GMR
The Original

MULTI-OMICS INTEGRATION FOR ELUCIDATING COMPLEX DISEASE MECHANISMS: A MACHINE LEARNING PERSPECTIVE

Dr. Indu Purushothaman[1], Dr. Shanmuga Priya M[2], Dr. Vinod Kumar P[3], Shitij Goyal[4], Ms. Niyati V. Thakkar[5], Jaskirat Singh[6]

[1]Assistant Professor, Department of Research, Meenakshi Academy of Higher Education and Research, Email: indu@maher.ac.in
[2]Professor, Pathology, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Enathur, Kanchipuram, Tamil Nadu – 631552, India, Email: spriyapatho@maher.ac.in
[3]Professor, Forensic Medicine, ORCID: https://orcid.org/0000-0001-9251-441X
[4]Quantum University Research Center, Quantum University, Roorkee, Uttarakhand – 247667, India, Email: shitij.qsb@quantumeducation.in, ORCID: https://orcid.org/0009-0002-5558-8238
[5]Assistant Professor, Faculty of Allied and Healthcare, Gokul Global University, Sidhpur, Gujarat, India, Email: nvthakkar.gpc@gokuluniversity.ac.in, ORCID: https://orcid.org/0009-0000-9270-140X
[6]Centre of Research Impact and Outcome, Chitkara University, Rajpura – 140417, Punjab, India, Email: jaskirat.singh.orp@chitkara.edu.in, ORCID: https://orcid.org/0009-0001-0914-4700

ABSTRACT

Complex diseases like cancer, neurodegenerative diseases, and metabolic syndromes develop on complex interactions involving biological processes through genomic, transcriptomic, proteomic, and metabolomic. Conventional single-omics techniques can tend to miss this complexity and result in incomplete mechanistic insight. Against this background, the multi-omics integration has come as a powerful tool to give an overarching picture of the disease biology, through integration of heterogeneous data at different levels of molecular biology. Nonetheless, multi-omics data are high dimensional, noisy, and heterogeneous, and they pose a big challenge in the analysis process, which requires sophisticated computational models. Machine learning algorithms, including classical algorithms like support-vector machines and random forests, deep learning, and network-based models, have shown great potential in revealing hidden patterns, finding biomarkers, and predicting the outcome of a disease. This review critically analyzes the existing multi-omics integration approaches, the importance of machine learning in enhancing classification and interpretability, and evaluation metrics including accuracy, F1-score, and ROC-AUC used to measure the performance of a model. The main insights are that integrative methods can greatly contribute to the knowledge of disease pathogenesis and allow to identify biomarkers more accurately compared to single-omics. In spite of these developments, there are issues of data standardization, model interpretability and clinical translation. To complete the gap between computational predictions and real-world biomedical applications, future studies should emphasize explainable AI, single-cell and spatial multi-omics, and scalable frameworks.

KEYWORDS: Multi-omics integration, Complex disease mechanisms, Machine learning, Genomics, Transcriptomics, Proteomics, Biomarker discovery, Data integration.

1. INTRODUCTION

Complex human diseases, such as cancer, neurodegenerative diseases and metabolic syndromes, are multifactorial in nature and occur as a result of complex interactions among many biological layers, including genes, transcripts, proteins and metabolites. Such diseases cannot be explained solely within the molecular lenses since their pathogenesis is dynamic regulatory networks and cross-level interactions that do develop over time. As an example, the heterogeneity of tumours in cancer or dysregulation of the molecular state in Alzheimer disease implies the coordination of changes at the genomic, transcriptomic, and proteomic levels and, hence, requires a systems-level analysis (Karczewski and Snyder, 2018; Hasin et al., 2017).

Single-omics methods are traditionally very useful in identifying particular molecular signatures; however, they have major deficiencies in the comprehensive nature of the disease mechanisms. These approaches can be rather fragmented in their insights and, they do not consider interdependences between various biological layers. Consequently, the use of either genomics or transcriptomics can result in partial or even erroneous explanations of the disease processes (Huang et al., 2017; Ritchie et al., 2015). Moreover, biological data are very high-dimensional and noisy, which also makes it more challenging to extract meaningful patterns when analyzed alone. In order to overcome them, multi-omics integration has become a strong paradigm, which is a combination of heterogeneous data sets of many different fields of the molecular world to offer one and complete picture of biological systems. The integrative methods allow discovering cross-omics interactions, enhancing the discovery of biomarkers, and understanding the progression and heterogeneity of the disease (Subramanian et al., 2020; Bersanelli et al., 2016). Similarity network fusion and latent factor models are examples of techniques that can reveal hidden connections between omics layers, and thus can be used to model diseases more accurately (Wang

et al., 2014; Vahabi and Michailidis, 2022). Moreover, new technologies of single-cell multi-omics have also broadened the scope of integrative analysis, being able to study cellular heterogeneity in unprecedented detail (Argelaguet et al., 2021; Baysoy et al., 2023).

In spite of these developments, multi-omics data integration and interpretation is extremely challenging because of problems like data heterogeneity, missing values and scalability problems. Machine learning, in this context, has become a vital facilitator of meaningful insights out of high-dimensional, intricate data. Deep learning models, traditional, support vector machines, and random forests are all powerful machine learning algorithms that can be used to perform feature selection, pattern recognition, and predictive modeling (Li et al., 2018; Ballard et al., 2024). In more recent years, the network based and deep learning methods have shown great ability of modeling some of the most nonlinear relationships and capture a biological complexity, which results in much better accuracy and interpretability of multi-omics analysis (Sartori et al., 2025).

Due to the swift development of multi-omics technology and computational algorithms, it is urgently necessary to develop an overview of existing methods which combine biological knowledge with modern machine learning methods. The purpose of this review is to critically review the current multi-omics integration strategies, to emphasize the role of machine learning in understanding the mechanisms of complex diseases, and to comment on the measures of evaluation that are usually applied to measure model performance. This work will help fill a gap between the computational methodology and biological interpretation in order to offer a systematic framework of future studies in integrative systems biology.

The review is a synthesis of multi-omics integration methods, to date, and more specifically an analysis of machine learning-based frameworks to study complex diseases. In contrast with a traditional review and its main emphasis on methods of integration, this work critically connects computational models with the biological interpretability and evaluatory measures to provide a holistic view that helps to develop methodological and methodological approaches as well as translate them into biomedical practices of interest.

2. METHODOLOGY OF REVIEW

The systematic and structured approach to this review was aimed at covering the existing literature on the topic of multi-omics integration and its application in the understanding of the mechanisms of complex diseases without any gaps. The methodology was created to locate, appraise, and combine all pertinent literature on computational strategies, especially machine learning-based algorithms, in the analysis of multi-omics data. High-quality peer-reviewed articles were identified through a thorough literature search that was conducted in large scientific databases such as PubMed, Scopus, and Web of Science. These databases were chosen since they have broad coverage of biomedical, computational biology and interdisciplinary research areas. A combination of keywords and Boolean operators was used as the search strategy to allow the search strategy to be as efficient as possible. Critical search terms were multi-omics integration, machine learning, deep learning, systems biology, complex disease mechanisms, biomarker discovery and data integration. These terms were varied and combined to capture all the studies that might be relevant, including both biological and computational sides of the problem.

In order to keep the relevance and quality of the review, inclusion and exclusion criteria were used. To be included in the studies, they (i) needed to be based on multi-omics data integration, (ii) they needed to use machine learning or computational methods, (iii) they needed to be a complex disease, which might include cancer, neurodegenerative, metabolic diseases, etc. and (iv) they had to be published in peer-reviewed journals. High quality review papers were taken into account (both original research articles and review papers) in order to offer a balanced picture. They eliminated studies that (i) only performed a single-omics analysis, (ii) were not published clearly methodologically, or (iii) were non-peer-reviewed sources that are either merely editorials, opinions, or abstracts of conferences without full-text accessibility.

The studies included in the review were mostly published in the past decade (20152025) to reflect the latest developments in the field of multi-omics technologies and machine learning methods and include a small number of earlier studies to give them a historical background. This period guarantees a trade-off between the classical practices and the new directions of deep learning and single-cell integration of multi-omics.

Though it is mainly a narrative review, there were also components of systematic review structure added to the work to improve its transparency and reproducibility. PRISMA-inspired selection process was adhered to; this included identification, the process of screening, eligibility assessment and eventual inclusion of the studies. Within the first segment, a wide range of articles had been found using the database and subsequently, they had to be filtered by deleting duplication and irrelevant records based on titles and abstracts. The next step was full-text screening, which was done to confirm adherence to the inclusion criteria. The last group of studies identified was used to conduct qualitative and thematic synthesis in this review. This systematic approach to methodology will give the review a broad, objective, and current synthesis of multi-omics integration strategies and machine learning techniques, and thereby allow solid insights into the complexity of disease mechanisms.

3. Overview of Multi-Omics Data

Biomedical research has undergone a great transformation with the development of multi-omics technologies that have allowed the analysis of biological systems on a number of molecular levels. In contrast to conventional single-omics methods, which emphasise the study of isolated components, like genes or proteins, multi-omics

analyzes a large number of data sets to give a systems-wide view of the mechanisms of complex diseases. This integrative approach is critical in the analysis of multifactorial diseases like cancer, neurodegenerative diseases, and metabolic syndromes where genetic, regulatory, and environmental factors interact in a very critical manner. Integrating the results of different fields of omics enables scientists to identify complex molecular interactions and regulatory networks that cannot be revealed by single analyses.

In its basic form, genomics offers information of the fixed genetic road map of a living organism with variations like the single nucleotide polymorphisms, insertions, deletions, and structural changes, which lead to the development of the disease. Nevertheless, functional outcomes cannot be fully explained with the help of genomic data since gene expression can be individually regulated. Transcriptomics is a technique that overcomes this drawback of capturing the RNA expression patterns, which amounts to gene activity under certain physiological or pathological conditions. Although transcriptomic data contains much useful information about cellular responses and the regulatory processes, it is very dynamic and varies with time and environmental conditions and it is more complex to analyze.

In addition to gene expression, proteomics and metabolomics provide essential understanding of functional and biochemical processes of biological systems. Proteomics deals with the analysis of proteins at large scale (protein abundance, structures, and post-translational modifications) and bridges the gap between molecular activity and cellular processes and pathways. Proteomic data, however, is usually hard to analyze because of technical constraints and the complication of protein interactions. Metabolomics, however, analyzes small-molecule metabolites, which are the final products of cellular processes, to present a picture of the physiological status of a system. The layer is extremely sensitive to environmental fluctuations and disease conditions and is thus very useful in biomarker-discovery research, but is variably sensitive and context-dependent, presenting further challenges.

Besides these layers, epigenomics is also applying a regulatory dimension by looking at changes in gene expression, which are heritable but do not require a change in the DNA sequence, e.g., DNA methylation and histone modifications. Combining such a variety of omics data is challenging because they are extremely dimensional, heterogeneous, and exhibit dissimilarities in data structure and scale. Missing values, batch effects and a small sample size are even more problematic on analysis. In turn, investigation of multi-omics necessitates sophisticated computational methods to be able to coordinate and analyze these datasets in order to have a more global apprehension of disease pathogenesis and be able to build precision medicine plans.

4. Multi-Omics Integration Strategies

Multi-omics data integration is a key procedure in deriving biologically significant information in multi-omics data. With the variety of data types, scales, and distributions between omics layers, a number of integration strategies have been created to successfully integrate this data. These methods can mainly be divided into early, intermediate and late integration strategies which vary based on the way and time of combination of the data provided by more than one source into the data analysis pipeline. It is necessary that a suitable strategy of integration be chosen as it has a direct impact on the performance of the model, understandability, biological meaning.

4.1 Early Integration (Feature-Level Fusion)

Early integration (alternatively feature-level fusion or feature fusion) is where features in multiple omics datasets are concatenated directly before analysis. This method combines genomics, transcriptomics, proteomics and other data sets upon preprocessing and normalization, and machine learning models can learn patterns based on the combined feature space. This is a reasonably simple technique to apply and allows capturing cross-omics correlations at feature level. Nevertheless, high dimensionality, noise effects, and variations between data distributions typically challenge early integration, potentially adversely affecting model stability and model performance.

4.2 Intermediate Integration (Model-Based Integration)

Intermediate integration, or model-based integration, is where multi-omics data is integrated into the process of modeling, as opposed to being integrated in an input or output step. This method usually uses sophisticated computational models like latent variable models, network-based models or deep learning models to learn common representations between multiple layers of omics. Other methods, including autoencoders, similarity network fusion, and multi-view learning models, are also in this category. Intermediate integration can capture the complex nonlinear relations and reduce dimensionality, which makes it perfectly suitable with high-dimensional biological data. Yet, such techniques are computationally heavy, and in many cases, may need to be carefully tuned and run on large datasets.

4.3 Late Integration (Decision-Level Fusion)

Late integration (also known as decision-level fusion) is an analysis of each omics dataset in isolation followed by integration at the decision stage. Under this method, the individual omics layers are trained on separate models whose outputs are combined via methods like voting, averaging or meta-learning. Such an approach brings the ability of flexibility in dealing with heterogeneous data and mitigates the effects of data incompatibility problems. Moreover, it allows using specialized models which are specific to any type of omics. Nevertheless, late

integration does not necessarily capture interdependencies between omics layers completely, and it may not be as capable of revealing underlying biological processes.

These three integration strategies are drastically different in the stage of data fusion and the complexity of calculations as shown in Fig 1. Early integration is simple and has scalability problems, intermediate integration provides greater biological insight at the cost of greater computational load and late integration is modular at the cost of potential cross-omics interactions. Comparatively, early integration is applicable with small datasets with nearly matching features whereas intermediate integration is more applicable with high dimensional complex datasets where the integrity of nonlinear relationships is central. The late integration has benefits whereby datasets are highly heterogeneous or the independent analysis pipelines are needed. Although each strategy has its advantages, there is no single strategy that works best across all research and no universal strategy that suits all research and experimental data and the needs of computation. Thus, the combination of several strategies into hybrid ones is currently also being considered to address the current limitations and enhance overall performance in multi-omics data analysis.

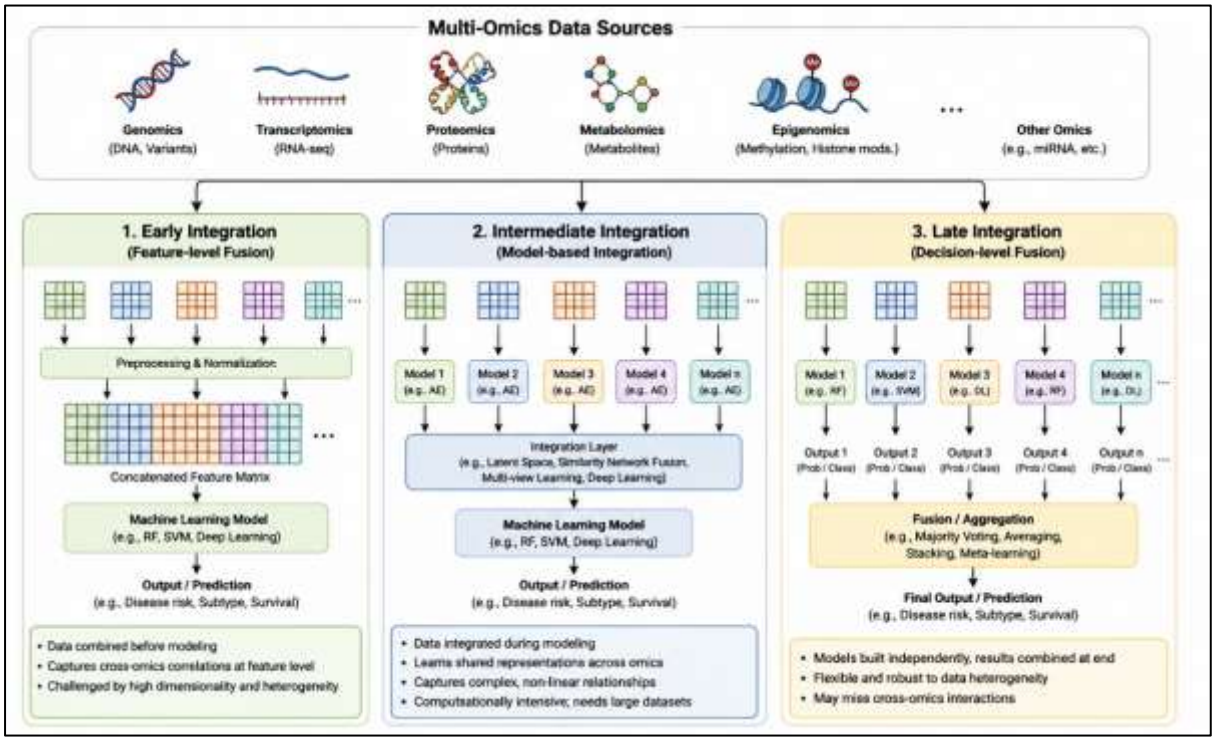

Fig 1. Multi-Omics Integration Strategies (Early, Intermediate, and Late Fusion).

5. Machine Learning Techniques in Multi-Omics Integration

The application of machine learning (ML) has become an essential part of multi-omics data integration and allows extracting meaningful patterns of high-dimensional and heterogeneous data. Multi-omics data is highly complex with nonlinear relationships and high feature space, making the advanced computational models to handle these complexities necessary. In addition to the ability to classify disease and biomarkerically explore diseases, ML techniques have been shown to increase the comprehension of complex biological interactions. As has been summarized in Table 1, various machine learning methods differ in ability, strengths and weaknesses as applied in multi-omics integration.

5.1 Traditional Machine Learning

Conventional machine learning algorithms have proven to be greatly used in the multi-omics studies because they are robust, interpretable and have fairly low computation costs. Random Forest (RF) is one of them and is typically utilized in the feature selection and classification activities because RF can handle high-dimensional data sets and offers information on the importance of features. The classification problems or the problems that can be successfully solved by Support Vector Machines (SVM) include the cases with small sample sizes, as the decision boundaries are determined in the high-dimensional spaces. In the same vein, k-Nearest Neighbors (kNN) is an easy but useful classification algorithm, based on similarity measures. Although they have these benefits, the conventional ML models can fail to identify multifaceted non-linear interactions among layers of various omics. Also, they may fail in their ability to work with very heterogeneous data, thus restricting their usefulness in multi-omics with full integration of data.

5.2 Deep Learning Approaches

Deep learning (DL) has become an attractive method of multi-omics integration because it can model nonlinear relationships and hierarchical feature representations are automatically learned. Autoencoders (AE) and

Variational Autoencoders (VAE) have become a common method of dimensionality reduction and latent feature extraction allowing the combination of various omics datasets into a common representation space. Deep Neural Networks (DNNs) are an extension of this, training more complex patterns with the benefit of accuracy to predictions in disease classification problems. As well, Convolutional Neural Networks (CNNs) which are more typically applied to image data have been extended to structured omics data to include spatial or hierarchical interactions. But deep learning models can be rather computation-intensive and need large datasets to work effectively. Moreover, they may be limited in the interpretability due to their black-box nature, as they are difficult to validate biologically and use in the clinical setting.

5.3 Network and Graph-Based Models

Networks give biologically relevant approaches towards integrating multi-omics through the representation of the relationships among molecular entities. Graph Neural Networks (GNNs) have received much interest due to their behavior of modeling complex biological networks interaction, e.g., the interaction between genes or proteins. These models use graph structures to learn representations that are indicative of the connectivity and functional relationships between biological components. Besides that, multi-omics information is regularly incorporated into a systems biology framework using gene regulatory networks (GRNs) and protein–protein interaction (PPI) networks. Such methods allow discerning relevant regulatory pathways and molecular interactions, and they provide more insights into the mechanisms of diseases. Although they have benefits, network-based approaches can be computationally intensive, and can sometimes depend on high quality prior biological knowledge, which may not always be present.

5.4 Multi-Omics-Specific Frameworks

There are a number of multi-omics data integration frameworks that have been designed. Similarity Network Fusion (SNF) is a popular approach, which combines various types of data building and connecting similarity networks to detect common patterns in the data sets. Multi-Omics Factor Analysis (MOFA) is a latent-variable modeling approach that helps to define sources of common variation across omics layers and provides dimensionality reduction and interpretation. On the same note, integrative clustering models as iCluster and supervised models like DIABLO (Data Integration Analysis for Biomarker discovery using Latent cOmponents) are good identification tools of disease subtypes and biomarkers. These frameworks are directly oriented to solving the issues of multi-omics data such as heterogeneity and high dimensionality. Nevertheless, they frequently need fine-tuning of parameters and might be limited in scalability with large data sets.

Table 1. Comparison of Machine Learning and Integration Approaches in Multi-Omics Analysis

Type	Application	Strengths	Limitations
Traditional ML	Classification, biomarker selection	Interpretable, robust, low computational cost	Limited in capturing complex nonlinear relationships
Deep Learning	Feature extraction, prediction	Handles high-dimensional data, learns complex patterns	Requires large datasets, low interpretability
Graph/Network Models	Biological interaction modeling	Captures network relationships, biologically meaningful	Computationally intensive, depends on prior knowledge
SNF / MOFA / iCluster / DIABLO	Multi-omics data integration	Designed for heterogeneous data, effective fusion	Parameter tuning required, scalability issues

6. Evaluation Metrics for Multi-Omics Models

Multi-omics integration models are tested with the help of a complex of computational and statistical measures that guarantee the correct forecast and the high quality of interpretation. To quantify classification, accuracy, precision, recall (sensitivity) and F1-score are some of the most common metrics. The accuracy is the average percentage of correctly identified samples, and precision and recall measure a false positive and false negative ratio. The accuracy in successful multi-omics models used in disease classification is usually in the range of 85-95 and the F1-scores are high (greater than 0.80), showing that the sensitivity and specificity were in equilibrium. These metrics give a baseline evaluation of predictive performance but might not be able to adequately assess model strength, particularly with unbalanced datasets.

In order to overcome them, more sensitive performance measures like the area under receiver operating characteristic curve (ROC-AUC) and precision recall curve (PR-AUC) are commonly employed. ROC-AUC tests the capability of classifying between classes using the model with all types of threshold settings with a value above 0.90 typically indicating a high level of discrimination. PR-AUC tends to work especially well in cases of class unbalance where one class of interest that is of more interest is the minority e.g. in disease detection. These measures give a more detailed analysis of the model performance and have been deemed to be necessary in multi-omics studies with heterogeneous datasets.

To analyze multi-omics data unsupervised, clustering measures are employed to determine the quality of groupings of integrated datasets. The Silhouette Score, with the range of values between -1 and 1, reflects how effectively the data is classified as belonging to the corresponding clusters with the value above 0.5 corresponding

to the clearly defined clusters. Likewise, the Adjusted Rand Index (ARI) and the Normalized Mutual Information (NMI) measure agreement between predicted clusters and known labels, and the value is usually greater than 0.70 indicating good clustering. In clinical uses, assessment can be carried out to the level of survival analysis, which involves Concordance Index (C-index), as a measure of predictive accuracy of risk models, where the index of above 0.70 indicates solid prognostic power. This analysis is also supported with Kaplan-Meier survival curves by providing a visual comparison of the survival distributions of patient groups.

In addition to the performance of computations, the biological validation is important in evaluating the applicability of the multi-omics models. The analysis of enriching pathways with the help of databases, e.g., Gene Ontology (GO) and KEGG, is a common technique used to find out whether the identified features are important in terms of functionality. The most common measure of statistical significance is a p-value, with a p-value of less than 0.05 usually used to determine significant biological relationships. Even though they can perform well in predicting, biologically unproven models do not necessarily yield practical information. Thus, it is important to note that high computational performance is not always associated with biological relevance and both should be taken into account together to facilitate robust and clinically relevant multi-omics analysis.

7. Applications in Complex Disease Mechanisms

Combining multi-omics data with machine learning has contributed greatly to the elucidation of the intricate processes of disease by facilitating the detection of cross-layer interactions and regulation of molecules. In contrast to the conventional methods that are mainly used in prediction, integrative multi-omics models can afford more information on disease etiology by connecting genetic variants to the functional consequences at the transcriptomic, proteomic, and metabolic scales. These methods enable the identification of important drivers of disease progression, the discovery of unknown biological networks, and enable the formulation of precision medicine strategies in a range of disease contexts.

Multi-omics integration has been especially useful in cancer systems biology, especially in treating and diagnosing tumor heterogeneity, which is a significant challenge in cancer treatment and diagnosis. Machine learning models can offer clear tumor subtypes and provide a contribution to the deep insights of the molecular processes of cancer progression by integrating genomic mutations, transcriptomic profiles, proteomic signal pathways, and epigenetic alterations. Notably, the methods allow finding strong biomarkers through the combination of signals in multiple layers of omics data, as opposed to individual-gene biomarkers. This has resulted in a better appreciation of oncogenic pathways, tumor microenvironment interaction and resistance mechanisms which will eventually aid in the development of targeted therapies and individualized treatment approaches.

Multi-omics combination in the neurodegenerative disease scenario, e.g. Alzheimer and Parkinson disease, is a critical source of insight regarding complex molecular mal-regulation of genes, proteins, and metabolic pathways. The application of machine learning algorithms to matrices of integrated data has demonstrated such important mechanisms as abnormal aggregation of proteins, mitochondrial dysfunction, and neuroinflammatory responses. These methods enable the identification of early-stage biomarkers and discovery of disease-specific pathways, not available by single-omics analysis, by linking transcriptomic changes to proteomic and metabolomic changes. Such integrative approach is crucial in comprehending the disease progression and the presence of therapeutic targets in diseases that have a gradual and multifactorial degeneration.

In the case of metabolic diseases, such as diabetes and obesity, multi-omics techniques can be used to understand how genetic predisposition, metabolic processes, and the environment interact. Slip learning-based combination of genomics, metabolomics, and proteomics data has identified important regulatory pathways in insulin resistance, lipid metabolism, and energy homeostasis. Such analyses give a more detailed insight into the pathophysiology of diseases, by establishing the role of distortions at more than one molecular level in causing a metabolic imbalance. On the same note, the integration of multi-omics in cardiovascular diseases has been applied to study intricate mechanisms, including inflammation, endothelial dysfunction, and lipid metabolism. By integrating datasets of varied type, scientists will be able to determine molecular signatures related to disease progression and discover the pathways related to conditions like atherosclerosis and heart failure.

In general, the combination of multi-omics with machine learning is not only likely to increase the accuracy of prediction, but, more importantly, it allows finding the underlying biological mechanisms that cause complex diseases. Such a transformation of the classical predictive models to the mechanistic interpretation is an essential breakthrough in the sphere of biomedical research, as it allows creating more efficient diagnostic instruments, specific treatment, and an individual approach to treatment.

8. Challenges and Limitations

Although the integration of multi-omics and machine learning has made great progress, a number of challenges remain that restrict their application to a complete explanation of complex disease mechanisms. Heterogeneity and noise of data is one of the main problems since multi-omics data are produced on various experimental platforms, which have varying scales, formats, and variability degrees. The variation in data distribution, methods of measurement, and batch effects create differences, which make integration difficult and can create biased or unreliable results. Also, the lack of data in omics layers only contributes to further difficulties, weakening the integrative analyses. The other important weakness is the lack of balance between small sample sizes and high

dimensionality commonly known as the curse of dimensionality. Multi-omics data generally has a high number of features (thousands to millions) but a small number of samples, posing a greater risk of overfitting and lowering the generalizability of models. Although machine learning methods can be used to a certain extent to counter this problem by feature selection and dimensionality reduction, the scarcity of sufficiently large and well-labeled datasets is a significant obstacle to creating predictive and mechanistic models.

It is also a complex issue that multi-omics research lacks standard pipelines and integration structures. Various studies usually use different preprocessing approaches, normalization options, and integration approaches, which complicate the comparison of results and generation of findings. The absence of standardization restrains the scalability and reproducibility of multi-omics studies, preventing their wider use in clinical practice and biomedical studies. Moreover, the choice of the right integration techniques can be heavily influenced by the particular characteristics of the data, and the field experience and a large amount of parameter optimization may be needed. Another critical challenge is model interpretability, especially as more sophisticated techniques using deep learning and complex networks become more popular. Although these models have high predictive power, they are black-box models and it is hard to know the biological processes behind the outcome. This reduces confidence in model outputs and acts as a barrier to the acceptance of model results in clinical practice, where model explainability is vital.

Lastly, concerns about limited translation of multi-omics findings into clinical practice are also a significant concern. Despite the high predictive accuracy and favourable biomarkers reported by many studies, only small proportion of studies are supported by practical clinical practice. The barriers to the implementation of multi-omics methods in the normal healthcare setting include data standardization, regulatory, cost, and large-scale validation studies. These limitations have to be tackled so that the gap between computational research and practical clinical implementation can be narrowed down and precision medicine can finally be achieved.

9. Future and New Trends and Directions

Multi-omics integration is a fast-changing field that has been propelled by newer technologies in high-throughput and computational intelligence. Among the most promising ones is the impossibility to study genomic, transcriptomic, and epigenomic profiles at the single-cell level with the help of single-cell multi-omics integration. This methodology can give unparalleled clarity in comprehending cellular heterogeneity and dynamic biological functions, especially in complicated illnesses like cancer and neurodegenerative diseases. Single-cell multi-omics can provide more information into the precise mechanisms of disease progression, cellular differentiation, and treatment response that is not available with bulk analysis due to cell-cell variation. The other major development is the space omics technologies, which adds spatial context to molecular profiling. In contrast to conventional omics technologies that are based on the loss of positional information, spatial transcriptomics and proteomics enable investigators to map molecular activity in tissue architecture. This is especially useful in the study of the microenvironment of tumors, tissue structure, and localized pathophysiology. Spatial omics together with machine learning may help uncover complex cellular interactions with the microenvironment, with the latter offering a more detailed picture of disease biology.

The future of multi-omics research is also being influenced by the integration of artificial intelligence (AI)-based frameworks. State-of-the-art machine learning and deep learning models are gaining popularity to process complex, high-dimensional data and reveal latent patterns across omics layers. Specifically, Explainable AI (XAI) is also becoming a focus because it is a response to the interpretability issue of deep learning models. Offering understanding of how models make decisions, XAI will improve transparency and trust in computational forecasting, which is needed to translate into clinical adoption and biological validation. Moreover, new methods like federated learning are involved in overcoming key challenges associated with data security and privacy in healthcare. Federated learning allows a highly scalable multi-omics analysis without patient data sharing among multiple institutions, thus keeping patient data confidential. Moreover, real-time and longitudinal multi-omics analysis is likely to be developed, redefining disease tracking and management of treatment. The proliferation of data collection and analysis over time can enable a researcher to monitor the development of the disease, discover the early biomarkers, and create adaptive therapeutic interventions. On the whole, these new trends point to a change towards more integrative, scalable and interpretable frameworks to multi-omics analysis. The integration of innovative omics technologies with AI-based methodologies will likely fill in the gap between computational findings and clinical uses and eventually offer more specific, personalized, and dynamic disease diagnosis and treatment.

10. CONCLUSION

Integration of multi-omics has become an urgent method to further develop the study of the mechanisms of a complex disease because it allows reaching a systems perspective of the biological processes on multiple layers of molecules. The integration of genomics, transcriptomics, proteomics, metabolomics and epigenomics data sets allows the investigator to reveal complex interactions and regulation networks that cannot be imagined in single-omics analyses. Machine learning is an essential component of this paradigm as it enables the derivation of meaningful patterns of high-dimensional and heterogeneous data, thus supporting biomarker discovery, disease diagnosis, and mechanistic insight. Nevertheless, even with the tremendous advancements, issues in data

heterogeneity, model interpretability, and lack of clinical translation remain the order of the day, and this indicates that more transparent and biologically interpretative computational frameworks are needed. Future opportunities involve creating more integrative, scalable and explainable systems that allow them to bridge the gap between computational models and clinical use in real practice to allow more specific and personalized approaches to disease diagnosis and treatment.

REFERENCES

1. Argelaguet, R., Cuomo, A. S., Stegle, O., & Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. Nature biotechnology, 39(10), 1202-1215.
2. Ballard, J. L., Wang, Z., Li, W., Shen, L., & Long, Q. (2024). Deep learning-based approaches for multi-omics data integration and analysis. BioData Mining, 17(1), 38.
3. Baysoy, A., Bai, Z., Satija, R., & Fan, R. (2023). The technological landscape and applications of single-cell multi-omics. Nature Reviews Molecular Cell Biology, 24(10), 695-713.
4. Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. BMC bioinformatics, 17(Suppl 2), S15.
5. Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. Frontiers in genetics, 8, 84.
6. Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. Nature Reviews Genetics, 19(5), 299-310.
7. Li, Y., Wu, F. X., & Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. Briefings in bioinformatics, 19(2), 325-340.
8. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics, 16(2), 85-97.
9. Sartori, F., Codicè, F., Caranzano, I., Rollo, C., Birolo, G., Fariselli, P., & Pancotti, C. (2025). A comprehensive review of deep learning applications with multi-omics data in cancer research. Genes, 16(6), 648.
10. Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. Bioinformatics and biology insights, 14, 1177932219899051.
11. Vahabi, N., & Michailidis, G. (2022). Unsupervised multi-omics data integration methods: a comprehensive review. Frontiers in genetics, 13, 854752.
12. Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature methods, 11(3), 333-337. This references can suitable for this article