enzymatic reactions provided a way forward. Protein biomarkers for complex diseases were comparatively described, with a focus on proteomics analyses of samples from the UK Biobank. Failure to make predictions into clinics is also discussed, providing recommendations for impactful future research. Abundant and accessible population-wide genomic, proteomic and metabolomic data makes it feasible to investigate prediction of complex diseases based on the cross-comparison of omic features with machine learning. In addition, the widespread availability of large-scale data creates a new challenge: As the dimensionality of available features is growing by an order of magnitude, their large-scale cross-comparisons must be interpreted with care, taking into consideration attributes intrinsic to specific omic pairs and the machine learning tools available to analyse them. One set of features that has so far seen less application are proteins, which are the products of genes and often the output of cellular modulation. Thus, the investigation of whether accounting for the biology of proteins could improve the prediction performance of any of the above disease attributes is relevant [7].

Models informed of genomic and metabolomic features were described to predict complex diseases, and comparative analyses exposed stroke, type 2 diabetes and atrial fibrillation, consistent with complex diseases involving multiple causative factors. Data from proteomics analyses of UK Biobank participants were presented, with an emphasis on predictions based on robust analyses across masses, retention time and intensities. Such observables would allow for simpler, more accurate and therefore more reproducible interpretation, applicable to future plasma peptidomics studies as well. Direct comparison of the potential of proteins, metabolites and genetic variants to predict complex diseases based on their relative cross-comparison with omic approaches. The potential of the proteins alone to provide accurate predictions of complex diseases was thus verified, especially synaptic potency and blood-spinal cord barrier-related proteins.

3.4. Metabolomics

Metabolomics is the study of quantifying metabolites and mapping their complex interactions within this domain, which is comprised of the total set of small molecules present in cells, tissues, organs and biological fluids. It is the final downstream component of the biochemical stages, involving genes, RNA, proteins and environmental factors, ultimately yielding phenotypic changes in an organism. Since metabolism crucially involves important physiological processes that diseases often alter, metabolomics analyses can be used to detect disease-driven changes from the levels of thousands of metabolites, enhancing current diagnostic methods and discovering specific, perturbed metabolic networks. The advantage of using metabolomics is derived from its provision of a functional readout of the physiological state of an organism. Importantly, metabolomics may hold the key to tackling the challenges associated with complex diseases, which are caused by an intricate interplay between an individual’s genes, environment and lifestyle. Most diseases lie under this umbrella term, which include cancer, cardiovascular disease, diabetes, arthritis, obesity and dementia. Rather, expression of certain correlational genes may increase risk of contraction, but does not guarantee incidence; instead, toxins from the environment, drugs consumed over one’s lifetime, poor diet and lack of exercise would likely lead to disease onset. Therefore, researchers of complex diseases must identify methods to overcome the challenges of deciphering the quantitative influence of risk-associated genes in comparison to non-genetic factors. Metabolomics offers a solution to this by allowing the individual influences of genetics, environment and lifestyle to converge onto the metabolome as a terminal downstream domain of products. This holistic approach allows metabolomics researchers to discover biomarker signatures that capture the multiple major factors driving the complex disease. These panels can help to diagnose at-risk complex disease patients and predict onset years before symptoms arise using prodromal metabolomes. Research for metabolic marker discovery spans a fast-growing array of prevalent disease areas, such as breast cancer, osteoarthritis and Alzheimer’s. Although rich quantitative datasets may contain valuable information, the extents of their utilities are limited by the appropriateness of the selected statistical and computational methods of analysis. Since these datasets contain hundreds of features, the value of an appropriate method would be derived from its ability to account for the effects of each metabolite in isolation, and in a multivariate manner with consideration of interaction-based effects. Thus, while recent advancements in analytical chemistry techniques have made it possible to quantify hundreds of metabolites within a reasonable time frame, these techniques must be coupled with fitting statistical and computational algorithms to translate the data into a practical application in the clinic. Unfortunately, the majority of metabolomics studies historically have not employed optimal methods for biomarker discovery, perhaps due to a lack of statistical and computational expertise among metabolomics researchers. Today, the existence of over 100,000 metabolites in the human body is reported. As analytical methods improve, the quantifiable metabolome and its associated datasets will continue to grow, raising the relevance of powerful, heuristic computational methods to the forefront. [8][9][10]

3.5. Integrative Analysis of Omics Data

The application of integrative analysis in multi-omics data is growing rapidly. These multi-omics data sets help identify and provide insights on the disease and stratified samples, pathways, and predictors of diseases. In recent years, more integrative analysis applications have been reported, showing its growing popularity in the diagnosis, prognosis, and treatment of diseases [5]. Personalized medicines leverage multi-omics data to provide insights on the personalized disease mechanism and identify personalized driver genes. For example, a study assessed the impact of tumor-mutated alleles on the functional activity of proteins. This approach identified and prioritized 5 driver genes in this patient, which were validated to play a crucial role in tumor cell growth. Other studies have developed methods to leverage multi-omics data for drug response prediction in diseases. Clinical assessment predictions combine the physician’s assessment and the omics data to predict treatment outcomes for depressive disorders. Under the evaluation of real clinical situations, their workflow predicted the therapeutic response by integrating mutation and metabolomics with clinical observations. This integrated approach harmonizes clinical and omics data and can provide novel therapeutic interventions for diseases with complex phenotypes.

Cancer is a complex and pervasive biological phenomenon involving the dysregulation of genes via multiple mechanisms. Oncogenic alterations may involve single genes or multiple alterations affecting distinct classes of genes [1]. This is unlikely to be fully explained by a single data type. By combining different “omes”, researchers can discover novel bio-molecular associations with disease-related patient phenotypes. For example, the genomic and transcriptomic composition could provide complementary and cross-validation insights into the tumor biology and could potentially present a more comprehensive profile of the multi-processes of tumorigenesis. In a tight collaboration with a data-generating institution, an integrative framework Data & Analytic Integrator (DAI) is developed to explore the relationship between different omics via different mathematical formulations and algorithms. DAI is underpinned by a combined data & analytic integration approach. Input data sets of different “omes” are first lighter-fined into a collection of data sets, allowing detailed personalized exploration of sub-networks using other network modeling tools or software.

4. Artificial Intelligence in Biomedical Research

Artificial intelligence (AI) has profoundly impacted numerous facets of society, including how data is created, processed, and used to make decisions in biomedicine. AI-enhanced solutions, particularly those based on deep learning, have opened up new horizons in biomedical research and discovery by providing sophisticated analysis and modeling capabilities. Several strategies for employing machine learning (ML), bioinformatics, and systems biology to integrate large quantities of data across multiple levels of biological processes have recently become accessible and applicable to biomedicine fields. Data-driven approaches capable of learning from high-dimensional data at the cellular level and predicting their disease associations at the tissue level have been widely adopted in drug development. AI-based approaches will likely be increasingly crucial to biomedicine as this paradigm shift in science and drug discovery continues with the faster and larger growth of biomedical data in the future [2].

AI has become a basic and key technology for drug discovery, meaning every striker would inevitably use such technologies and platforms. The tutorial covers AI and its application scopes in the industry and academia, focusing specifically on strategies and tools for analyzing multiomics data with AI technologies and effective cooperation with non-AI scientists. Multiomics approaches and their applications in biomarker discovery and drug testing. AI-based deep-learning methods enable biomarker identification and drug response prediction from multiomics data. AI-based multiomics data analysis tools provide additional value in case studies.

The scholars anticipate that knowledge learned from this tutorial with hands-on experience will broaden the audience's perspectives, helping them explore innovative ideas for AI, multiomics, multimodality, and other techniques in future discovery.

4.1. Machine Learning Techniques

Over the years, numerous machine learning approaches have been developed for the analysis of multi-omics data. This section discusses the major categories of machine learning whereby representative methods for each category are provided. Under each category, a detailed analysis of the method's background and its cancer-related applications is conducted. Moreover, a summary table of detailed works is included to provide a comparative view of existing methods' features, advantages, and limitations, which can be helpful for researchers in selecting appropriate methods for their research needs.

With the extraordinary success of deep learning in computer vision, natural language processing, and many other machine learning-oriented tasks, deep learning-based methods have also been developed for the analysis of multi-omics data, including multi-modality gene expression, single-cell RNA sequencing and spatial transcriptome. Most efforts focus on developing novel neural architectures for integrative analysis tasks, which usually require a large amount of training data. Exploring more machine learning models for multi-omics data, especially conventional approaches, is an open challenge but essential for real-world applications where the current data size is often small. Since many conventional approaches can be regarded as a simplifying assumption on the latent structure of the target domain, the theoretical foundations of conventional approaches should also be established for the integration of multi-omics data analysis. The latest advances in graph neural networks may hold promise [2].

Together with the International Life Sciences Institute (ILSI) have developed a research framework for risk assessment of the potential adverse effects of plants produced by new biotechnology methods. To assist in implementing the research framework, a decision-support tool called Nidus was developed. Available human-health, environmental, and agronomic decision-modifiers were digitized into a user-friendly format with guidance on how to locate, assess, and utilize the data. The development, implementation, and illustrated use of Nidus offers a replicable approach for agribusinesses, NGOs, and government agencies to assist risk assessors worldwide in fulfilling their responsibilities to protect human health and the environment in the age of agricultural innovation.

4.2. Deep Learning Applications

Deep learning (DL) health informatics applications review is presented here, focusing on the supervision of extracting multi-layer and multi-resolution parameters in the field of neurodegenerative diseases (ND). The high-level learning techniques that have been proposed in ND have been classified into math/statistics, signal processing/image analysis, knowledge representation/graph theory/network science, and DL. They involve many possible data attributes. Examples of their applications to ND retrieval and classification of medical imaging, genetics-sequencing, radiomics, E-EG/MEG signals, and texts are presented. While matchmaking classic attributes of data analysis methods and data attributes clarifies what can be analyzed with each method, building DL systems is not easy for biologists due to their complexity. Therefore, such habitats are also identified with state-of-the-art off-the-shelf turnkey software tools [11]. In particular, deep learning (DL) and artificial intelligence (AI) approaches that have been developed for the fully intelligent analysis of multi-omics and multi-features of complex cancers are reviewed. The public resources available for AI-based cancer multi-omics systems biology studies are compiled, as well as benchmarking guidelines. Furthermore, the most significant challenges and unmet needs in this emerging field for future development and use are highlighted [2]. The complexity of cancer in the context of systems biology is discussed, followed by descriptions of the state of data generation and systematic analysis applicable to cancer at multi-omics levels. A framework is proposed that incorporates emerging AI- and graph-based technologies for systems-level interpretation and multi-omics integration of the cancer-wide human molecular constituent, as well as fundamental considerations for robust application. Finally, the impact that applications in this field can have on precision cancer prevention and treatment is illustrated with a few examples from the ongoing research efforts. Given the rapid development of big data application technology and the need for effective prevention/therapeutics for all kinds of cancers, a promising future is envisioned. [12][13][14]

4.3. Natural Language Processing in Health Data

Natural Language Processing (NLP) methods can convert the unstructured text notes in electronic health records (EHR) to structured data fields. These methods can help advance clinical decision-making for chronic diseases by facilitating cohort identification for ongoing monitoring outside of clinical trials. A generalizable approach was developed, successfully extracting variables underlying treatment and clinical outcome in nephrotic syndrome from EHR free text notes [15]. This proof-of-concept study serves as a use case for broader applications of NLP in studying chronic diseases. NLP offers potential for improved risk stratification, cohort identification, and outcome assessment, yet it has not been applied extensively to offer pseudonymous cohort information which can be made publicly accessible.

Text notes resulting from clinical encounters, discharge summaries, pathology reports, imaging readings, and other narratives often contain information consistent with intent of clinical trials. Electronic health records (EHR) for chronic diseases – settings of treatment failure or disease progression – can mature for years, generating large text corpuses amenable to mining. Natural language processing (NLP) provides methods to convert unstructured text to structured fields and standardized entity types. NLP applications on EHR are rare with respect to issues of intervention, clinical context, and disease phenotyping. Approaches to health text mining, event timeline construction, extraction of cancer quality-of-life surveys, and systematic use of process mining have emerged, yet research areas remain wide open. NLP of EHR notes has potential for improved trial feasibility, population robustness, and targeting of higher risk patients not currently in interventional settings [16].

Complex chronic diseases – computational frameworks for text mining could be built based on freely available resources such as administrative datasets and patient-generated health data. Clinicians and researchers could refine relevant variables leading to datasets clearly demonstrating effectiveness of a priori targeted therapy strategies. This unprecedented access to clinical evaluation and learning data could inform care paths tailored to differences in intervention type, risk factors, and patient profiles. In trials on chronic diseases to date, routine text notes in free form narrative sometimes asymmetrically disclosed to patient care could be considered a valuable source for standardization and compliance reasoning analysis.

5. Biomarker Discovery Process

The Biomarker Discovery Process (BDP) is essential in biomedical research. It aims to find candidate biomarker molecules that can subsequently be validated and used in clinical practice for patient stratification in a precision medicine context. Robust bioinformatics pipelines have been established that take some high-dimensional omics datasets and yield interpretable result lists of biomarkers and accompanying bioinformatic analyses [17]. These pipelines are usually based on meta-analytic statistics. They reproduce essential steps in most hands-on pipelines in painstaking detail and shine light on potential pitfalls. Nonetheless, they cannot be used in an “all-in-one” fashion or by the majority of life scientists yet.

Academic researchers in biomarker discovery studies for patient stratification using omics data are slower than vendors to seize the power of AI-enhanced solutions. This is partly due to the fact that new machine learning approaches are steeped headlong into established but deficient bioinformatics tools. In organizational terms, the AI-olution space is fragmented, the integration costs between providers are significant, and the availability of user-friendly, open-source standalone software is limited. However, niche open-source efforts like the “SwathX” offer untapped opportunities for scientifically painting a thousand colors of patient profiles, and this rich palette is still waiting to be distributed to the art sectors of the biomarker world [18]. With the increasing scale, complexity, and integration of omics data, conceptual clarifications, user-friendly software, and appropriate example use cases are still in high demand. Closely associated priorities include fostering towards explainable AI-augmented solutions designed for high-dimensional heterogeneous omics input, and propensity to adopt more open-source software that is responsive to the shifting landscape of multi-omics.

5.1. Identification of Potential Biomarkers

The identification of potential biomarkers of complex diseases has proven a daunting task due to the data gap as well as the complexity of both the biomarker and the disease [7]. Early on, candidate biomarkers were typically followed up based on the scientific literature, clinical experiences or selected from multiomics analyses based on simple statistics such as univariate p values. The cleared path for prioritising candidates was left unexplored. Here, this gap is filled with an interactive online tool. Analyses have identified 90 million common genetic variants, 1453 proteins, and 325 metabolites across 30 complex diseases, and systematic comparisons have been made of the individual potential of both the candidate sets and the predictor variables. Since complex diseases are characterised by both complexities in the disease and in the biomarker, machine learning has been used as the preferred methods to build a pipeline for the evaluations.

It was found that the cohort size balanced the need for hue, diversity and coverage of the candidate biomarker datasets. Consequently, UK Biobank was chosen with some 1.7 million genetic variants, 1453 proteins, and 325 metabolites involved in the incidence and prevalence of broad range of complex diseases. UK Biobank has recently made extensive phenotypic and multiomics data available to researchers. This publication includes a major part of these newly obtained data from both the cohort and the experiments for the benefit of the research community, including the on-line tool for biomarker prioritisation.

With the rapid development of data generation technologies, the amount of omics data concerning human health and disease is growing exponentially. Meanwhile, healthcare is concurrently shifting from intervention towards prevention with further patients’ stratification per case and subsequently to tailor made intervention. The escalating needs of biomarker discovery to assess risk, select patients and monitor predictive intervention have confronted trials with omics data generated as an obverse. Genomic studies have buoyed up expectations by finding high throughput, low cost and commercially viable variants. However, hundreds of thousands of candidate variants can be primitively filtered down by selection based on the literature or biochemical pathways, yet the odds of being diagnostic are extremely slim.

5.2. Validation of Biomarkers

Recent advances in omics technologies offer exciting new opportunities for biomarker discovery. Omics data have proven successful for patient stratification in clinical applications related to oncology, where the complexity of tumor mechanisms has inspired multi-omics biomarker design [17]. Unraveling the underlying mechanisms of complex diseases is fundamental for reaping the potential of multi-omics data for biomarker discovery. Patient stratification necessitates tight collaboration of medical research, applied life science and bioinformatics, as well as orchestration of diverse data acquisition and analysis methods. Key elements for involvement of academia, industry and clinical partners in collaborative research are discussed, alongside the future challenges and opportunities of biomarker discovery for patient stratification using multi-omics data.

While most clinically validated biomarker models derived from omics data have been developed for personalized oncology, first applications for non-cancer diseases show the potential of multivariate omics biomarker design for other complex disorders. Distinctive characteristics of prior success stories enable the derivation of specific recommendations for future studies. Five cancer biomarker models are discussed and show that reverse engineering of designs for complex macromolecular samples can successfully characterize the molecular features of polygenic diseases such as cancer. However, the competitive advantage of a stratification approach ultimately hinges on genuine understanding of disease-related processes. Given the knowledge gaps and the availability of new multi-omics and AI technologies, ample opportunities for discovery currently arise.

5.3. Clinical Utility of Biomarkers

While tremendous advances have been made in knowledge and rapid assay developments for omics profiling, there are still significant barriers to clinical translation of candidate biomarkers, especially for complex diseases [3]. These barriers include (a) a lack of standardization and rigorous assessments of assay features like specificity, sensitivity, dynamic range, and reproducibility, which is particularly relevant for "omics" techniques, owing to the complexity of the sample processing protocols and the inherent heterogeneity of biological specimens; (b) difficulty in reconciling the often modest performance of potential classifiers and their limited clinical utility; and (c) the failure to provide actionable insights that lead to clinically meaningful outcomes. Broadly speaking, the vicious cycle of desperate need and disappointingly few successes largely arises from a lack of rigorous statistical underscore at all steps of the biomarker discovery cycle and broad over-reliance on a small number of methods that have quality control and statistical interpolation with ad hoc selection of methods. Therefore, success in continuously diving deeper into the biomes is likely to depend critically on increased scrutiny and rigor for such methods. New biomarkers need to be developed utilizing tailored assay technologies and subject to rigorous and thorough evaluation. A multi-pronged approach is required that will take into account the high level of complexity and uncertainty in the pre-analytical, technical, and biological data. Standards and controls shall need to be developed that will enable and facilitate the generation of high-quality data by all parties. Bioinformatic and statistical methods need to be rigorously assessed so that the sensitivity and specificity of biomarker signatures are optimized. Rigorous validation of biomarker assays through clinically relevant mechanistic studies shall be key for translation.

6. Case Studies in Biomarker Discovery

Advancements in multi-omics data acquisition technologies together with machine learning-powered analytical methods have motivated the development of a novel generation of integrative biomarker discovery studies. While most of these studies focus on patient stratification, only few are concerned with the investigation of complex diseases and associated diagnostic biomarker signatures. Among a large number of publications dealing with assay, method, or test development for the measure of circulating biomarkers, only a small part actually identifies and characterizes novel peptide, lipid, or metabolite biomarkers for diagnostics, and presents first validation studies. Many groups are investigating biomarker panels for differentiated diagnoses of complex diseases. This offers the chance to enlarge available resources regarding biofluid biomarkers, broaden perspectives on the design of biomarker discovery studies, identify success criteria, and facilitate collaborative view on developments in the field [17]. Three specific disease areas were targeted for the selection of illustrative case studies, namely dialysis-related complications, metabolic disorders leading to obesity, and liver disease. Subsequently, a further limitation was applied to consideration of human studies only, which focus either on untargeted multiplexed omics approaches or inferential proteomics studies. Finally, some high-profile papers with large data sets were prioritized also to highlight the challenges in the field that arise with such resource demanding instrumentation. For each selected case study a general introduction is given together with an overview of devices, omics platforms, and assays used therein. Next, the key elements of sample preparation, analysis, and data pre-processing are reviewed while carefully addressing specific considerations not typically propagated in details in earlier work. Subsequently, the integrative analysis of multi-omics data and modelling of discovery biomarker panels are outlined, followed by a description of the validation procedure employed in the respective study. Finally, each case study is concluded with a discussion of key insights and challenges encountered.

6.1. Cancer Biomarkers

An essential challenge of cancer research is identifying cancer type without the need for patients' biopsy. Considered classically as cell-type-specific, distinguished noncoding RNA and protein biomarker types deliver individual signatures that are highly devoted to corresponding cancer types, conferring robust discrimination against molecular and clinical challenges. Here, a noncoding RNA-based cancer biomarker panel together with a machine-learning neural network is presented to classify RNA-sequencing data from different cancer types. Benchmarked on 163,859 samples of 33 cancer types, the deep learning neural network sufficed to achieve state-of-the-art discrimination of 1,442 RNA-sequencing data of conditional NN- and PCR-measured forms, outperforming the most feature-rich present biomarker type. Comprehensive runs of equal conditions supported a pan-cancer approach, and a stage-specific biomarker subset competed computationally with default essential, powerful noncoding RNAs, microRNAs, and proteins. With the engagement of the tested old stock market market indices, successful multi-asset class detection was drawn [19].

Machine learning can analyze biomedical big data generated from nowadays advanced technology. As an essential subfield of artificial intelligence, computer science, and data science, it aims to replicate and simulate the human brain to solve complex problems. The strength of machine learning lies in automatically recognizing the pattern of new, unseen data that were not provided earlier. It can also enhance the understanding of biological systems by integrating, analyzing, and visualizing big datasets and screening potential drug targets [2]. This presents an account of machine learning methods and tools applied to analyze multi-omics big data in cancer research, including clinically predictive algorithms to access personalized therapies and genomic epidemiologic studies supporting precision prevention strategies.

6.2. Cardiovascular Disease Biomarkers

Understanding the genomics of cardiovascular diseases (CVDs) is a relatively new field of research, yet the pressing need for such investigations to aid precision medicine is well-recognized. Given the intricate characteristics of CVDs, it is important for scholars and medical practitioners to make new discoveries regarding CVDs that lead to personalized interventions, even for patients suffering from similar disease classes. The appropriate utilization of machine learning (ML) methodologies can yield novel understandings of CVDs. Important insights include enabling improved personalized treatments based on predictive analysis. In this model, newly utilized transcriptomic biomarkers with strong evidential significance are anticipated to exhibit additional power for publicly available gene expression datasets. A series of transcriptomic biomarkers for the discovery of CVDs were domestically studied and advanced to this point, yet evidence across further cohorts remains elusive [20].

In this study, a novel combination of traditional statistics and cutting-edge AI/ML techniques was proposed to identify significant biomarkers by analyzing the complete transcriptome of CVD patients. First, gene expression data from 1,052 patients, encompassing complete data and deep clinical features, were accessed. After robust gene expression data pre-processing, three statistical tests were utilized to assess the differences in transcriptomic expression and clinical characteristics separately between healthy individuals and CVD patients. To help facilitate the development of unbiased models, the transcriptomic features were categorized. Then, using the recursive feature elimination (RFE) classifier, additionally termed the Random Forest model, the transcriptomic features were modeled independently to assign rankings based on how they relate to the case–control variable. The top ten percent of commonly observed significant biomarkers, based on statistical and ML questioning, were evaluated using four unique classes of ML classifiers. Each ML classifier was hyper-parameter-tuned, and after maximizing recall metrics through test/validation split, they were ensembled to accurately differentiate between patients and healthy individuals.

In this study, a newly developed approach was demonstrated to robustly unveil novel transcriptomic biomarkers that play important roles in the discovery of CVDs. Unbiased development of competing classifiers was conducted, and the highest accuracy and robustness were achieved. As a result, 18 transcriptomic biomarkers were uncovered that highly significantly distinguished the CVD population used in this study to predict disease with up to 96% accuracy. Further, the results were cross-validated using clinical records collected from patients of the same cohort. The clinically derived variables were generally highly significant in drawing important conclusions regarding disease prevalence. To summarize, 18 highly significant biomarkers, especially CAMK2N1, GPR137B, and MAP2K3, were uncovered and served as potential indicators for the early detection of CVDs.

6.3. Neurodegenerative Disease Biomarkers

With the continuous aging of the population, neurodegenerative diseases are becoming a major burden of society. With the progressive loss of neurons, neurodegenerative diseases typically manifest as cognitive disorder, movement disorder, both cognitive and movement disorder, dysphagia, or ataxia [21]. The most extensively studied neurodegenerative diseases are AD and PD. As the most common form of neurodegenerative dementia, AD accounts for about 60-80% of dementia cases in elders. AD is a progressive disease that starts long before the onset of symptoms. The clinical pre-stage presents mild cognitive impairment with subtle memory loss that may not interfere with daily living activities. Although it is hard to simulate daily life activities, blood biomarkers have gained more attention recently due to their convenience, less cost, and less risk sampling [22].

Blood plasma samples are easy to collect from patients. Candidate protein biomarkers are screened and quantified by an antibody array. Differentially expressed proteins are selected as input features of a multilabel classification model to distinguish the Alma subset of AD, MCI, normal controls, and discriminator samples. Cross-validation is applied to ensure the model fits the dataset well and test accuracy on unseen external testsets. To further investigate the relative stage in the disease when a sample is annotated as AD or MCI, the pseudo time series of the plasma samples is deduced and profiled based on well-fitted Langevin process. Neurodegenerative diseases manifest as a progressive loss of function or structure of neurons that results in dysfunction and death of neurons. The death of neurons in the nervous system is irreversible and ultimately leads to the functional loss of networks and organs, resulting in clinical symptoms.

7. Ethical Considerations in Omics Research

Omics research has become a prominent field of study in recent years, leading to the emergence of ethical considerations that were once rarely discussed. Ethical issues in omics must be viewed against the backdrop of established ethics in the life sciences, which progress from simple and straightforward to complex, obscure, and unstandardized [2]. The ethical issues in omics are similar because they affect the lives of many vulnerable individuals, which obliges researchers to move from the realm of facts to the realm of values. Omics research will lead to the addressing of a range of issues, creating societal moral concerns and specific ethical issues for each omic. In a wider approach, the general presumptions of modalities of disruption in society are necessary. Omics research will affect many lives, often in unforeseen ways. Omics research may lead to the unintended effect of disempowering many individuals, raising questions about how autonomy is situated in a societal paradigm of dissatisfaction. Omics for all will address issues of access, applicability, and ownership/responsibility. This will also lead to moral concerns about the fair distribution of omics benefits. Omics also carries risks of stigmatization, labelling, enhancement, and genetic surveillance, raising questions about its societal desirability. These normative questions about why, what, and how omnics should be pursued are necessary.

The acquisition of new research methodologies requires a substantial adjustment period. A transition from a heroic age of a new science to a normalized application of it is required. The academic rhetoric surrounding ground-breaking advances and the dwarfing of existing work is replaced by a slowly growing critical mass of literature. As the road forward is undertaken, voices of concern increasingly surface and principles for the responsible development of that research field are called for. In the past five years, the omics research space has matured considerably. In tandem with this growth, the literature on ethical, legal, and social issues has significantly expanded, indicating that omics research is now an established research field. Yet an unorganized patchwork of ethical considerations still persists.

7.1. Patient Privacy and Data Security

Integrating multi-omics big data and artificial intelligence (AI) technologies in biomarker discovery for healthcare has received great attention as it may facilitate early diagnosis and timely intervention of complex diseases. However, it also raises fresh challenges and concerns involving patient privacy and data security. The integration of these emerging technologies necessitates the collection, fusion, and analytics of massive data with high dimensionality, and the involvement of third parties, such as cloud computing service providers. This may cause data security/privacy problems, such as unauthorized parties accessing sensitive information in omics data and patient health records. Various approaches have been proposed to tackle those privacy concerns using privacy-preserving or secure data sharing or access mechanisms. For example, data anonymization is widely used in practice to protect individuals against reidentification attacks when sharing data containing individually reidentifiable information. Privacy-preserving mechanisms such as homomorphic encryption, secure multi-party computation, etc., can be adopted to enable AI-driven big data analytics while protecting data without compromising privacy. Prior to deployment of these security-privacy-preserving technologies in practice, many strict, responsible, and ethical requirements shall be fulfilled, including (1) governing proper use of data with strict compliance of laws, regulations, and business practices, (2) transparently informing patients how their personal data may be used, (3) preventing biased treatments arising from skewed database and models, (4) being traceable and verifiable by an auditing entity during a service evaluation. All these mechanisms aim at ensuring patients' trust over the responsible use of their personal information/assets.

One effective solution to provide patients with secure access to distributed medical data while preserving their augmented privacy is the Secure Patient-Side Privacy Architecture. It uses paged compression to minimize the patient information retrieved at the remote end and a ternary extended Key-Hash-Message Authentication Code to preserve privacy in the compressed access patterns. With the SPPA approach, the patient's right of transferring and managing his/her health secrets is embedded in the user end. Patients are able to ensure and manipulate the privacy of their health data in the remote data service without introducing any complexities to the data service providers. This will open up tremendous prospects of health analytics and ensure full patient privacy at the same time. A major obstacle for collaborative research projects analyzing large-scale heterogeneous patient data from different clinics is that individual hospital information systems each contain patient derived confidential data. Here, a secure pseudonymisation system protecting access to personal healthcare data that allow for the discovery of patterns is presented. It permits the assembly of a range of patient data, from genomics, clinical records, and medical imaging, while respecting privacy and encouraging collaborations amongst hospitals and researchers. This allows for the subsequent matching of patient data with the available analysis tools utilising high performance computing resources to facilitate the discovery of insightful information.

The project now assembles a range of patient-derived data, from genomics, via clinical records, to imaging, whilst respecting privacy and encouraging collaboration among hospitals and researchers. For a better understanding of this assembled data, the secure access system has to be interfaced with the data analysis workflows used for clinical decision support. [23][24][25]

7.2. Informed Consent in Research

Ubiquitous health data—resulting from habitual monitoring through personal devices, new collection technologies and partly because of regulatory investments—are predicted to be widely mined using Artificial Intelligence (AI) algorithms. By discovering population-wide health risks, urgent and relevant public health measures are envisioned. Such predictions can motivate health-promoting behavior in individuals and clinical diagnosis. Nevertheless, essential ethical questions arise. Using a European citizen panel approach, this paper reports on citizens' attitudes towards research based on mining ubiquitous health data. The citizen panel (N = 27), convened for an iterative series of informed discussions and deliberations, first highlighted ethically important issues. Its subsequent qualitative analysis provides insights into three focal considerations with a strong influence on a variety of views: (i) sharing ubiquitous health data for research purposes is almost acceptable as long as trust is established, (ii) actors ought to be trustworthy, ideally reflecting the public good, and (iii) citizens regard a clear elaboration of terms like 'data mining' and 'individually impossible disclosure' as necessary precursors for donation decisions [26].

8. Future Directions in Multi-Omics and AI

Biomarkers play a critical role in the early diagnosis of complex diseases. However, their discovery is often hindered by biological complexity and sample limitations for validating results. Integrative multi-omics approaches provide a viable solution, but the potential to remedy the overwhelming data scenario is hampered by the need for suitable AI models. AIMD aims to tackle this challenging scenario by developing an integrative AI framework, covering all steps from data processing to predictive biomarker discovery and validation for large-scale cancer multi-omics studies. AIMD has been successfully applied for breast cancer subtype identification and gene signatures (feature) detection across multiple platforms. The new article aims to boost future research in the area with deep details on multi-omics data acquisition, AI models for integrative learning, and computational tools for future applications. New parameter tuning and model development will be added to enhance model efficiency and robustness in future works. It can create a new bioinformatics platform through collaborations for intuitive and automatic model design and hyper-parameter tuning. Integrative studies aided by multi-omics analysis, mathematical modeling, and deep learning-AI-based strategies will lead to the development of a more comprehensive understanding of cancer metastasis. Multi-omics data is multi-dimensional in nature and "big" in size. The data need to be stored anonymously maintaining quality. The use of omics data is mostly limited to transcriptomics, copy number variations, and DNA methylations because of their abundance in different data portals. Repositories like store and share different types of transcriptomics and genomics data. The Clinical Proteomic Tumor Analysis Consortium provides proteomics data corresponding to TCGA cohorts. The precision medicine initiative, launched in 2015 in United States, aims to shift from "one-size-fits-all" treatment to tailored treatment for cancer patients. Precision medicine uses a more individualized molecular approach and enriches pharmacogenomics. This individualized approach requires the assembly and analysis of the individual's molecular signatures, which could be manifested in the form of multiple types of omics data representing the status of various biomolecules for this individual. AI and other deep learning tools and techniques can be utilized to optimize the utilization of patients' derived multi-omics data to extract target bio-entities and fit the targets with drug–target interaction data to extract relevant drugs and doses in the omics data landscape. Technologies like nanotechnology are boosting the attempt to targeted drug delivery. This well-defined approach is beneficial for discovering "new" drug candidates for targeted therapy of ribonucleotide reductase inhibitors (and other enzymes that catalyze nucleotide triphosphates dephosphorylation to protect cells from hyperphosphorylated RNA species) in viral and other diseases (). [27][28]

8.1. Emerging Technologies

Over the past several years, the rapid development of novel sequencing technologies broadened the range of biological layers accessible for high-throughput measurements and consequently increased the amount and complexity of primary biological data. In principle, relevant biological information can be obtained from genome-wide measurements of nucleic acids, RNAs, proteins, and metabolites thanks to rapid advancements in appropriate biosensors based on sequencing, chromatography, nuclear magnetic resonance, and mass spectrometry. Owing to possible omics layers being numerous and data complexity being daunting, integrative data-driven analysis of multi omics matrices has become a prominent and ambitious field of study. “Multi-Omics” denotes sets of data matrices preprocessed from different modalities that profiled sample-to-sample variations of abundances of different molecular entities. A typical Multi-Omics application applies an analytical pipeline to preprocessed data matrices for questions of scientific interest and biological relevance. Since both omics datasets of any types are in matrices of samples × features (entities), they can be analyzed in a pairwise manner by relying on data integration or distilling approaches that produce sample-proxy matrices. The newly produced sample-proxy matrix from two entities are then brought into consideration of data-driven exploratory tasks such as clustering and discriminative analysis.

Software tools and servers aim at effectively applying multi omics approaches on systems biology questions or medicine problems. Data merging tools accompany a server and a standalone software for merging, integration, and network analysis of paired omics data. A comprehensive data mining and visualization server for an arbitrary number of individual omics datasets is available. Knowledge-based approaches implement pathway enrichment analysis for gene networks, textual databases, miRNA-gene interaction databases, or drug-target databases, and so on. There are also some general Multi-Omics tools that focus on machine-learning-based approaches and can be integrated pipelines. Multi-Omics analysis extends biological insights to more explanatory interpretation models, yet inevitably introduces a leak of reproducibility. Existence and co-development of public Multi-Omics data with thorough omics profiling of the same sample across diverse modern cloud-out infrastructure and analysis strategies facilitate transparent description of every specific data-driven analysis. Recently, a comprehensive implement tool has been made available to automate the adaptive generation of the corresponding reproducible analysis protocol based on scripts for every general application with multi omics data. [29][30][31]

8.2. Integration with Clinical Practice

Precision medicine aims to empower clinicians to predict the most appropriate course of action for patients with complex diseases like cancer, diabetes, cardiomyopathy, and COVID-19 [32]. Multi-Omics, Clinical, and Biological Analytics can provide insight into the underlying biology of diseases beyond the single-dimension evaluation of clinical or omics data. Understanding the patient's metabolomics and genetic make-up in conjunction with clinical data will significantly lead to determining predisposition, diagnostic, prognostic, and predictive biomarkers ultimately providing optimal and personalized care for diverse diseases. In clinical settings, a multi-omics profile can be provided for patients with a complex disease, thereby determining coordinated clinico-multi-omics analytics power for a timely model of clinical and multi-omics data to find statistical patterns to identify underlying biologic pathways, modifiable risk factors, and actionable information that support the early detection and prevention of complex disorders, and the development of new therapies for better patient care. It could tap into the colorful spectrum of biological, clinical, lifestyle, and environmental information prior to long and costly adverse outcomes to finely segment at-risk populations through tailored, coordinated monitoring and logistical and behavioral intervention for wellness and the attenuation of cardiometabolic risk factors long before the onset of disease; the outcome of using a multi-omics-integrative platform could ultimately revolutionize medicine, re-positioning it from a symptom-driven “reactive” approach, to a predictive “proactive” one to identify modifiable risk factors and enable earlier interventions.

Finding less invasive methods to improve diagnosis and prediction of diseases are an important goal of biomarker identification. Translating these research insights into clinical practice has the potential to improve health care [33]. However, every study conducted has its own set of requirements and constraints. Gathering all requirements and engineering a generic solution that fits all requirements seems infeasible. Furthermore, analysts and physicians emphasized the importance of solutions that fit into their individual workflows and facilitate interdisciplinary communication and analysis. This seems especially crucial since without tools that are tailored to their individual workflows, analysts and physicians could have a hard time utilizing research insights.

9. Conclusion

Bioinformatics-assisted approaches are sought for data collection, curation, storage, and analysis to discover reliable molecular signatures or biomarkers of complex diseases like cancer at an early stage. Especially proteomics, metabolomics, and transcriptomics have been claimed reliable to conduct screening trials for MS. As multi-omics data continue to grow exponentially in size due to increasing availability of high-throughput molecular tools from both academia and industry, advanced AI/ML-driven data integration and machine knowledge extraction approaches enable more reliable biomarker discovery for improved understanding of a disease. Application of advanced AI-driven data analytics in MS biomarker discovery is still in its infancy. Potentials are enormous to transform the current data-poor knowledge environment into knowledge-rich bio-complexity understanding and patient benefit.

The AI-driven data analytics and ML-driven knowledge discovery approaches presented here are expected to significantly improve the current data-to-knowledge translation system by using emerging joint data integration, analytics, and knowledge extraction techniques. The success of integrative analysis and multi-omics discovery of complex biomarkers strictly relies on machine readiness, completeness, and compatibility of the biomedical big data; appropriate integration, analytics, and knowledge extraction approaches; deep learning of appropriate thresholded data; scalable hard and soft realization of the selected candidate biomarker sets and their hard establishment in the in-house screening platforms; and QMS-compliant validation and commercialization of the discovered biomarker sets. AI-ML-based tasks that require human expertise and priori knowledge are highlighted.

AI/ML approaches could be fundamental to serve as supporting pieces or solution paths for more traditional methods or heuristics within a hybrid biodesign framework to maximize the compromise between interpretability, efficiency, and effectiveness of the multi-omics discovery methodologies proposed above including flexible architecture configurations, canvasses, and task instances. IMOMs would also be applied to integration and analysis of novel cancer multi-omics data types since, unlike transcriptomes, glycomics, and lipidomic data, methylomes, and metabolomes have not been yet studied deeply in multi-omics data integration and knowledge discovery. On the other hand, co-integration and co-analysis approaches for joint input and hidden multi-task models and multi-view models would be critically explored to use existing integration of data for preknowledge acquisition or knowledge refinement.

References:

1. A. Bhardwaj and K. Van Steen, "Multi-omics Data and Analytics Integration in Ovarian Cancer," 2020. ncbi.nlm.nih.gov
2. N. Biswas and S. Chakrabarti, "Artificial Intelligence (AI)-Based Systems Biology Approaches in Multi-Omics Data Analysis of Cancer," 2020. ncbi.nlm.nih.gov
3. V. Bhakta Mathema, P. Sen, S. Lamichhane, M. Orešič et al., "Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine," 2023. ncbi.nlm.nih.gov
4. S. Tabakhi, M. Naimul Islam Suvon, P. Ahadian, and H. Lu, "Multimodal Learning for Multi-Omics: A Survey," 2022. [PDF]

5. I. Subramanian, S. Verma, S. Kumar, A. Jere et al., "Multi-omics Data Integration, Interpretation, and Its Application," 2020. ncbi.nlm.nih.gov
6. A. CASAMASSIMI, M. RIENZO, S. ESPOSITO, A. Federico et al., "Transcriptome Profiling in Human Diseases: New Advances and Perspectives," 2017. [PDF]
7. M. Smelik, Y. Zhao, X. Li, J. Loscalzo et al., "An interactive atlas of genomic, proteomic, and metabolomic biomarkers promotes the potential of proteins to predict complex diseases," 2024. ncbi.nlm.nih.gov
8. A. K. Jain, S. A. Busgang, C. Gennings, et al., "Environmental toxicants modulate disease severity in pediatric metabolic dysfunction-associated steatohepatitis," *Journal of Pediatric*, vol. 2024, Wiley Online Library. [HTML]
9. A. A. Tamer, "Bacterial Metabolites in Liver Fibrosis: Integrating Metabolomics with Clinical Applications," EC Microbiology, 2025. ecronicon.net
10. M. Spick, H. M. Lewis, C. F. Frampas, K. Longman, and others, "An integrated analysis and comparison of serum, saliva and sebum for COVID-19 metabolomics," *Scientific Reports*, 2022. nature.com
11. A. Termine, C. Fabrizio, C. Strafella, V. Caputo et al., "Multi-Layer Picture of Neurodegenerative Diseases: Lessons from the Use of Big Data through Artificial Intelligence," 2021. ncbi.nlm.nih.gov
12. M. J. Iqbal, Z. Javed, H. Sadia, I. A. Qureshi, A. Irshad, et al., "Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future," *Cancer Cell*, vol. 39, no. 1, pp. 1-15, 2021. springer.com
13. D. Painuli and S. Bhardwaj, "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review," Computers in Biology and Medicine, 2022. [HTML]
14. M. Sufyan, Z. Shokat, and U. A. Ashfaq, "Artificial intelligence in cancer diagnosis and therapy: Current status and future perspective," Computers in Biology and Medicine, 2023. [HTML]
15. B. Adamson, M. Waskom, A. Blarre, J. Kelly et al., "Approach to machine learning for extraction of real-world data variables from electronic health records," 2023. ncbi.nlm.nih.gov
16. S. Sheikhalishahi, R. Miotto, J. T Dudley, A. Lavelli et al., "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review," 2019. [PDF]
17. E. Glaab, A. Rauschenberger, R. Banzi, C. Gerardi et al., "Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review," 2021. ncbi.nlm.nih.gov
18. M. Leclercq, M. Leclercq, B. Vittrant, B. Vittrant et al., "Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data," 2019. [PDF]
19. A. Wang, R. Hai, P. J Rider, and Q. He, "Noncoding RNAs and deep learning neural network discriminate multi-cancer types," 2021. [PDF]
20. W. DeGroat, H. Abdelhalim, K. Patel, D. Mendhe et al., "Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine," 2024. ncbi.nlm.nih.gov
21. J. Zhang, X. Zhang, Y. Sh, B. Liu et al., "Diagnostic AI Modeling and Pseudo Time Series Profiling of AD and PD Based on Individualized Serum Proteome Data," 2021. ncbi.nlm.nih.gov

22. M. Karaglani, K. Gourlia, I. Tsamardinos, and E. Chatzaki, "Accurate Blood-Based Diagnostic Biosignatures for Alzheimer’s Disease via Automated Machine Learning," 2020. ncbi.nlm.nih.gov

23. Z. Dudová, N. Conte, J. Mason, D. Stuchlík, R. Peša, "The EurOPDX Data Portal: an open platform for patient-derived cancer xenograft data sharing and visualization," BMC Genomics, vol. 23, no. 1, 2022. springer.com

24. T. Gao, X. He, J. Wang, J. Liu, X. Hu, C. Bai, S. Yin, et al., "Self-assembled patient-derived tumor-like cell clusters for personalized drug testing in diverse sarcomas," Cell Reports, 2025. cell.com

25. P. Belleau, A. Deschênes, N. Chambwe, et al., "Genetic ancestry inference from cancer-derived molecular data across genomic and transcriptomic platforms," *Cancer*, vol. 2023. aacrjournals.org

26. M. C. Rivas Velarde, P. Tsantoulis, C. Burton-Jeangros, M. Aceti et al., "Citizens' views on sharing their health data: the role of competence, reliability and pursuing the common good," 2021. ncbi.nlm.nih.gov

27. S. Tang, K. Yuan, and L. Chen, "Molecular biomarkers, network biomarkers, and dynamic network biomarkers for diagnosis and prediction of rare diseases," Fundamental Research, 2022. sciencedirect.com

28. J. Doroszkiewicz, M. Groblewska, and B. Mroczko, "Molecular biomarkers and their implications for the early diagnosis of selected neurodegenerative diseases," *International Journal of …*, 2022. mdpi.com

29. M. Akiyama, "Multi-omics study for interpretation of genome-wide association study," Journal of Human Genetics, 2021. [HTML]

30. A. Hussein, M. Prasad, and A. Braytee, "Explainable AI Methods for Multi-Omics Analysis: A Survey," arXiv preprint arXiv:2410.11910, 2024. [PDF]

31. A. Morabito, G. De Simone, R. Pastorelli, "Algorithms and tools for data-driven omics integration to achieve multilayer biological insights: a narrative review," Journal of Translational Medicine, vol. XX, no. YY, pp. ZZ-ZZ, 2025. springer.com

32. Z. Ahmed, "Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis," 2020. ncbi.nlm.nih.gov

33. M. Höhn, H. Lücke-Tieke, J. Burmeister, and J. Kohlhammer, "Towards medhub: A Self-Service Platform for Analysts and Physicians," 2023. [PDF]