Cross-Modal Data Fusion and Pathway Activity Inference: Machine Learning Approaches to Integrated Multi-Omics Data Analysis for Biological Discovery

Dr. Maria Cláudia Barbosa, Associate Professor of Computer Science, Federal University of Minas Gerais (UFMG), Brazil

1. Introduction to Multi-Omics Data Integration

Cellular phenotypes emerge as a result of complex interactions between a wide variety of molecules. To understand the performance of these biological systems may require a comprehensive knowledge of genetic, transcriptomic, epigenetic, proteomic, and many other mechanisms involved in cell regulation and function. Therefore, correlating gene expressions at the transcriptional level with their counterpart proteins resulting from these records can lead to a better understanding of cellular function. Furthermore, other omic data, such as the DNA and RNA sequences and proteins' combinatorial and post-translational modifications, also provide knowledge regarding molecular structure and functions. Integrating these molecular datasets may reveal the underlying complex biological systems of human health and disease, as well as drug behavior and interactions in the body. Multiple omics methods have the potential to provide a more comprehensive view of cellular systems than traditional single-layer approaches.

Without a doubt, one of the key requirements for research and clinical applications is the integration of multi-omics data layers. However, multi-omics data vary in terms of data generation technologies, generated data resolutions, and sources of technical variation and systematic biases. Therefore, integrating such huge and diverse data is not a trivial task. In this respect, the fields of data science, mathematics, and statistics strive to integrate, model, and identify the patterns hidden in this complex multi-omics landscape. For instance, machine learning algorithms are employed to integrate multi-omics data and identify the corresponding multi-omics clusters, which enable better understanding and personalized treatment of multiple sclerosis. The first step of unlocking these opportunities is the integration of multiple omics data with data

analysis methodologies and techniques. The need for rigorous methodology development for multi-omics integration is still evident.

1.1. Definition and Significance of Multi-Omics Data

The origin of the term "omics" is derived from "system biology," aiming to gain knowledge of the biology of a whole organism or subsystem by collecting, displaying, and providing recommendations concerning the complete set of a certain biological molecule. As a result, omics studies aim to identify the functions and interactions of individual biological components in a holistic manner and focus on finding tendencies or patterns over time or space. In particular, the term "multi-omics" refers to the integrated analysis of the relationships and interactions of multiple omics data types or "omic layers" in order to gain a "pan-omics" insight that combines several types of biological knowledge. Gene integration, followed by proteomic, transcriptomic, genomic, and metabolomic analysis, was then performed to look for evidence of interactions. This deep profiling of various biological activities has enormous potential and holds broad applications for personalized medicine, precision medicine, predictive medicine, and participatory medicine. This kind of omics study may serve as the driver of the future of biology for the whole organism. The individual components of interest are observed independently within a single omics layer, but when we look at all of the relationships, an additional layer of information is revealed. As a result, it has been suggested that omics layers are neither repeatable nor redundant. In order to deal with multi-omics studies, levels of high-quality multidisciplinary efforts and large-scale measurements are necessary. Only then can researchers obtain high-quality biological data and utilize experimental and bioinformatics analyses. Overall, the multi-omics approach is critical in developing futuristic prediction methods that might drive clinical approaches in future biomedical research and the era of precision medicine.

1.2. Challenges in Integrating Genomic, Transcriptomic, and Proteomic Data

Integration of data primarily at the genomic, transcriptomic, and proteomic levels encounters several challenges at the molecular, biological, statistical, computational, and analytical levels. Differences in scale, type, and measurement techniques of the omics data platforms contribute to data heterogeneity and result in integrative systems biology with its own set of challenges and considerations. The biology itself is complex, leading to intertwined mechanisms of biological networks across the omics data types that are

characterized by unobserved parameters. In turn, inference drawn from functions, meanings, or any downstream biological interpretation or insight can be biased or not generalizable when using different platforms or molecules. From a computational and statistical standpoint, large-scale omics projects generate a vast amount of information in various dimensions with a low signal-to-noise ratio. Moreover, biological metadata coming from different consortia or laboratories exhibit inconsistency in standardization and curation, leading to additional noise in the data matrix. Consequently, the need to consistently integrate, harmonize, and disentangle such datasets from multiple biological and information systems levels is required. In practice, missing data can be attributed to analytical noise, unobserved biological mechanisms, and variance in data quality and measurement instruments, standardization norms, and protocols. These are challenges in the biological field of multi-omics integration that lead to research questions and objectives. The sheer complexity that arises when dealing with large-scale biological data requires the development of innovative and inclusive approaches, accompanied by robust statistical and machine learning techniques to propagate an integrative understanding of biological systems. These can ensure that insightful knowledge is extracted from multi-omics sources and avoid epistemic biases and pitfalls in reporting and communication.

2. Machine Learning Fundamentals

Multi-omics analyses generate multi-omic layers that require machine learning to tease out potential biologically relevant joint patterns of associations. Within the machine learning framework, there are three main paradigms: supervised learning, unsupervised learning, and semi-supervised learning. The choice of machine learning approach depends on various considerations, including the complexity of the questions researchers aim to answer, the structure and nature of the data, the size and features of the datasets, the desired interpretability and scalability of results, as well as the presence of a true outcome of interest. Dimensionality reduction by summarizing the patterns of the multi-omics data, usually high-dimensional in nature, through feature selection or feature extraction strategies are well-documented examples of intermediate steps in the machine learning framework required for reducing the dimensionality of the data and dealing with the issue of the curse of dimensionality. Principal Component Analysis and t-Distributed Stochastic Neighbor Embedding are some commonly used approaches for dimensionality reduction that are extremely useful for visualizing clusters and

subclasses of samples based on their multi-omic profiles. Additionally, reducing the dimension may improve model training efficiency, generalizability, and stability, especially in datasets that have known co-correlation structures between the features, thus addressing the issue of multicollinearity among co-occurring features and addressing heteroscedasticity. Feature selection for multi-omics integration aims to reduce undesirable systemic noise present in high-dimensional omics data that result from the lab processing workflow, and the selection of only optimal atomic features is required for the development of data-driven models with good interpretability as well as improving model prediction performance.

2.1. Supervised, Unsupervised, and Semi-Supervised Learning

Machine learning can be categorized into three main learning paradigms: supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning, which is the dominant learning paradigm, uses historical data whose outcomes are known to predict new data for predictive modeling or regression analysis. In supervised machine learning, an algorithm learns from trained labeled data, which is a series of feature-label pairs. The models based on supervised learning have been frequently used in multi-omics and have proved their worth in various studies. Unsupervised learning is a type of machine learning used to draw inferences from data sets consisting of features without any labeled responses. Instead, it focuses on both the data and its properties that could underline the attractors, hidden variables, or some other intrinsic structures. Analogously, much of the biological data are unlabeled and typically include various data types in vast collections of samples. Unsupervised learning could be used for exploratory analysis and yield valuable insights by providing a detailed reduction of multi-dimensional datasets, assembling similar samples into groups, or ordering samples and features.

Semi-supervised learning comprises methods integrating labeled and unlabeled data for training a classifier. While supervised learning is more commonly used, semi-supervised learning is particularly beneficial for biological data analysis, which may be incomplete with labels. In essence, semi-supervised learning attempts to generalize based on a mixture of labeled and unlabeled samples and can help alleviate the labeling burdens. In omics research, the availability of labels for downstream analysis is often rare. Since biological data—in general and molecular data—in particular, is usually high-

dimensional with small numbers of samples, computational approaches of unsupervised, supervised, or semi-supervised learning have been broadly used for research and clinical-based analysis.

2.2. Feature Selection and Dimensionality Reduction Techniques

Feature selection and dimensionality reduction are crucial techniques in any machine learning pipeline. Feature selection aims to identify a subset of input features that is used for predictive modeling. The smaller subset of features only contains the most relevant predictor variables, which has the potential to improve model accuracy. Many feature selection techniques have been proposed and implemented into machine learning algorithms. Feature selection can be used to alleviate some of the difficulties arising from the collection, measurement, and analysis of large-scale multi-omics biological data for feature interpretation and visualization. A simple taxonomy of feature selection methods may be based on the manner in which candidate features are evaluated by splitting them into three principles: filter methods, wrapper methods, and embedded methods. Nowadays, omics technologies enable the collection of more data, which provides detailed insights into biological systems. Although more data is helpful in understanding biological systems, many datasets are collectively referred to as "high-dimensional data" or "big data." Indeed, the increasingly high dimension of data is one of the reasons that make analysis and modeling challenging. The high dimensionality of data increases computational complexity, making it difficult to develop a machine learning model and speculate on the properties and structure of data. The simple remedy to all these issues is to reduce the dimensionality of data without losing much information or finding a small subset of variables that possess the most discriminating information. Having a high number of features also incurs a risk of overfitting and removes interpretability of the machine learning model. The subsequent sections provide various strategies to tackle these attributes and focus on bettering the integration of multi-omics data.

3. Integration Approaches in Multi-Omics Data Analysis

A large number of approaches have been proposed to integrate the different levels of omics data. In general, these methods can be grouped into two main classes: concatenation-based integration and network-based integration, which sometimes intersect or draw from one another. Concatenation-based approaches are the most

common and typically consist of the aggregation of the various levels of omics data into a single large dataset, which is then processed and analyzed jointly. This methodology allows for all information available across the different datasets to be used in completing large-scale data analyses and modeling, taking full advantage of existing statistical procedures and methods. However, biological signals can often be drowned out by data noise, some methods cannot address complexity within the system, and innovation in computational methods is currently outpacing the generation of new data, making it challenging to process all of the available data collected. Finally, network-based approaches integrate omic data into biological networks, which are used to model interactions and relationships between the sequences, proteins, metabolites, etc. present in the different omics layers. These methods are well-established and exploit the structure of biological networks to highlight or elucidate relationships between different levels of biological entities. Placing omic changes in the context of these networks can assist in detecting regulatory mechanisms, differences between normal and disease states, and model complex biological interactions. However, these approaches may not work well for all research questions or may be computationally demanding, and provide different network structures and relationship representations depending on their use case.

3.1. Concatenation-based Integration Methods

Concatenation-based methods are direct strategies that merge multi-omics data into a new dataset for integrative analysis. The simplicity and straightforwardness of this approach make it appropriate for non-longitudinal data, small sample sizes, and convenience for common analytical techniques in machine learning that are independent of data integration, as well as a reduction in data handling and computationally more tractable models. Several combination methods have been designed to integrate multi-omics data types using concatenation. The advantage of using concatenation in the mappings is that data can be handled by the same framework due to the same variable type and dimensions, to which simple case-control and latent variable analysis can be applied without complex constraints. Conversely, the limitations of concatenation are as follows. First, the use of the new dataset can lead to an information loss of the parent datasets because it may not contain relevant biological or population information. Second, multi-omics datasets with different input spaces may require handling of multiple levels of heterogeneity among tumors or cells, such as subtype classification,

and a single picked port or latent variable does not guarantee the optimal representation of different bio-omics spaces. As a result, the contribution in a new paper can be overshadowed by weakly coupled publication data. Although the addition was intended to be flexible and hence a great advantage, it is necessary to proceed with caution because it may result in data interpretation bias. There are various examples of these two strategies. For example, you can see the concatenation architecture designed for capturing multi-omics associations in the integrative analysis model, where all networks were connected.

3.2. Network-based Integration Methods

One approach to analyze multi-omics data is to explore various types of networks, using distinct edges that depict the interaction of different types of biomolecules commonly investigated through associations among various omics layers. This adds a biological level of interpretation, confirms deciphered networks, and provides insightful results when incorporated into the pathway and gene set enrichment analyses. Network analysis is beneficial in deciphering complex interactions among multi-omics datasets. It provides easy segregation of members from the same group by laying them out on local regions in the network, which in turn helps establish associations between exacted members, often missed by standard network-based approaches. It is also performed by various frameworks and tools that have directly or indirectly performed analyses in many multi-omics genomics and bioinformatic studies. However, a few challenges must be met, such as those related to the construction of the network due to its size and complexity, and analyses like centrality score determination.

Biological context and prior knowledge. Network-based integration approaches are widely used in multi-omics studies for the construction of networks. These methods use or hypothesize regulatory networks to incorporate different omics data. Integration may be conducted at the level of regulatory signals, activities of networks, or the phenotypic level. In any case, the inferred interactome is dictated by biological contexts that are either assumed or well-studied previously. Each gene is assumed to participate in the activity of pathways in which they are involved without assuming directionality of interactions. Builds upon machine learning to predict the regulation of genes resulting from network-based regulatory interactions. Any approach using the method is making the assumption that a protein-to-protein association network produced by a high-

throughput assay is a good proxy for the interactions that occur. Other approaches have made explicit use of gene-specific context and assumed that genes in the same pathway interact in both the transcriptomic and interaction network space more frequently than expected by pure chance.

4. Deep Learning Models for Multi-Omics Data

Introduction Several model-based methodologies have been developed to analyze and integrate multi-omics data emerging from advances in high-throughput experiments. More recently, deep learning models are among the advanced methodologies that are being developed and employed to integrate and analyze this data. Deep learning has been applied to many fields and has the potential to capture complex patterns and relationships in large-scale datasets. Deep learning models are becoming increasingly popular due to their capability to learn high-level abstractions from complex data and achieve state-of-the-art results across a vast range of tasks. Deep learning often refers to algorithms consisting of many layers of computation, including but not limited to convolutional architectural features, linear and nonlinear transformations, variational autoencoders, restricted Boltzmann machines, and generative adversarial networks, which allow systems to result in data-driven representations. The use of multi-omics deep learning models can identify hidden groups or networks for compounds associated with phenotypic similarities, prioritized genetic mutations, and the identification of driver genes that are potentially responsible for neoepitope formation and associations with histopathological response, progression-free survival, and overall survival. Modern machine learning techniques may integrate traditional machine learning models with deep learning models, leveraging both shallow and deep neural networks. Different machine learning techniques, architectures, and special strategies were tailored for multi-dataset analysis, in terms of data-driven or model-based methods. Traditional machine learning and data-driven network-based approaches characterize the potential to learn multilayer hierarchies of data-driven features or network links hidden in the dataset, but fail to learn informative hierarchies in the face of complex datasets. Deep learning can indeed surpass shallow networks by learning deeper architectures that exploit not only diverse omic data representing different information levels but also multi-modality information connected with the heterogeneous molecular level data. Particularly, data-driven omic models without other information's interpretability effects are not widely applied in research, but their use could be fostered over the years

to come. Thanks to new findings within research perspectives in data diversity, deep learning could lead to new avenues to improve analysis approaches of omic studies.

4.1. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are at the core of deep learning models and have raised great interest in multi-omics data integration over the past years. CNNs are of great interest for their ability to learn spatial hierarchies of varying complexity. Their original application was for image-like structures, and given the nature of some types of multi-omics datasets, particularly genomics and proteomics, they can be directly applicable to these data. Several adaptations have been proposed for extending a sequential CNN to multiple inputs, with shared and unshared layers across the input channels. The design of the shared and unshared networks may vary. Some simply concatenate the input tensors into a single tensor and then process using convolutional layers with the proposed filter kernels. This type of approach can also be combined with multi-modal learning, further incorporating a joint block, such as a fully connected or second CNN layer that processes the features that are learned from the shared multitask layer.

The main advantage of a CNN is the ability to perform feature learning and representation in a sequential manner, each time learning complex features from smaller and simpler patterns. This is a pivotal feature if we aim to capture sub-patterns in high-dimensional data. By adjusting the metrics, CNN models can allow for the identification of a typical pattern in multi-omics data or can even perform analysis under a different range of scales simultaneously. Therefore, convolutions can summarize the whole dataset into a single number or sequence of numbers in a manner that captures biological information while discarding complex influence on the data. In addition to that, the use of CNNs in genomics and proteomics typically enhances the model's predictive performance and can shed deeper biological insights by visualizing convolved patterns and filters.

4.2. Recurrent Neural Networks

In recent years, specialized deep learning models for sequential data emerged, which are known as Recurrent Neural Networks (RNNs). RNNs are introduced for sequential and temporal learning, which exist in multi-omics studies where the order of omics data samples is essential. The architecture of an RNN consists of a step, and it can also

maintain information across time steps, which means RNNs are suitable for handling dynamic biological processes. One example of omics that can be analyzed using RNNs is time-series data, which is prevalent for omics studies, such as genomics, transcriptomics, and proteomics analysis in a longitudinal study. RNNs are useful for learning the representation of temporal patterns that exist in time-series data. Furthermore, RNNs permit modeling and predicting the next sample within the series because of their capacity to perform sequential processing.

Another advantage of RNNs is their ability to work with a collection of datasets over time. While most machine learning models are only able to process data with a specific size, such as CNNs with fixed-size input image tensors or LSTMs with a fixed number of sequence steps, RNNs can process an entire collection of data that has various sizes. RNNs process the data in sequential order, where each data sample is placed by a time step index. This nature gives RNNs an advantage in handling continuously growing and evolving multi-omics datasets that are always updated and added with new data collections. Nevertheless, training such an RNN on multi-omics data is not straightforward because we may face an exploding and vanishing gradient issue that occurs when the product of large or small derivatives happens in the backpropagation during training. Several RNN architectures, such as LSTMs and GRUs, have been developed to overcome this problem. LSTMs are composed of memory cells that can maintain long-term memory, which makes the gradient of the error remain constant over time.

5. Case Studies and Applications

present a plethora of case studies in which integrative tools for multi-omics data can be applied. Specifically, they introduce a number of applications that have been or are continuously being developed. select several case studies based on the application of machine learning-based multi-omics data integration, categorizing these studies according to the biomedical question that they answer. For instance, tackle the problem of identifying subtypes in skin cutaneous melanoma with the help of multi-omics data integration. Another common biomedical question that the studies emphasized tackle is related to prognosis prediction. have developed ICPM (Integrative Cluster of Proteomics and Methylation) to identify different prognostic groups of cancer patients.

Interestingly, the prognosis in these categories is well linked to the immune system's response in the patient.

stress the concept of personalized treatment suggestions for a particular drug that targets specific pathways in particular patients. For instance, have developed an ensemble machine learning-based model to predict pemetrexed treatment response of solid cancer patients based on the proteomics data, methylation, and clinical features. The study demonstrates that multi-omics data integration is capable of yielding insightful recommendations for clinical medicine. Not only that, it also sketches groundbreaking biological discoveries. A case in point is the patient shingles or 'insultoma' in the cardiovascular system that was unveiled due to the multi-omics integration studies that either subsequently perturbed molecular networks or analyzed the molecular phenotypes. More case studies showcasing the practical importance of incorporating multi-omics data in integrative models can be found in each of the excellent review papers on multi-omics data fusion and analysis. In sum, the studies presented illustrate the journey from the aim of multi-omics data integration to personalized treatment strategy suggestions and their pioneering biological findings.

5.1. Cancer Subtyping and Prognosis Prediction

One of the clinical potentials of integrating multi-omics data is cancer subtyping and prognosis prediction. There are different levels of heterogeneity in cancer. Currently, histological methods are mainly used to classify cancer. Since cancer subtypes with different biological behaviors may have similar histopathological characteristics, integrated analysis of multiple omics data at different levels can provide more molecular information for more accurate categorization. Machine learning methods are powerful in mining deep data features. They have been widely used for multi-omics-based cancer stratification and prognosis prediction. Tumors with identical morphology may behave differently. Identifying and researching tumor properties is crucial for individual treatment decisions. Through heterogeneously integrated or homogeneous data, personalized subgroups are identified to distinguish the various population prognoses. Prognostic models may describe the role of each marker in grouping prognosis. Machine learning is a powerful method to predict patient survival by integrating tumor histological features and multi-omics data.

Because the proliferation of gene expression data is too massive to process, it is necessary to perform a comprehensive multi-omics integrative analysis to fully understand the underlying biology of the tumor. Many research studies combining multi-omics data can further stratify and develop prognostic models with better prediction performance. Several integrated multi-omics predictive model development studies are available, mostly in standard format and not easy to interpret or compare. When integrating multiple omics data, the predictive resolution and the ability to assess how much a predictive model can predict in the clinical setting without overfitting is required. Deep learning models integrate multi-omics data and present results in simple and understandable formats. Biosystems use multi-omics learning predictive analysis with datasets to develop a joint prediction model. The expression analysis has been published to predict the model with molecular and clinical data. The predictive approach also includes how to visualize and interpret the developed combined model for use in the clinic, which is a major advantage of the presented analysis. Studies must go beyond practicalities, such as predicting cancer subtype. In our future work, we will increase our sample size and collect data for various cancer types because this study can affect actual clinical practice.

5.2. Drug Response Prediction

Multi-omics data integration can further leverage precision oncology by predicting drug responses of a patient. In particular, the integration of genetic, proteomic, and metabolomic profiles can comprehensively present perturbations in cancer cells and tumor-infiltrating immune cells, providing a complete understanding of how a particular patient will respond to different therapeutic interventions. Based on these data, state-of-the-art machine learning models could be employed to predict the effect of specific treatments for individual patients by uncovering gene expression, protein, or metabolite patterns associated with treatment response or resistance.

Predictive models for drug response prediction were demonstrated to achieve superior predictive accuracy compared to random chance. However, they might struggle to uncover universally valid, biomedically interpretable, and generalizable features of unknown mechanistic relevance or investigate the underlying causality of the identified predictive features, representing future challenges to tackle. This approach paves the way for personalized medicine, in which tailored drug therapies are generated to

improve patient outcomes. Several case studies also show real-life applications of these predictive models. They predict tumor-intrinsic drug responses and stratify patients into high- and low-risk patient groups of recurring diseases for personalized disease management. However, using multi-omics data to accurately predict an effective drug treatment strategy is non-trivial due to the extreme variability and complexity of the patient mutational landscape and the tumor microenvironment.

Conclusion and Future Perspectives Can we infer an untreated patient's response to a particular treatment from their molecular profile using existing multi-omics prediction models for cancer patients and successfully transpose these initial results showing a marginally effective actual treatment based on an untreated molecular profile? Furthermore, more advanced methodologies that do not rely only on the selection of small subsets of features but directly integrate multi-modal data types in drug prediction models are being developed and will take off a large part of the proposed future research challenges.

6. Future Direction

As research in multi-omics data generation advanced, the cost of conducting these experiments decreased as well. The instruments employed for conducting these experiments have also multiplied during the last decade. With more affordability in conducting these omics and more data availability through various data repositories, we expect a lot of interesting research that will show an array of methodologies utilizing these data and their applications in various fields. Along with the advances in sequencing technologies and their data types, some bioinformatics and systems biology-based conferences are focusing on multi-omics data integration and analysis. There is a growing interest in the intersection of artificial intelligence, machine learning, and multi-omics research. There is a need to standardize the disbursement of data and their methodologies not only in the form of publications but for data sharing and algorithms aligned with multimodal data types and data generation technologies.

With technological advances in generating data, it is essential to focus on the algorithms or methodologies to be developed to process the data. This is essential in the application of multi-omics in disease or aging. These 'omics' are always heterogeneous, and integration is the first barrier that has to be addressed. Instantly identifying respective elements that differ in various criteria normally results in huge data sparsity. Machine

learning algorithms to handle such complexity may also have to be estimated. These are some open challenges in research. Additionally, estimates made using huge data are generally robust to the experimental bias introduced in using data from various platforms. Data standardization and representation of heterogeneity and protocols used for data generation still limit their reproducibility in peer-reviewed publications. Progress in various methodologies is expected, and the use of machine learning is rapidly growing in data analyses of different omics methodologies. Given the nature of multi-omics or polyomics, the possibility of personalized medicine is very high, generating great interest in healthcare research. With the incoming technology on the market and industry, we predict that in the next ten years, the polyomics field will show a paradigm shift that may reveal significant synergistic results in various interdisciplinary fields. We predict that developing a platform to represent and support multi-omics studies may hold a significant impact. Innovative tools are expected to be built in these platforms that are open and useful for data generation, sharing, or retrieval, to draw research conclusions that are interpretable and provide end-to-end representations. The results of these goals may prove to be transferable or replicable across the fields worldwide.

7. Conclusion

In sum, this essay reviewed different approaches to integrate multi-omics and machine learning. All these methods have been essential in improving our understanding of phenotypes by correlating with high-throughput molecular markers. The results obtained from a combination of omics datasets have been translated into putative drug targets, tumor subtyping, biomarkers, etc., which are essential for better clinical outcomes and personalized medicine. Many researchers have expressed the consensus that a combination of various layers of omics data is essential for a holistic view and better understanding of biological systems. This field is still in its development, and many challenges are being encountered. Many studies are being conducted daily to integrate the various pathways and functional links. It is concluded that such integrated analyses will provide a new way for biomedical researchers. Integration of multi-omics datasets supported and evolved the biological network approach. To move from individually published studies to deriving useful and experimentally testable hypotheses, we really need adequate amounts of large-scale and high-quality datasets in conjunction with robust analytical methodologies and bioinformatics pipelines.

Heterogeneity, interoperability, and scalability are the key challenges of multi-omics data integration. We need to develop enhanced and robust systems across various laboratory systems, study designs, and patient subgroups. We expect that all concepts and methodologies discussed will help in narrowing the gap between the interpretation of multi-omics and clinical research. The systems biology-based hypothesis needs to underpin these approaches from multi-omics data. We strongly believe that researchers should think about how multi-omics data is incorporated from the beginning of experimental design to the identification of causally related regulatory molecular phenomena.