AI-Driven Integration of Multi-Omics Data for Next-Generation Precision Medicine

Sophia Anne Coleman [1*], Ethan Robert Mitchell [2]
[1] Institute for Translational Nanomedicine and Cancer Therapeutics, Harvard Medical School, United States
[2] Department of Biomaterials and Nanocarrier Research, University of California San Diego, United States

* Corresponding Author: **Sophia Anne Coleman**

Article Info

Volume: 01
Issue: 05
September -October 2025
Received: 07-08-2025
Accepted: 10-09-2025
Published: 06-10-2025
Page No: 10-14

Abstract
Precision medicine aims to individualize therapeutic strategies by accounting for the molecular, environmental, and clinical heterogeneity inherent across patient populations. The concurrent maturation of multi-omics technologies—encompassing genomics, transcriptomics, proteomics, metabolomics, epigenomics, and microbiomics—has generated unprecedented volumes of complementary biological information that, when integrated, offer a holistic molecular portrait of disease. However, the complexity, high dimensionality, and heterogeneity of multi-omics datasets render traditional analytical frameworks insufficient. Artificial intelligence (AI), encompassing machine learning (ML), deep learning (DL), graph neural networks (GNNs), and transformer architectures, has emerged as the indispensable computational engine for cross-layer omics integration, latent feature extraction, and clinically actionable insight generation. This review comprehensively examines current AI methodologies for multi-omics data integration, their translational applications in oncology, cardiovascular medicine, and rare disease genomics, and the technical, ethical, and regulatory challenges impeding clinical deployment. We further articulate a forward-looking framework for responsible, equitable AI-omics integration in next-generation precision medicine.

Keywords: Multi-Omics Integration, Artificial Intelligence, Precision Medicine, Deep Learning, Graph Neural Networks, Genomics, Biomarker Discovery

1. Introduction

The concept of precision medicine—delivering the right intervention to the right patient at the right time—has evolved from a theoretical ambition to an achievable clinical paradigm, driven by extraordinary advances in high-throughput biological measurement technologies. Genomic sequencing costs have declined by more than six orders of magnitude since the Human Genome Project, democratizing whole-genome and whole-exome sequencing for both research and clinical contexts. Simultaneously, mass spectrometry, nuclear magnetic resonance spectroscopy, single-cell RNA sequencing, and chromatin accessibility profiling have opened transformative windows into protein expression, metabolic flux, transcriptional heterogeneity, and epigenetic regulation across tissues, cell types, and disease states.

No single omics layer, however, captures the full molecular complexity of a disease process. Genomic variants may predispose to disease without transcriptional activation; metabolic dysregulation may manifest downstream of otherwise silent proteomic perturbations. It is the integrative analysis of multiple omics strata—coupled with longitudinal clinical phenotyping, imaging, and electronic health record (EHR) data—that promises a genuinely comprehensive molecular diagnosis and a principled basis for patient stratification and individualized therapy.

AI methodologies are uniquely positioned to address the analytical demands of multi-omics integration: they can learn non-linear, high-dimensional cross-modal interactions; incorporate prior biological knowledge through graph structures; operate in federated environments that preserve data privacy; and generate probabilistic, interpretable clinical recommendations. This review synthesizes the current landscape of AI-driven multi-omics integration, evaluating both methodological advances and

translational clinical achievements, and identifying the key challenges and governance imperatives for responsible implementation.

2. Multi-Omics Data: Layers, Technologies, and Characteristics

Multi-omics data encompass six principal biological information layers, each probing distinct molecular scales and biological processes (Table 1). Genomics characterizes the static blueprint of hereditary variation—single-nucleotide polymorphisms (SNPs), copy number variants (CNVs), structural rearrangements, and somatic mutations—using whole-genome (WGS) or whole-exome sequencing (WES). Transcriptomics measures dynamic mRNA expression through RNA sequencing (RNA-seq) and, at single-cell resolution (scRNA-seq), reveals cellular heterogeneity that bulk analyses obscure. Proteomics quantifies the expressed protein complement via liquid chromatography-tandem mass spectrometry (LC-MS/MS), bridging the genotype-phenotype gap with functional biological entities.

Metabolomics characterizes the downstream biochemical phenotype through nuclear magnetic resonance (NMR) spectroscopy and high-resolution mass spectrometry (HRMS), producing metabolite profiles reflective of cellular energetics, signaling, and environmental exposures. Epigenomics surveys DNA methylation landscapes, histone modifications, and chromatin accessibility through whole-genome bisulfite sequencing (WGBS) and ATAC-seq, revealing gene regulatory states that modulate expression without altering the DNA sequence. Microbiomics—increasingly recognized as the sixth omics dimension—characterizes host-microbiome interaction networks through 16S rRNA amplicon sequencing and shotgun metagenomics, with profound implications for immune modulation, pharmacogenomics, and metabolic disease.

Table 1: Multi-Omics Data Layers — Technologies, Data Volumes, and Primary AI Methods

Omics Layer	Biological Target	Technology	Data Volume	Primary AI Method
Genomics	DNA variants & SNPs	WGS / WES	~200 GB/sample	GNNs, Random Forest
Transcriptomics	mRNA expression	RNA-seq / scRNA-seq	~10 GB/sample	Transformers, VAE
Proteomics	Protein expression	LC-MS/MS	~5 GB/run	Deep Neural Nets
Metabolomics	Metabolite profiles	NMR / HRMS	~1 GB/sample	Autoencoders, PCA
Epigenomics	DNA methylation	ATAC-seq / WGBS	~50 GB/sample	CNN, LSTM
Microbiomics	Gut microbiota	16S / Metagenomics	~15 GB/sample	Bayesian Networks

The defining analytical challenges of multi-omics data are fourfold: extreme high dimensionality (often exceeding millions of features), systematic batch effects and measurement noise across platforms, missing data across patient cohorts with incomplete omics profiling, and the fundamental biological complexity of cross-layer interactions that span timescales from milliseconds to developmental epochs. These challenges necessitate AI frameworks that can simultaneously perform dimensionality reduction, cross-modal alignment, noise modeling, and graph-structured relational learning within a unified computational architecture.

3. AI Methodologies for Multi-Omics Integration

3.1. Deep Learning Architectures

Deep learning has transformed multi-omics analytics by enabling automated hierarchical feature extraction from high-dimensional biological datasets without requiring domain-specified feature engineering. Convolutional neural networks (CNNs) excel at capturing local spatial dependencies in sequence and imaging data, while recurrent architectures—particularly long short-term memory networks (LSTMs)—model temporal dependencies in longitudinal omics profiles. Transformer models, originally developed for natural language processing, have demonstrated exceptional capacity for learning long-range dependencies across omics sequences; models such as DNABERT and BioFormer apply self-attention mechanisms to genomic sequence data, achieving state-of-the-art performance in regulatory element prediction and variant effect scoring.

Variational autoencoders (VAEs) and generative adversarial networks (GANs) enable unsupervised latent representation learning across omics modalities, projecting heterogeneous data into shared embedding spaces amenable to downstream clustering, classification, and association studies.

Multi-modal fusion architectures—combining early fusion (concatenation at the raw feature level), intermediate fusion (latent space alignment), and late fusion (decision-level ensemble)—have been systematically evaluated across cancer genomics benchmarks, with intermediate fusion generally achieving the most favorable accuracy-interpretability balance.

3.2. Graph Neural Networks and Knowledge Graphs

Biological systems are inherently relational: genes interact through regulatory networks, proteins engage in complex interaction webs (PPI networks), metabolites participate in enzymatic pathways, and cells communicate through ligand-receptor signaling. Graph neural networks (GNNs) represent the natural AI architecture for modeling these relational structures, propagating information across edges weighted by interaction confidence or co-expression correlation. In multi-omics integration, heterogeneous graphs simultaneously encode genomic variants, protein nodes, and metabolite vertices, with edges reflecting cross-layer regulatory relationships derived from curated databases such as STRING, KEGG, and Reactome.

Knowledge graph-augmented GNNs—which incorporate structured biomedical ontologies alongside learned data-driven representations—have demonstrated particularly strong performance in drug-target interaction prediction and rare disease gene prioritization, where labeled training data are scarce and biological prior knowledge is essential to constrain the hypothesis space. Models such as MOGONET and SUPREME achieve cross-omics node classification by jointly optimizing within-layer graph convolution and cross-layer attention weighting, effectively learning which omics modalities carry the highest informational value for each clinical prediction task.

3.3. Federated and Privacy-Preserving Learning

Training robust multi-omics AI models requires large, diverse patient cohorts spanning multiple institutions, ethnic groups, and disease subtypes. However, genomic and clinical data are subject to stringent privacy regulations—including HIPAA, GDPR, and the NIH Genomic Data Sharing Policy—that prohibit direct data centralization across institutional boundaries. Federated learning addresses this constraint by training models locally at each participating site and aggregating only model parameter updates at a central server, preserving raw data sovereignty while achieving near-equivalent model performance to centralized training. Differential privacy mechanisms, which inject calibrated noise into gradient updates, provide formal mathematical privacy guarantees at the cost of modest accuracy trade-offs.

4. Clinical Applications

4.1. Oncology and Cancer Genomics

Cancer is fundamentally a disease of molecular heterogeneity—driven by somatic mutation accumulation, epigenetic reprogramming, and clonal evolution—making it the primary domain for AI-driven multi-omics integration. Integrative analyses of TCGA (The Cancer Genome Atlas) multi-omics data using deep learning have identified novel cancer molecular subtypes with significantly different survival trajectories and therapeutic vulnerabilities compared to histopathology-based classifications alone. In breast cancer, multi-omics AI models integrating somatic copy number alterations, DNA methylation, mRNA expression, and protein phosphorylation have stratified patients into biologically coherent subtypes with superior predictive accuracy for chemotherapy response and distant recurrence.

Liquid biopsy platforms that combine cell-free DNA (cfDNA) methylation profiling with plasma proteomics and microRNA signatures—analyzed through multi-modal deep learning—have demonstrated capacity for multi-cancer early detection with high specificity, offering a paradigm-shifting approach to population-level cancer screening that avoids the morbidity of invasive tissue sampling. In precision oncology, AI-driven drug sensitivity prediction models integrating patient-derived tumor genomics with pharmacogenomic interaction databases enable in silico identification of optimal therapeutic regimens prior to empirical treatment, reducing patient exposure to ineffective therapies and their associated toxicity.

4.2. Cardiovascular and Metabolic Disease

Cardiovascular disease risk prediction has been substantially enhanced by integrating polygenic risk scores (PRS) derived from GWAS genomic data with proteomic biomarker panels and metabolomic lipid signatures. AI models integrating these cross-omics features with EHR-derived clinical variables have achieved area under the curve (AUC) values exceeding 0.90 for 10-year major adverse cardiovascular event (MACE) prediction in prospective validation cohorts—substantially outperforming traditional clinical risk scores such as the Framingham Risk Score and SCORE2. In type 2 diabetes, metabolomic-transcriptomic integration has revealed patient endotypes with differential insulin secretion defects and peripheral resistance profiles, enabling endotype-specific pharmacological targeting beyond the conventional glycemic management paradigm.

4.3. Rare Disease Genomics and Drug Discovery

Multi-omics AI is transforming rare disease research by enabling systematic genotype-phenotype correlation mapping across molecularly heterogeneous patient populations. Knowledge graph-augmented GNNs have demonstrated capacity to prioritize causative genes from whole-exome sequencing data in unsolved rare disease cases, reducing diagnostic odyssey duration from a median of 5–7 years to months in pilot clinical implementations. In drug discovery, AI-driven proteomics-genomics integration identifies novel druggable protein targets with compelling genetic evidence—combining Mendelian randomization frameworks with proteomic quantitative trait loci (pQTL) mapping to establish causal evidence for therapeutic targets prior to costly experimental validation.

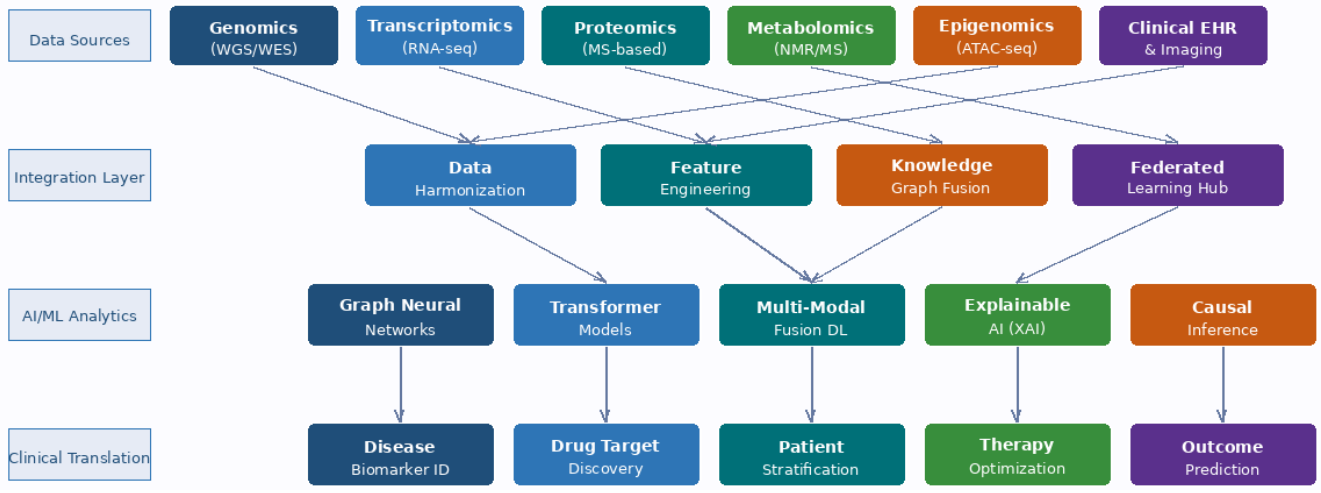

Fig 1: AI-Driven Multi-Omics Integration Pipeline — from raw biological data sources through harmonization, AI/ML analytics, to clinical precision medicine translation.

5. Challenges and Future Directions

5.1. Technical and Computational Challenges

Despite compelling research demonstrations, several technical barriers constrain clinical translation of AI-driven multi-omics integration. Batch effects—systematic non-biological variation introduced by differences in sample collection protocols, sequencing platforms, and processing dates—can profoundly confound cross-cohort analyses if not rigorously corrected through computational harmonization methods such as ComBat, Harmony, or scVI. Missing data

across omics modalities—arising from cost constraints, sample degradation, or assay failure—requires sophisticated imputation strategies that must not introduce spurious biological signals that distort downstream inference.

Model interpretability remains a critical concern: while deep learning architectures achieve high predictive accuracy, the opacity of their decision processes limits clinical acceptance and impedes regulatory approval. Explainable AI (XAI) methodologies—including SHAP (SHapley Additive exPlanations), LIME, and integrated gradients—provide post-hoc attribution of model predictions to specific omics features, improving mechanistic transparency. However, these methods themselves carry assumptions and limitations that must be understood by clinical end users. Table 2 summarizes key AI integration approaches with their clinical applications and performance benchmarks.

Table 2: AI Integration Approaches — Omics Modalities, Clinical Applications, and Performance Metrics

AI Approach	Omics Modalities	Clinical Use Case	Performance	Key Limitation
Graph Neural Networks	Genomics + PPI	Drug target discovery	AUROC: 0.94	Scalability
Transformer Models	Multi-omics + EHR	Cancer subtyping	F1: 0.91	Interpretability
Multi-Modal Fusion DL	Imaging + Omics	Tumor classification	Accuracy: 93%	Data alignment
Federated Learning	All omics layers	Privacy-safe training	Δ AUC < 2%	Communication cost
Variational Autoencoder	Transcriptomics	Latent disease repr.	Silhouette: 0.78	Batch effects
Causal Inference AI	Genomics + Metabol.	Biomarker causality	SHD: 12.3	Confounding

5.2. Ethical, Regulatory, and Equity Considerations

The genomic data embedded within multi-omics datasets carries irreducible patient re-identification risk, raising profound ethical imperatives around informed consent, data stewardship, and secondary use governance. Regulatory frameworks governing AI-based clinical decision support—including the FDA's Software as a Medical Device (SaMD) guidance and the EU AI Act's high-risk AI classification—impose prospective clinical validation requirements that current multi-omics AI platforms have yet to systematically satisfy. Algorithmic bias constitutes a particularly urgent equity concern: multi-omics reference datasets have historically overrepresented individuals of European ancestry, meaning that AI models trained on these corpora may underperform for non-European populations—precisely those with historically unmet clinical need.

Addressing these challenges requires deliberate investment in diverse biobank infrastructure, continuous model performance auditing across demographic subgroups, and inclusive governance frameworks that engage patient communities in research design and data stewardship. The path from algorithmic innovation to equitable clinical implementation is ultimately as much a social and organizational challenge as a computational one.

6. Conclusion

AI-driven multi-omics data integration represents the defining methodological frontier of next-generation precision medicine. By synthesizing complementary molecular information across genomic, transcriptomic, proteomic, metabolomic, epigenomic, and microbiomic layers, AI systems are enabling molecular disease characterization, therapeutic target identification, and patient stratification at a depth and scale previously unattainable. Graph neural networks, transformer models, and federated learning architectures have demonstrated compelling clinical utility across oncology, cardiovascular medicine, and rare disease genomics, with performance metrics consistently surpassing those achievable through single-omics or traditional statistical analyses.

Realizing the full precision medicine promise of multi-omics AI demands coordinated progress across four interconnected dimensions: computational methodology, data infrastructure, regulatory science, and health equity. Methodological advances in interpretable, causally-grounded AI must be matched by investment in diverse, longitudinal biobank cohorts; regulatory validation frameworks must evolve to accommodate the dynamic, adaptive nature of AI-based clinical tools; and equity-centered design principles must ensure that the molecular medicine revolution reaches all patient populations equitably. The convergence of AI and multi-omics science is not merely transforming what we can measure—it is fundamentally redefining what is medically possible.

References

1. Subramanian I, Verma S, Kumar S, *et al*. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051.
2. Reel PS, Reel S, Pearson E, *et al*. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv. 2021;49:107739.
3. Vamathevan J, Clark D, Czodrowski P, *et al*. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–477.
4. Ritchie MD, Holzinger ER, Li R, *et al*. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16(2):85–97.
5. Tran KA, Kondrashova O, Bradley A, *et al*. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. 2021;13(1):152.
6. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet. 2017;8:84.
7. Lappalainen T, MacArthur DG. From variant to function in human disease genetics. Science. 2021;373(6562):1464–1468.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.
9. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
10. Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589.
11. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–332.
12. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46(20):10546–10562.

13. Chen RJ, Lu MY, Wang J, *et al*. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis. IEEE Trans Med Imaging. 2022;41(4):757–770.
14. Raza K. Artificial intelligence against COVID-19: a meta-analysis of current research. Stud Big Data. 2022;80:165–176.
15. Denny JC, Collins FS. Precision medicine in 2030—seven ways to transform healthcare. Cell. 2021;184(6):1415–1419.
16. Nicora G, Vitali F, Dagliati A, *et al*. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. Front Oncol. 2020;10:1030.
17. Zhang L, Lv C, Jin Y, *et al*. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front Genet. 2018;9:477.
18. Miotto R, Wang F, Wang S, *et al*. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236–1246.
19. Wainberg M, Merico D, Delong A, *et al*. Deep learning in biomedicine. Nat Biotechnol. 2018;36(9):829–838.
20. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nat Med. 2020;26(1):29–38.
21. Li Z, Li J, Yu P. scMDC: single cell multi-omics clustering with deep learning. Genome Biol. 2022;23(1):92.
22. Raza K, Singh NK. A tour of unsupervised deep learning for medical image analysis. Curr Med Imaging. 2021;17(9):1059–1077.
23. Wiens J, Saria S, Sendak M, *et al*. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337–1340.

How to Cite This Article
Coleman SA, Mitchell ER. AI-driven integration of multi-omics data for next-generation precision medicine. Int J Biol Biomed Res. 2025;1(5):10-14.