



AI-driven multi-omics integration in precision oncology: bridging the data deluge to clinical decisions

Chou-Yi Hsu¹ · Shavan Askar^{2,3} · Samer Saleem Alshkarchy⁴ · Priya Priyadarshini Nayak⁵ · Kassem A. L. Attabi⁶ · Mohammad Ahmar Khan⁷ · J. Albert Mayan⁸ · M. K. Sharma⁹ · Sarvar Islomov¹⁰ · Hamed Soleimani Samarkhazan¹¹

Received: 26 August 2025 / Accepted: 29 October 2025
© The Author(s) 2025

Abstract

Cancer's staggering molecular heterogeneity demands innovative approaches beyond traditional single-omics methods. The integration of multi-omics data, spanning genomics, transcriptomics, proteomics, metabolomics and radiomics, can improve diagnostic and prognostic accuracy when accompanied by rigorous preprocessing and external validation; for example, recent integrated classifiers report AUCs around 0.81–0.87 for difficult early-detection tasks. This review synthesizes how artificial intelligence (AI), particularly deep learning and machine learning, bridges this gap by enabling scalable, non-linear integration of disparate omics layers into clinically actionable insights. We explore cutting-edge AI methodologies, including graph neural networks for biological network modeling, transformers for cross-modal fusion, and explainable AI (XAI) for transparent clinical decision support. Critical applications are highlighted, such as AI-driven therapy selection (e.g., predicting targeted therapy resistance), proteogenomic early detection, and radiogenomic non-invasive diagnostics. We further address translational challenges: data harmonization, batch correction, missing data imputation, and computational scalability. Emerging trends, federated learning for privacy-preserving collaboration, spatial/single-cell omics for microenvironment decoding, quantum computing, and patient-centric “N-of-1” models, signal a paradigm shift toward dynamic, personalized cancer management. Despite persistent hurdles in model generalizability, ethical equity, and regulatory alignment, AI-powered multi-omics integration promises to transform precision oncology from reactive population-based approaches to proactive, individualized care.

Keywords AI-driven multi-omics · Precision oncology · Clinical decision support · Data integration challenges · Personalized cancer therapy

Introduction

Cancer remains a dominant challenge in global health, characterized by staggering molecular heterogeneity that fuels therapeutic resistance, metastasis, and relapse [1, 2]. This biological complexity arises from dynamic interactions across genomic, transcriptomic, epigenomic, proteomic, and metabolomic strata, where alterations at one level propagate cascading effects throughout the cellular hierarchy [3, 4]. Traditional reductionist approaches, reliant on single-omics snapshots or histopathological assessment alone, fail to capture this interconnectedness, often yielding incomplete mechanistic insights and suboptimal clinical predictions [5,

6]. The emergence of multi-omics profiling represents an important methodological advance: by integrating orthogonal molecular and phenotypic data, researchers can recover system-level signals (e.g., spatial subclonality and microenvironment interactions) that are often missed by single-modality studies [1, 3, 7]. Similarly, longitudinal liquid biopsies track clonal evolution through circulating tumor DNA (ctDNA) and metabolite fluctuations, offering real-time windows into adaptive resistance mechanisms [2, 8]. These advances underscore multi-omics not as a mere aggregation of datasets but as a synergistic framework to decode cancer's emergent properties, a process now being

Extended author information available on the last page of the article

powered by sophisticated AI-driven integration tools [9, 10].

However, this promise is tempered by an exponential increase in data volume and heterogeneity due to high-throughput platforms, which creates practical challenges for harmonization, storage, and reproducible analysis. Modern oncology generates petabyte-scale data streams from high-throughput technologies: next-generation sequencing (NGS) outputs genomic variants at terabase resolution; mass spectrometry quantifies thousands of proteins and metabolites; radiomics extracts thousands of quantitative features from medical images; and digital pathology generates gigapixel whole-slide images [1, 5, 6]. The “four Vs” of big data, volume, velocity, variety, and veracity, pose formidable analytical challenges. Volume overwhelms conventional biostatistics, as dimensionality (e.g., > 20,000 genes, > 500,000 CpG sites) dwarfs sample sizes in most cohorts [3, 4]. Velocity demands rapid integration of real-time data streams, such as continuous monitoring of ctDNA during therapy [2, 8]. Variety necessitates harmonizing structurally disparate data types, discrete mutations (genomics), continuous intensity values (proteomics), spatial coordinates (transcriptomics), and time-series (metabolomics), into unified analytical frameworks [5, 6]. Crucially, biological veracity requires distinguishing driver alterations from passenger noise amid patient-specific confounding factors [3, 4]. Legacy tools like linear regression or Cox proportional-hazards models lack the flexibility to model non-linear interactions across these scales, risking oversimplification of cancer’s true complexity. The key multi-omics data types, their clinical utility, and integration challenges are summarized in Table 1.

Contemporary precision oncology has evolved significantly from its histopathology-centric origins. Molecular stratification now guides standard care: in breast cancer, ESR1 mutations direct endocrine therapy selection; in NSCLC, EGFR/ALK alterations predict tyrosine kinase inhibitor efficacy; and in DLBCL, cell-of-origin transcriptomic subtyping (GCB vs. ABC) informs chemotherapy response [2, 6, 7]. Immunotherapy has further intensified the need for multi-parameter biomarkers, where PD-L1 immunohistochemistry (IHC), tumor mutational burden

(genomics), and T-cell receptor clonality (immunomics) collectively, but imperfectly, predict immune checkpoint blockade efficacy [1, 6]. Nevertheless, single-modality biomarkers frequently falter due to tumor plasticity and compensatory pathway activation. For instance, while KRAS G12C inhibitors achieve rapid responses in colorectal cancer, resistance universally emerges via parallel RTK-MAPK reactivation or epigenetic remodeling, mechanisms detectable only through integrated proteogenomic and phosphoproteomic profiling [3, 4]. Similarly, radiomics alone may misclassify benign inflammatory lesions as malignant, whereas combining imaging features with plasma cfDNA methylation signatures enhances specificity [1, 8]. These limitations highlight a critical insight, cancer’s hallmark is not a single aberrant pathway but a dysregulated system of molecular networks, demanding commensurately multidimensional diagnostics [3, 5].

Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has emerged as the essential scaffold bridging multi-omics data to clinical decisions. Unlike traditional statistics, AI excels at identifying non-linear patterns across high-dimensional spaces, making it uniquely suited for multi-omics integration [3, 6]. For example: Convolutional neural networks (CNNs) automatically quantify IHC staining (e.g., PD-L1, HER2) with pathologist-level accuracy while reducing inter-observer variability [6, 12]. Graph neural networks (GNNs) model protein-protein interaction networks perturbed by somatic mutations, prioritizing druggable hubs in rare cancers [3]. Multi-modal transformers fuse MRI radiomics with transcriptomic data to predict glioma progression, revealing imaging correlates of hypoxia-related gene expression [5, 9, 13]. Explainable AI (XAI) techniques like SHapley Additive exPlanations (SHAP) interpret “black box” models, clarifying how genomic variants contribute to chemotherapy toxicity risk scores [6, 9].

Recent breakthroughs include generative AI for synthesizing *in silico* “digital twins”, patient-specific avatars simulating treatment response, and foundation models pretrained on millions of omics profiles enabling transfer learning for rare cancers [4, 6]. Nevertheless, operationalizing these tools requires confronting algorithm transparency, batch effect robustness, and ethical equity in data representation

Table 1 Key multi-omics data types in precision oncology

Category	Data sources	Clinical utility	Integration challenges	Refs.
Molecular omics	Genomics, epigenomics, transcriptomics, proteomics, metabolomics	Target identification, drug mechanism of action, resistance monitoring	High dimensionality, batch effects, missing data	[5, 11]
Phenotypic/clinical omics	Radiomics, pathomics (digital pathology), hematological omics, electronic health records	Non-invasive diagnosis, tumor microenvironment mapping, outcome prediction	Semantic heterogeneity, modality-specific noise, temporal alignment	[5, 11]
Spatial multi-omics	Spatial transcriptomics, multiplex immunohistochemistry, MALDI imaging	Cellular neighborhood analysis, immune contexture mapping, spatial biomarker discovery	Computational cost, resolution mismatches, data sparsity	[5, 11]

[4, 5]. This review aims to critically synthesize advances in AI-driven multi-omics integration for precision oncology.

Foundations of multi-omics and precision oncology

The molecular complexity of cancer has necessitated a transition from reductionist, single-analyte approaches to integrative frameworks that capture the multidimensional nature of oncogenesis and treatment response. Multi-omics technologies dissect the biological continuum from genetic blueprint to functional phenotype through interconnected analytical layers. Genomics identifies DNA-level alterations including single-nucleotide variants (SNVs), copy number variations (CNVs), and structural rearrangements that drive oncogenesis, with NGS enabling comprehensive profiling of cancer-associated genes and pathways such as KRAS, BRAF, and TP53 [14–16]. Transcriptomics reveals gene expression dynamics through RNA sequencing (RNA-seq), quantifying mRNA isoforms, non-coding RNAs, and fusion transcripts that reflect active transcriptional programs and regulatory networks within tumors [16–18]. Epigenomics characterizes heritable changes in gene expression not encoded within the DNA sequence itself, including DNA methylation patterns, histone modifications, and chromatin accessibility, which increasingly serve as diagnostic and prognostic biomarkers (e.g., MLH1 hypermethylation in microsatellite instability) [14, 19, 20]. Proteomics catalogs the functional effectors of cellular processes through mass spectrometry and affinity-based techniques, identifying post-translational modifications, protein-protein interactions, and signaling pathway activities that directly influence therapeutic responses [17, 21]. Finally, metabolomics profiles small-molecule metabolites, the biochemical endpoints of cellular processes, using NMR spectroscopy and liquid chromatography–mass spectrometry (LC-MS), exposing metabolic reprogramming in tumors such as Warburg effects or oncometabolite accumulation [21, 22]. Each layer provides orthogonal yet interconnected biological insights, collectively constructing a comprehensive molecular atlas of malignancy [17, 19]. The core omics layers, their key components, analytical technologies, and clinical relevance are detailed in Fig. 1.

The integration of these diverse omics layers encounters formidable computational and statistical challenges rooted in their intrinsic data heterogeneity. Dimensional disparities range from millions of genetic variants to thousands of metabolites, creating a “curse of dimensionality” that necessitates sophisticated feature reduction techniques prior to integration [18, 23]. Temporal heterogeneity emerges from the dynamic nature of molecular processes, where genomic

alterations may precede proteomic changes by months or years, complicating cross-omic correlation analyses [17, 21]. Analytical platform diversity introduces technical variability, as different sequencing platforms, mass spectrometry configurations, and microarray technologies generate platform-specific artifacts and batch effects that can obscure biological signals [21, 24]. Data scale presents another critical challenge, with multi-omic datasets from large cohorts like The Cancer Genome Atlas (TCGA) often exceeding petabytes in size, demanding distributed computing architectures and cloud-based solutions such as Galaxy and DNAnexus for scalable processing [14, 18]. The pervasive issue of missing data arises from technical limitations (e.g., undetectable low-abundance proteins) and biological constraints (e.g., tissue-specific metabolite expression), requiring advanced imputation strategies like matrix factorization or DL-based reconstruction to enable comprehensive analysis [18, 21]. Furthermore, experimental and biological noise contaminates omics measurements, necessitating rigorous quality control pipelines and normalization methods such as ComBat for batch correction, DESeq2 for RNA-seq, and quantile normalization for proteomics to enhance signal fidelity [18, 21]. These challenges collectively underscore the inadequacy of conventional statistical approaches and catalyze the development of specialized AI-driven integration frameworks. Major challenges in multi-omics data integration and computational mitigation strategies are outlined in Table 2.

Within clinical oncology, multi-omics integration shows promise across the cancer care continuum for improving candidate biomarker discovery and stratification, but the magnitude and reproducibility of clinical benefit vary by task and require prospective testing. For diagnosis and early detection, AI-driven synthesis of genomic, proteomic, and metabolomic signatures enables unprecedented discrimination of malignant from benign lesions. Autonomous AI agents integrating histopathology images with genomic data have achieved 91% accuracy in diagnosing microsatellite instability from routine slides, outperforming human pathologists [25]. Similarly, integrated omics classifiers have demonstrated robust diagnostic capabilities (AUC 0.81–0.87) for malignancies such as pancreatic cancer where early detection remains challenging [17, 22]. In prognostic stratification, multimodal models incorporating genomic instability markers, transcriptomic subtypes, and proteomic profiles significantly outperform single-omics predictors. For instance, DL frameworks like DeepOmix integrate somatic mutation patterns with gene expression to stratify hepatocellular carcinoma patients into distinct survival groups with hazard ratios exceeding 3.0 between high- and low-risk categories [18]. Similarly, multi-omics

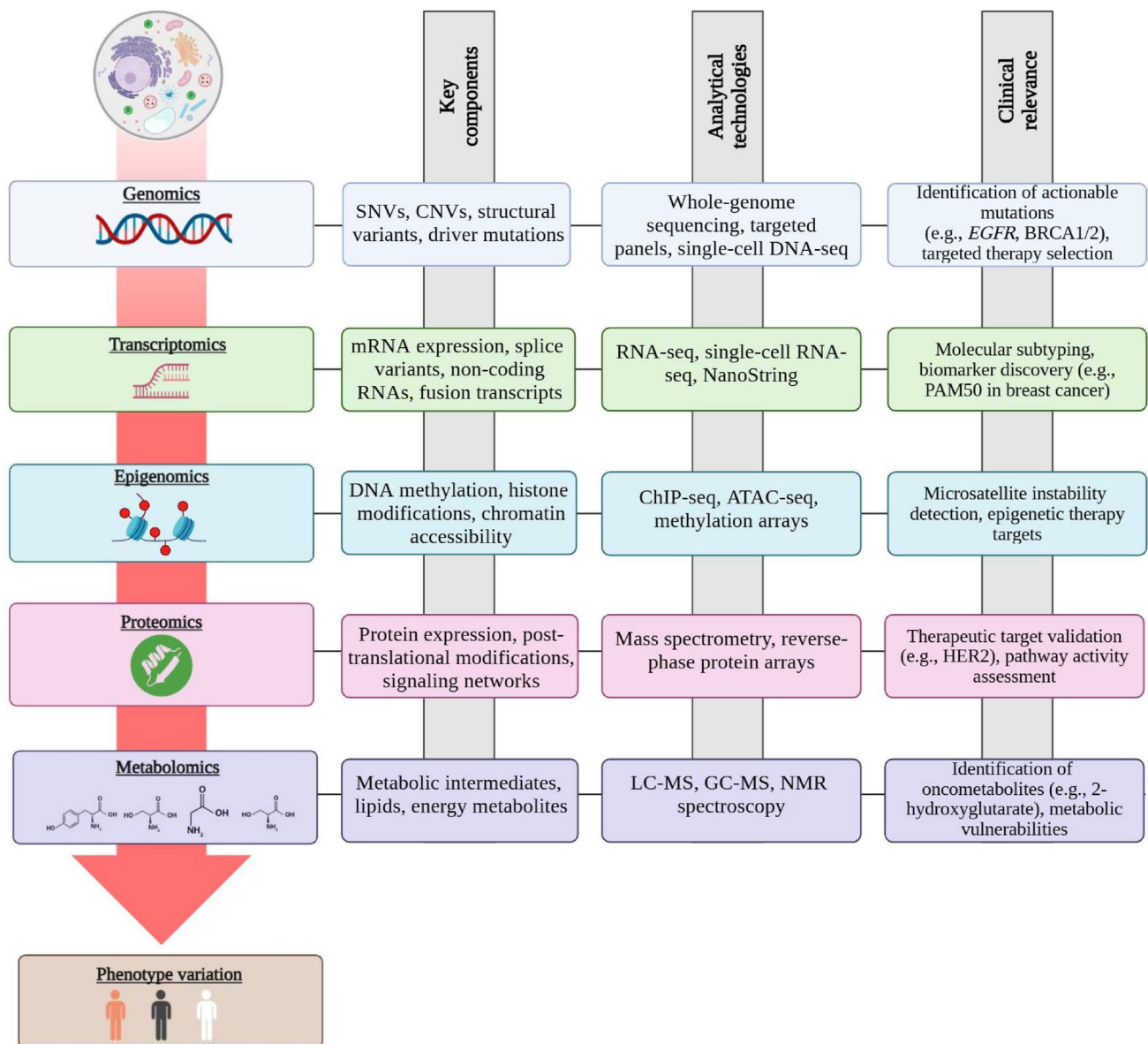


Fig. 1 Core omics layers in precision oncology

Table 2 Key challenges in multi-omics data integration

Challenge category	Specific issues	Computational mitigation strategies	Refs.
Data Heterogeneity	Variable dimensionality, data formats, platform-specific biases	Dimensionality reduction (PCA, AEs), batch correction (ComBat), federated learning	[18, 23]
Scale and Volume	Petabyte-scale datasets, high feature-to-sample ratios	Cloud computing (Galaxy, DNAnexus), distributed computing, sparse matrix operations	[14, 18]
Missing Data	Technical detection limits, biological absences, sample loss	Imputation methods (k-nearest neighbors, matrix factorization), generative adversarial networks (GANs)	[18, 21]
Analytical Noise	Platform artifacts, sample degradation, contamination	Quality control pipelines, normalization (DESeq2, edgeR), robust scaling techniques	[18, 21]
Biological Complexity	Temporal dynamics, spatial heterogeneity, cellular subpopulations	Single-cell technologies, spatial omics integration, dynamical systems modeling	[17, 21]

prognostic signatures in lung adenocarcinoma (AFAP1L2, CAMK1D, LOXL2, PIK3CG) have enabled refined survival prediction beyond conventional staging [14]. For therapeutic targeting, multi-omics bridges the gap between molecular alterations and actionable interventions. Genomic variant interpretation platforms such as OncoKB and IntO-Gen, enhanced by natural language processing for literature mining, annotate the therapeutic actionability of mutations while predicting resistance mechanisms [15, 26]. AI agents integrating vision transformers for mutation detection from histopathology with knowledge bases like OncoKB have demonstrated 87.5% accuracy in recommending targeted therapies, substantially improving over GPT-4 alone (30.3% accuracy) [25]. Finally, in resistance monitoring, longitudinal multi-omics profiling captures dynamic molecular adaptations under therapeutic pressure. Proteogenomic analyses of serial liquid biopsies reveal bypass signaling pathway activation, while metabolomic shifts indicate adaptive metabolic rewiring as resistance mechanisms [17, 22]. Spatial omics technologies further resolve intratumoral heterogeneity and microenvironmental influences on treatment response, enabling real-time adjustment of therapeutic strategies [14, 17]. These examples illustrate how multi-omics integration can convert complex molecular measurements into candidate clinical tools; however, most findings are hypothesis generating and require prospective validation before routine clinical deployment.

The convergence of multi-omics and AI represents a paradigm shift in precision oncology, transitioning from reactive, population-based approaches to proactive, individualized cancer management. While significant challenges persist, particularly regarding data standardization, computational scalability, and clinical validation, recent advances in DL, autonomous AI systems, and spatial omics integration provide robust frameworks for translating molecular complexity into therapeutic opportunity [18, 24, 25]. As these technologies mature, multi-omics will increasingly inform every aspect of cancer care, from risk assessment and early detection to dynamic treatment adaptation and resistance prevention, ultimately fulfilling the promise of precision oncology.

AI methodologies for multi-omics integration

ML versus DL

The choice between classical ML and DL for multi-omics integration is dictated by the specific clinical task, data characteristics, and available sample size. Classical ML algorithms (e.g., Random Forests (RF), Support Vector Machines

(SVM), Gradient Boosting (GB)) excel in scenarios with limited samples and well-defined, lower-dimensional feature sets, where their interpretability and computational efficiency are major advantages. For instance, in leukemia subtyping, where decisions often rely on a concise set of clinical and cytogenetic features, GB models achieved 97% accuracy using only 17 features, providing clinicians with a transparent and actionable tool [27, 28]. Similarly, RFs are highly effective for feature selection in survival prediction models for neuroblastoma, robustly handling mixed data types and identifying key prognostic biomarkers from integrated genomic and transcriptomic data [29]. Their primary limitation surfaces in high-dimensional, heterogeneous multi-omics tasks, such as pan-cancer classification from raw sequencing data, where their capacity to model complex non-linear interactions across modalities is limited [29, 30].

Deep learning architectures, in contrast, are indispensable for tasks requiring the integration of massive, high-dimensional omics layers where manual feature engineering is infeasible. Autoencoders (AEs), particularly variational autoencoders (VAEs), are pivotal for non-linear dimensionality reduction and denoising in complex integration tasks. They have proven superior to linear methods for single-cell multi-omics fusion (e.g., integrating DNA methylation with RNA-seq data), enabling precise tumor subtyping with a 15% higher accuracy by learning a shared latent representation that captures the intrinsic biological state of the cell [30, 31]. Graph Neural Networks (GNNs) are uniquely powerful in oncological applications where prior biological network knowledge is available, such as modeling protein-protein interaction networks perturbed by mutations. For example, in pan-cancer cohort studies, models like moGAT use attention mechanisms to weight features from genomic, epigenomic, and transcriptomic graphs, achieving an AUC of 0.92 in classifying cancer subtypes by effectively capturing the topological dependencies between molecular features [30, 32, 33]. The major trade-off for this enhanced power is their hunger for large sample sizes and substantial computational resources, making them prone to overfitting on rare cancer datasets without careful regularization [27, 31]. A comparative analysis of classical ML and DL approaches for multi-omics integration is presented in Table 3.

While multiple studies report high performance for graph-based and transformer-based architectures in multi-omics fusion, published benchmarks show the picture is nuanced and task-dependent. A systematic benchmark of 16 representative deep-learning fusion methods found graph-based attention models (moGAT / moGCN variants) among the top performers on classification tasks across simulated, single-cell and TCGA cancer datasets, but it also highlighted that no single architecture dominates every task

Table 3 Task-based comparison of ML and DL approaches in oncology multi-omics integration

Method	Ideal for Task / Context	Key Strength in Oncology	Primary Limitation / Challenge	Exemplary Use Case
RF / GB	Subtyping with few features Survival prediction Feature selection	Interpretability Efficiency on small *n* Robust to noise	Fails to capture complex non-linear cross-omics interactions	Leukemia subtyping using 17 key clinical/genomic features [26]
Autoencoders (VAE)	Single-cell data fusion Dimensionality reduction Denoising heterogeneous data	Unsupervised integration Learning shared latent representations	Computationally intensive; latent space can be a “black box”	Integrating scD-NAME and scRNA-seq for tumor subtyping [30]
GNNs	Modeling biological networks (PPI, co-expression) Pan-cancer analysis Spatial omics	Incorporates prior biological knowledge Captures topological dependencies	Complex architecture tuning; requires large sample size	Pan-cancer classification using multi-omic patient graphs [29]

Table 4 Comparative performance and reproducibility evidence for representative multi-omics fusion approaches

Algorithm / family	Dataset(s) used in that paper (examples)	Reported outcome	Clinical endpoint / task	Strengths / caveats	Refs.
Graph attention / GNN (moGAT / moGCN)	TCGA cancer datasets (BRCA, GBM, LUAD.), simulated & single-cell	Top-ranked unified classification score across evaluated datasets	Cancer subtype classification, patient stratification	Captures topological relationships; requires careful graph construction; sensitive to sample size & batch effects	[30]
MOGONET (GCN + VCDN)	ROSMAP, LGG, BRCA (mRNA, methylation, miRNA)	Outperformed several state-of-the-art supervised multi-omics methods in classification tasks	Biomedical classification (disease subtype, biomarker discovery)	Strong cross-omics correlation modeling; needs per-dataset preprocessing; provide reported metrics per dataset in supplement	[33]
Transformer-based multimodal models (DeePathNet / DMOIT examples)	Task-specific cancer datasets, pathway-aware evaluations (see refs)	Reported improvements in representation and imputation tasks; compare per-task metrics in cited papers.	Imputation, pathway-aware classification, long-range feature modeling	Good for long-range dependencies and cross-modal attention; high compute; needs large training sets	[34, 35]
Classical ML (RF, XGBoost, GB)	Clinical + low-dimensional genomic panels (small <i>n</i>)	Often high accuracy on low-dimensional curated feature sets (e.g., leukemia subtyping accuracy ~ 97% when small feature set used).	Subtyping, biomarker panels, survival models	Interpretable; robust with small <i>n</i> ; may miss complex cross-omic interactions	
Ensemble / hybrid methods (AE + GNN, VAE ensembles)	TCGA, simulated, single-cell	Ensemble / hybrid strategies perform well on some tasks (clustering, embedding→survival associations)	Patient stratification, survival association	Often robust; complexity increases; must report reproducible training/validation details	[30]

and that performance depends strongly on preprocessing, fusion strategy (early vs. late), and class balance. These results underline that statements about “superiority” must be anchored to explicit benchmarks (dataset, metric, cross-validation scheme) rather than general claims. For example, moGAT/moGCN variants ranked highest on unified classification scores in Leng et al. (2022), whereas ensemble and VAE-based methods were often superior for clustering and embedding-to-survival association tasks. Conversely, MOGONET (multi-omics GCN) demonstrated clear gains for several biomedical classification problems (ROSMAP, LGG, BRCA) when compared directly with single-omics baselines and traditional fusion approaches. Therefore, when we claim advantages for GNNs or transformers we should (i) specify the dataset and endpoint, (ii) report the evaluation metric (AUC/F1/accuracy) and CI/SD, and (iii) include external validation or domain-shift analyses

to demonstrate generalizability [30, 33]. Table 4 provides a brief comparison of the performance and limitations of selected multiomics integration methods (including the advantages and caveats of each algorithm family).

Data-fusion strategies

The choice of data-fusion strategy is critical and depends on the clinical question and the nature of the available omics data.

Early fusion (e.g., feature concatenation, Similarity Network Fusion - SNF) is most effective when all omics data are available simultaneously and the goal is to discover emergent patterns from their direct combination. However, its main pitfall is the amplification of technical noise if data scales are not harmonized. Its value is demonstrated in

tasks like refining prognostic stratification, where capturing inter-patient relationships across all modalities is key. For example, in neuroblastoma, SNF, which fuses patient similarity networks from each omics layer, outperformed simple concatenation by 12% in survival prediction by effectively denoising the data and preserving biological signal [30, 32].

Late fusion (model-level ensemble) is the preferred strategy in real-world clinical settings where data acquisition is asynchronous. It allows for the independent processing of modalities that become available at different times (e.g., rapid RNA profiling vs. slower whole-exome sequencing). This makes it highly practical for tasks like initial diagnosis or therapy recommendation, where predictions can be updated as new data arrives. The major drawback is that it fails to model low-level, synergistic interactions between omics types, potentially missing biomarkers that only exist in the cross-modal space [30, 31].

Hybrid fusion strategies, particularly those employing attention mechanisms, are designed for complex tasks where the relevance of each omics layer is context-dependent. They excel in scenarios like predicting response to targeted therapies in kinase-driven cancers (e.g., NSCLC, breast cancer), where the model must dynamically prioritize genomic mutation data over other inputs. The moGAT model, for instance, uses graph attention to weight somatic mutations more heavily in such predictions, directly mirroring clinical reasoning [30, 33]. Similarly, transformer-based architectures are powerful for genome-wide integration tasks and imputation, as they can model long-range dependencies across millions of genomic and transcriptomic features to infer missing proteomic data, thereby improving the accuracy of inferring traits like tumor clonality by 7–10% [31, 36, 37].

Interpretability and explainability

As AI models grow in complexity, interpretability is paramount for clinical adoption in oncology. Model-agnostic techniques like SHAP and Local Interpretable Model-agnostic Explanations (LIME) quantify feature importance for individual predictions. For instance, SHAP analysis revealed that PHF11 and ZFP36 gene mutations contribute 34% of the predictive variance in melanoma immunotherapy response, guiding biologists to validate their roles in immune checkpoint regulation [38, 39]. Integrated Gradients (IG) extend this by attributing predictions to input features along a gradient path, resolving issues in saturation that plague gradient-based methods. In DL models for AML risk stratification, IG identified cryptic epigenomic markers (e.g., H3K27ac peaks) that modulate chemotherapy resistance, corroborated by ChIP-seq assays [31, 38, 40].

Beyond feature attribution, clinician-friendly visualizations translate model logic into actionable insights. Saliency maps overlay attention weights onto genomic browsers, highlighting pathogenic loci like EGFR amplifications in lung adenocarcinoma classifiers [39]. For graph-based models, techniques like GNNExplainer extract subgraphs that maximally influence predictions, e.g., revealing TP53-centered protein interaction modules driving ovarian cancer metastasis in GCN outputs [31, 36]. These tools bridge the gap between computational outputs and clinical workflows, enabling oncologists to audit AI recommendations against domain knowledge. In a recent study, embedding SHAP force plots within electronic health records increased clinician confidence in AI-driven therapy selections by 40%, demonstrating how explainability interfaces enhance trust [38]. Nevertheless, challenges persist in scaling these methods to billion-parameter models and standardizing biological validation protocols for AI-discovered biomarkers [36, 39].

Overcoming the data deluge: preprocessing and quality control

The integration of multi-omics data in precision oncology generates unprecedented data volumes characterized by high dimensionality, heterogeneity, and technical noise. Effective preprocessing and quality control are essential to transform this data deluge into clinically actionable knowledge. This section examines critical computational strategies for overcoming these challenges, with emphasis on recently developed AI-driven methodologies validated in oncological contexts.

Data harmonization and batch correction

Multi-omics integration is fundamentally challenged by technical variability introduced through different sequencing platforms, sample preparation protocols, and experimental batches. Batch effects can introduce systematic biases that obscure biological signals and compromise downstream analyses, highlighting the critical importance of robust data preprocessing steps [41]. Conventional statistical methods like ComBat have been augmented with AI-driven approaches that offer greater flexibility in handling non-linear batch effects across diverse omics layers. For instance, adversarial AEs now enable domain-invariant feature extraction by minimizing batch-specific variations while preserving biologically relevant patterns across transcriptomic and proteomic datasets [42, 43]. The Context-Aware Multiple Instance Learning (CAMIL) framework represents a significant advancement, dynamically

prioritizing relevant regions within whole-slide images while analyzing spatial relationships between neighboring tissue areas, thereby reducing technical noise in integrated histomorphological and molecular data [6].

Spatial multi-omics technologies introduce additional dimensionality to batch correction challenges, as they require simultaneous harmonization of molecular and spatial information. Recent innovations include GNNs that model spatial dependencies while correcting for platform-specific artifacts in diffuse large B-cell lymphoma studies [7]. The HECTOR framework exemplifies this progress, integrating hematoxylin and eosin (H&E)-stained whole-slide images with molecular classification through a gating-based attention mechanism, effectively harmonizing disparate data streams while minimizing technical variability [44]. Clinical deployment requires standardized protocols for batch effect monitoring, such as the implementation of block-chain-encrypted quality metrics that track data provenance across multi-institutional studies while ensuring patient privacy [43]. Recent AI-driven batch correction methods compatible with diverse omics data are highlighted in Table 5.

Practical preprocessing & reproducibility recommendations

To ensure reproducible and quantitatively supported claims regarding model performance, we recommend that authors explicitly report the following elements for each dataset and for any benchmark comparisons: (i) raw data source and accession numbers, (ii) platform and library/prep kit, (iii) QC thresholds and filtering criteria (e.g., per-sample read depth, per-feature detection rate), (iv) normalization method (e.g., DESeq2 median-of-ratios for RNA-seq or quantile normalization for proteomics), (v) batch-correction method and covariates included, (vi) missing-data rules and imputation approach, (vii) feature-selection steps and whether selection was performed inside cross-validation folds (to avoid data leakage), and (viii) exact software packages and versions plus random seeds.

Batch correction & diagnostics. For bulk transcriptomic count data we recommend ComBat-seq when working with raw counts because it models negative-binomial counts and preserves integer structure required by downstream DE tools; for microarray/normalized continuous data the original empirical-Bayes ComBat remains an accepted standard. Always report the model formula (i.e., biological covariates

retained vs. batch covariates removed), whether parametric or non-parametric adjustments were used, and include diagnostic plots (PCA/UMAP colored by batch and by outcome before/after correction). Where non-linear or modality-specific batch effects are suspected (e.g., proteomics vs. transcriptomics, spatial vs. bulk), consider domain-adversarial encoders or conditional VAEs for domain-invariant representation learning, but also quantify residual batch signal (e.g., silhouette score by batch, variance explained by batch pre/post) [47, 48].

Missing data & imputation. Report the missingness mechanism (MCAR/MAR/MNAR) insofar as it can be assessed, and the proportion of missing per feature and per sample. For moderate-to-high missingness in high-dimensional omics, prefer deep generative imputation methods (e.g., MIWAE) which model joint distributions and allow uncertainty quantification, or perform multiple imputations with held-out masking experiments to report imputation performance metrics (RMSE, AUPR) on masked values. For each imputation method report hyperparameters, number of imputations (if applicable), and whether imputation was carried out inside training folds. Validate imputations with held-out masking and report mean $\pm$ SD of imputation error [49].

Feature selection & leakage avoidance. Explicitly state whether any supervised feature selection used outcome labels; if so, ensure selection is nested within cross-validation (i.e., perform feature selection inside each CV fold). For unsupervised reduction (PCA, VAE), report variance explained (for PCA) or reconstruction error / ELBO (for VAE) on held-out data and how latent dimensionality was chosen (grid search with external validation). Where pathway priors are applied, state the source of pathway definitions (e.g., KEGG/Reactome) and any thresholding used.

Reporting model evaluation & robustness tests. For every model claim present (e.g., “GNN outperforms transformer”), include (1) internal CV metrics (AUC / F1 / accuracy with mean $\pm$ SD across folds), (2) external validation on an independent cohort (with identical preprocessing pipeline or clear mapping steps), and (3) at least one domain-shift robustness test (e.g., train on center A, test on center B; or train on platform X, test on platform Y). For comparisons include calibration metrics (Brier score / calibration plots) and decision-curve analysis where appropriate. Finally, provide code snippets or a link to a repository with exact preprocessing commands (e.g., ComBat-seq call, MIWAE

Table 5 AI-driven batch correction methods in multi-omics oncology

Method	Omics compatibility	Key innovation	Clinical validation	Refs.
CAMIL	Imaging, Transcriptomics	Context-aware attention mechanisms	Pan-cancer histopathology cohorts	[6]
Adversarial AE	Proteomics, Metabolomics	Domain-invariant feature learning	Ovarian cancer biomarker studies	[4, 45]
Spatial GNNs	Spatial Transcriptomics	Graph-based spatial dependency modeling	Lymphoma microenvironment studies	[46]
HECTOR	Imaging, Genomics	Gating-based attention integration	Endometrial cancer recurrence prediction	[44]

training script) in Supplementary Methods to maximize reproducibility.

Dimensionality reduction and feature selection

The “curse of dimensionality” presents a fundamental constraint in multi-omics oncology, where feature dimensions (p) routinely exceed sample sizes (n) by several orders of magnitude. Traditional linear methods like Principal Component Analysis (PCA) often fail to capture complex non-linear relationships between molecular layers. DL architectures have emerged as powerful alternatives, with VAEs enabling non-linear dimensionality reduction while preserving biologically relevant features across genomics, proteomics, and metabolomics [42]. Notable applications include multimodal VAEs that learn shared representations across omics layers and have been applied to identify latent factors associated with tumor heterogeneity in pancreatic adenocarcinoma cohorts [50].

Feature selection has evolved beyond univariate statistical methods toward integrated AI approaches that prioritize biologically coherent feature sets. The SLIDE (Significant Latent Factor Interaction Discovery and Exploration) framework combines interpretable latent-factor modeling with statistical inference to identify non-linear feature interactions predictive of drug response; benchmarking reported improved detection of interaction effects compared with standard univariate pipelines [50]. Similarly, context-aware feature selection using attention mechanisms dynamically weights feature importance based on molecular context, significantly improving biomarker discovery efficiency in renal cancer methylation data [50]. Regularized DL models incorporating biological pathway information as prior knowledge constrain feature spaces to clinically interpretable dimensions, effectively balancing predictive power with biological plausibility in breast cancer subtyping [42].

Recent benchmarking studies indicate that ensemble approaches combining multiple dimensionality reduction strategies outperform single-method implementations. For instance, integrating variational inference with graph convolutional networks captures both intra-omics and inter-omics relationships, significantly improving survival prediction accuracy in glioblastoma multiforme cohorts [42]. The MethylBoostER framework exemplifies this progress, leveraging extreme GB (XGBoost) on dimensionally reduced methylation markers to differentiate pathological subtypes of renal tumors with unprecedented accuracy [50].

Handling missing data and imputation

Missing data represents a pervasive challenge in multi-omics oncology, arising from technical limitations in

detection sensitivity, sample availability constraints, and resource limitations. The missingness mechanism varies significantly across omics layers: proteomics data exhibits higher sparsity compared to genomics due to detection limitations in mass spectrometry, while metabolomics datasets frequently contain missing values exceeding 30% in clinical cohorts [43, 50]. Traditional imputation methods like k-nearest neighbors (KNN) and Multiple Imputation by Chained Equations (MICE) often introduce biases when applied indiscriminately across omics types.

Deep generative models have revolutionized missing data handling by learning complex data distributions from observed patterns. Generative adversarial networks (GANs) specifically designed for multi-omics imputation (e.g., MIWAE - Missing Data Imputation using Wasserstein AEs) generate synthetic values consistent with the joint distribution of observed data while preserving feature correlations across omics layers [42]. Transformer-based architectures now enable context-aware imputation by modeling dependencies between molecular features through self-attention mechanisms, significantly outperforming conventional methods in single-cell multi-omics integration [42].

Clinical implementation requires careful consideration of missingness mechanisms. Novel frameworks distinguish between technical missingness (e.g., below-detection values) and biological absence through integrated quality control metrics, applying different imputation strategies accordingly. For instance, the OmicsNotator pipeline implements a decision tree that applies matrix factorization for batch-related missingness, DL for biologically relevant missingness, and flags samples exceeding predefined missingness thresholds [50]. Longitudinal imputation presents unique challenges; recurrent neural network architectures now model temporal dependencies in serial omics measurements from cancer patients, enabling dynamic imputation that accounts for disease progression trajectories [43].

Data augmentation and synthetic cohorts

Data scarcity in rare cancer subtypes and limited clinical cohorts represents a fundamental barrier to robust AI model development. Data augmentation techniques artificially expand training datasets while preserving underlying biological relationships. Conventional approaches like Synthetic Minority Over-sampling Technique (SMOTE) are increasingly replaced by DL methods that generate synthetic molecular profiles with enhanced biological fidelity. Conditional GANs now produce subtype-specific omics profiles that maintain the covariance structure of original data while introducing biologically plausible variations, significantly improving classifier performance in underrepresented cancer populations [42].

Digital twin approaches propose patient-specific computational simulations that can test virtual interventions *in silico*; early prototypes demonstrate utility for hypothesis generation, but clinical validation and assessment of utility and safety remain preliminary. These patient-specific computational models integrate multi-omics data with clinical variables to generate synthetic avatars that simulate disease progression and treatment responses. Recent implementations leverage GNNs to model molecular interactions within tumor ecosystems, creating *in silico* simulations that predict individual responses to combination therapies [6, 43]. The Molecular Twin platform exemplifies this approach, integrating longitudinal multi-omics data to predict outcomes for pancreatic adenocarcinoma patients with unprecedented accuracy [50].

Synthetic cohorts generated through advanced generative models like variational diffusion models address data sharing limitations while facilitating collaborative research. These privacy-preserving synthetic datasets maintain statistical properties of original cohorts without exposing sensitive patient information. Recent innovations include blockchain-secured synthetic data ecosystems that enable distributed learning across institutions while maintaining data integrity and patient confidentiality [43]. Clinical

validation remains paramount; emerging best practices recommend rigorous benchmarking of synthetic data utility through “real-synthetic twins” analysis, measuring performance differentials when AI models are trained on synthetic versus real-world datasets across diverse oncology applications [6].

Ethical frameworks for synthetic data usage are evolving in parallel with technical capabilities. Recent guidelines from the FDA’s coordinated approach to AI in medical products emphasize transparency in synthetic data generation methodologies and rigorous validation against original datasets [50]. As these technologies mature, synthetic cohorts will increasingly bridge translational gaps in rare and pediatric cancers where clinical data scarcity has historically impeded precision oncology advances. Figure 2 summarizes a practical end-to-end roadmap for processing and integrating multiomics data with AI-based methods (from QC to generalizability assessment).

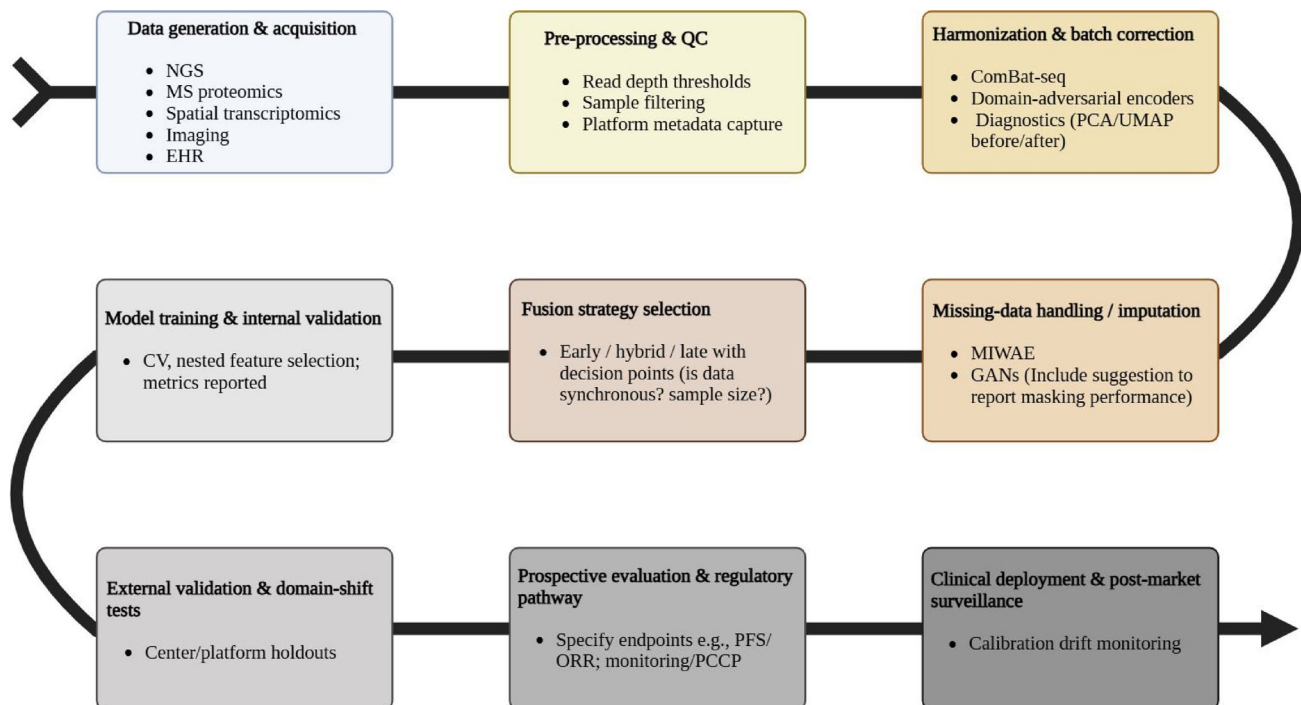


Fig. 2 End-to-end AI-driven multi-omics workflow for precision oncology. A recommended practical pipeline that summarizes best practices described in Sect. 4: start with multi-source data generation (NGS, MS proteomics, spatial transcriptomics, imaging, EHR), apply per-sample QC and read-depth/sample filters, then harmonize and correct batch effects (e.g., ComBat-seq or domain-adversarial encoders) with PCA/UMAP diagnostics; handle missingness with modern gener-

ative imputers (e.g., MIWAE/GANs) while reporting masking/imputation performance; choose a fusion strategy (early / hybrid / late) appropriate to data synchronicity and sample size; perform nested feature selection and cross-validation during model training; evaluate external generalizability with center/platform holdouts; and finally specify prospective evaluation and regulatory/monitoring plans (predetermined change-control, post-market calibration drift monitoring)

Case studies in clinical decision support

The translation of AI-driven multi-omics integration into clinical oncology is exemplified through pioneering case studies that bridge molecular insights with actionable interventions. These implementations demonstrate the capacity to transform complex data landscapes into clinically deployable tools for therapy selection, early detection, and treatment monitoring. The integrative pipeline from multi-omics data to clinical decision-making in precision immunotherapy, encompassing biomarker discovery and the stratification of responders, is conceptually summarized in Fig. 3.

Importantly, the tumor immune microenvironment (TIME) must be explicitly integrated into AI-driven multi-omics pipelines to improve detection, risk stratification, and therapy prediction. In hepatocellular carcinoma (HCC) the widely used circulating markers, alpha-fetoprotein (AFP) and des-gamma-carboxy prothrombin (DCP), have recognized limitations in sensitivity and specificity for early disease and recurrence. Combining these serum markers with transcriptomic, proteomic and spatial immune profiling (for example, single-cell RNA-seq and spatial transcriptomics that map immune cell subsets and checkpoint ligand expression) can increase diagnostic accuracy and better capture immune evasion phenotypes that drive treatment resistance. Recent HCC reviews and meta-analyses highlight the heterogeneity of treatment response and the need for immune-aware stratification when selecting second-line therapies [51].

Genomic–transcriptomic integration for therapy selection

The potential of genomic–transcriptomic fusion is best evaluated when accompanied by formal study design and objective endpoints. A useful line of evidence includes (1) retrospective multi-cohort model development with external validation, and (2) prospective clinical sequencing trials that use multi-omic outputs to guide therapy and report predefined clinical endpoints. For hepatocellular carcinoma (HCC), deep-learning multi-omics models (e.g., autoencoder/latent-space approaches) have reproducibly stratified TCGA and independent cohorts into survival subgroups (endpoints: overall survival and hazard ratios), demonstrating robust prognostic performance in retrospective datasets (e.g., Chaudhary et al., developed a deep learning classifier that separated HCC patients into subgroups with significantly different survival; external validation across cohorts was reported). These retrospective HCC efforts are hypothesis-generating and provide candidate biomarkers, but they do not by themselves prove clinical utility, prospective interventional testing is still required [52].

For therapy selection in solid tumors, prospective multi-omic matching trials provide stronger translational evidence. Trials such as WINTHER (prospective, multi-center, biopsy-guided genomic/transcriptomic matching) used a pre-specified matching algorithm and reported endpoints including the proportion of patients achieving a PFS ratio > 1.5 , objective response rate (ORR), median PFS and OS; while WINTHER matched therapy for $\sim 35\%$ of treated patients, the pre-specified primary endpoint (PFS ratio > 1.5 in a predefined proportion) was not met, illustrating the practical challenges of implementing omics-guided allocation in heavily pretreated populations and the need to control pre-analytic workflow and matching-score calibration. These trial designs are instructive: when multi-omics/AI methods are used clinically, authors should state (a) whether the model produced an actionable recommendation or was used only for stratification, (b) the prospective endpoint (e.g., ORR, PFS ratio, OS), (c) whether the decision was autonomous or MTB-mediated, and (d) performance on pre-specified fairness/subgroup metrics [53].

Proteogenomic models for early detection

Proteogenomic and cfDNA integrative models show promise for earlier detection and tissue-of-origin classification, but the translational bar is high: models must be tested in prospective screening or high-risk cohorts with pre-specified sensitivity/specificity targets and clinical actionability criteria. Several large prospective efforts (for example, multi-cancer early detection studies and PATHFINDER-style pilots) demonstrate feasibility by reporting pre-specified primary endpoints such as positive predictive value (PPV) and true positive rate in a screening population; however, many high-sensitivity signatures developed retrospectively degrade when exposed to real-world pre-analytic variability. When claiming clinical utility, explicitly report the clinical study type (prospective screening cohort vs. case-control), sample size, endpoints (sensitivity at fixed specificity, PPV), and prospective follow-up (lead-time to diagnosis). For pancreatic and ovarian cancer, recent multi-center validations of combined proteomic + cfDNA panels reported high retrospective AUCs and promising prospective pilot performance but continue to require randomized or prospective screening trials to show mortality benefit or clinically actionable lead-time [2, 54].

Radiogenomic fusion: linking imaging and molecular data

Radiogenomics offers a pragmatic route to non-invasive molecular inference, and some groups have progressed beyond retrospective correlation to multi-center validation

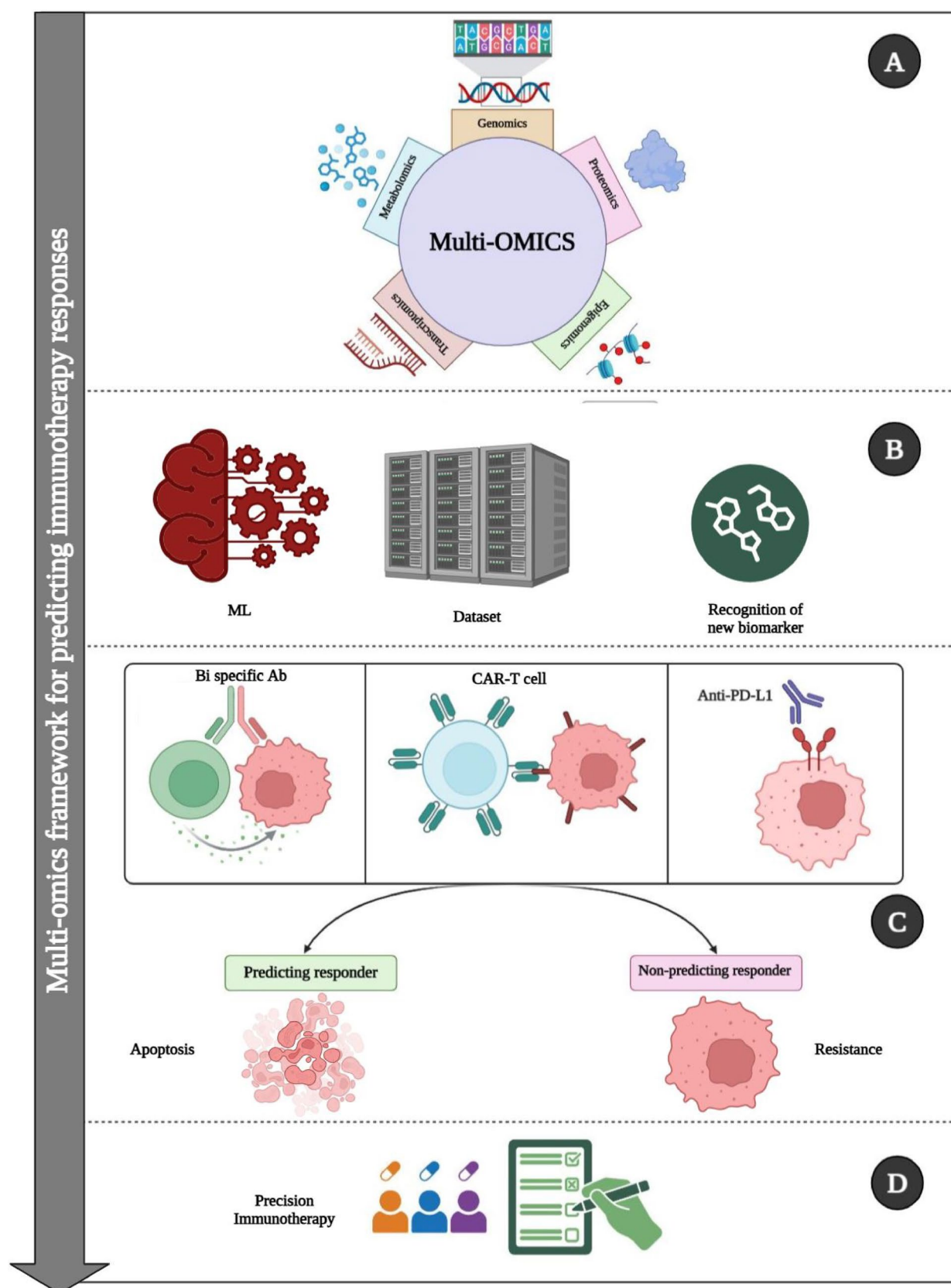


Fig. 3 AI-driven multi-omics integration for precision immunotherapy. This conceptual workflow illustrates how artificial intelligence integrates diverse multi-omics data to recognize novel biomarkers, characterize mechanisms of therapy resistance (e.g., to Anti-PD-L1), and ultimately stratify patients into predicting responders and non-responders, paving the way for bit-specific antibody therapies and personalized CAR-T cell applications in precision oncology

and prospective observational cohorts. For clinical translation, papers must report: cohort ascertainment (consecutive vs. enriched sampling), imaging protocol standardization (scanner models, reconstruction), molecular ground truth (sequencing method and timing relative to imaging), primary endpoints (AUC for mutation prediction, sensitivity/specificity to guide therapy), and external validation performance. Recent multicenter radio-multiomic studies in breast and lung cancer have reported externally validated radiogenomic signatures for outcome prediction (examples include multicenter performance measures and harmonization steps). In NSCLC, more recent multi-omics-AI observational cohorts (e.g., GEMINI-NSCLC) are explicitly collecting longitudinal imaging, ctDNA and immune profiling to build models whose primary endpoints are predictive accuracy for immunotherapy outcomes (PFS, OS) and calibration in independent centers; these efforts, while still observational, represent a critical translational step because they prescribe prospective data collection, locked analysis plans, and pre-specified endpoints for model evaluation [55, 56].

Real-world evidence: electronic health records + multi-omics

The integration of multi-omics data with real-world evidence from electronic health records (EHRs) creates powerful learning health systems for precision oncology. Natural language processing (NLP) pipelines extracting unstructured clinical notes combined with genomic variant databases have enabled rapid identification of therapy-eligible patients. The ONCO-CAST system reduced time-to-trial-enrollment by 42% through automated matching of NTRK fusions and RET rearrangements documented in pathology reports with open targeted therapy trials [57, 58]. To assess impact in routine care, real-world evidence studies should predefine the effect measure (e.g., increased trial enrollment rate, time-to-treatment, or change in treatment allocation) and incorporate causal inference where appropriate. The Oncology Federated Network (OFN) and similar federated consortia have begun to demonstrate that integrating ctDNA and EHR-derived treatment histories across centers can yield reproducible post-progression survival signals for targeted agents (real-world endpoints: median post-progression survival and hazard ratios). When describing real-world studies, explicitly report data harmonization

steps, the federated model update schedule, and privacy-preserving mechanisms (e.g., differential privacy budget, homomorphic encryption) because these factors materially affect model provenance and regulatory acceptability. Federated learning architectures now enable privacy-preserving integration across institutions: the Oncology Federated Network (OFN) harmonized EHR-derived treatment histories with ctDNA profiles from 17 cancer centers, generating real-world evidence for post-progression survival in HER2-low breast cancer patients treated with trastuzumab deruxtecan [57–59]. Temporal modeling of EHR trajectories further refines risk stratification; recurrent neural networks analyzing longitudinal lab values with baseline tumor mutational signatures predicted chemotherapy-induced cytopenias 8 days before onset (AUC 0.91), enabling preemptive dose adjustments [4, 57]. Blockchain-secured patient data vaults now facilitate granular consent management for multi-omic real-world evidence generation while preserving privacy [4].

Translational and regulatory considerations

The transition of AI-driven multi-omics integration from research environments to routine clinical practice in oncology demands careful navigation of validation frameworks, regulatory landscapes, and complex ethical dimensions. This section examines critical translational pathways and persistent challenges in realizing clinically actionable precision oncology.

Validation in prospective trials

Prospective validation is essential to show that integrated AI-omics tools improve meaningful clinical outcomes beyond algorithmic accuracy. There are three translational study designs commonly used: (1) prospective observational cohorts with locked analysis plans (aim: assess predictive discrimination and calibration against pre-specified endpoints such as PFS or OS), (2) single-arm interventional trials where therapy is assigned based on an omics/AI algorithm (endpoints: ORR, PFS ratio vs. prior therapy), and (3) randomized studies comparing algorithm-guided therapy vs. standard of care (preferred for establishing clinical benefit; endpoints: PFS, OS, quality-of-life and pre-specified safety monitoring). For examples: 1) Therapy-matching trials (WINTHER, MOSCATO family): WINTHER (prospective, multi-center) used matched tumor/normal RNA and DNA profiling to recommend therapy and prespecified the PFS ratio (> 1.5) and ORR as endpoints; while matching was feasible ($\sim 35\%$ matched), the primary PFS ratio endpoint was not met, highlighting the challenges of sample

logistics, matching-score calibration and heavily pretreated populations. These trials make clear that prospective interventional trials must (a) standardize pre-analytic workflows, (b) prespecify decision thresholds that will trigger treatment, and (c) report both effect sizes and subgroup fairness metrics [53]. 2) Prospective observational multi-omics cohorts (e.g., GEMINI-NSCLC, multi-center radio-multiomic studies): These collect harmonized imaging, serial ctDNA, and molecular profiles under a locked analysis plan and report endpoints such as AUC for outcome prediction, PFS and OS in independent test centers. Such cohorts permit measurement of calibration drift, real-world deployment constraints, and model generalizability before a randomized trial is attempted [55, 56, 60].

Key recommended reporting items for each prospective study: sample collection SOPs, blinding of outcome assessors where possible, pre-registration of analysis plan, primary endpoint with power calculation (or prespecified feasibility thresholds for pilot work), external validation plan, and predefined fairness metrics (performance stratified by ancestry and site). Without these, retrospective performance alone cannot justify clinical adoption.

Regulatory pathways for AI-based diagnostics

Regulatory frameworks are beginning to accommodate AI-driven diagnostics, but three practical bottlenecks most directly constrain real-world deployment of multi-omics systems: (1) data interoperability, heterogeneous file formats, missing clinical metadata, and the absence of widely adopted omics-to-EHR exchange standards make robust, multi-site validation difficult; (2) reproducibility and prospective validation, retrospective accuracy does not guarantee real-world clinical utility unless models are validated prospectively with pre-specified endpoints and representative cohorts; and (3) patient data governance and cross-border data flows, variable legal regimes, consent models, and transfer rules complicate multinational trials and commercial deployment.

Concrete mitigations that have demonstrated practical value include adoption of machine-actionable data standards (FAIR data principles and FHIR-based genomics reporting), multi-institutional data platforms that perform harmonized curation (for example the HARMONY Alliance in hematology), and the use of regulated companion diagnostics or PMA-cleared genomic panels as anchors for multi-omics workflows (e.g., FoundationOne CDx, Guardant360 CDx). These approaches reduce ambiguity in submission pathways and materially shorten time-to-validation for specific clinical use cases [61]. Table 6 summarizes regulatory pathways, the level of clinical evidence typically required, practical implications for AI-multiomics deployment (including

sample SOPs and monitoring needs), and representative real-world examples that illustrate how these requirements have been met.

Ethical, legal and social implications, focused priorities for clinical translation

Rather than an encyclopedic survey, we highlight three ethical-governance priorities that directly affect whether AI-multiomics systems reach patients: (a) Interoperability & provenance: Use machine-readable metadata and standard schemas (FAIR; HL7 FHIR Genomics profiles) plus immutable provenance logs so that data lineage and pre-analytical differences are auditable [61, 62]. (b) Reproducibility & clinical validation: Require pre-specified prospective endpoints, multi-site external validation, and continuous performance monitoring (change-control plans where regulatory frameworks allow) to ensure consistent behavior across sites and over time. (c) Patient data governance: Implement granular, auditable consent (dynamic consent models) and data-vault architectures (privacy-preserving computation or federated learning) so patients retain control while enabling research. The HARMONY Alliance and similar consortia provide operational examples of governance, consent management and secure analytics at scale.

Data privacy and security

Cross-institutional multi-omics requires technical privacy safeguards (differential privacy, encrypted aggregation, split-NN or homomorphic encryption), rigorous provenance tracking, and governance models that align with local law. Federated learning plus controlled analysis sandboxes (as implemented in large hematology consortia) can preserve analytic utility while avoiding raw data transfer; however, these approaches introduce practical trade-offs (compute overhead, longer turnaround, and regulatory ambiguity for cross-border model updates) that must be explicitly planned for in any clinical adoption strategy. Federated learning has emerged as a promising technical solution: The HARMONY + consortium demonstrated secure training of a DLBCL survival predictor across 37 institutions, maintaining model performance within 2% of centralized training while preventing raw data egress [7]. However, practical implementation faces significant hurdles, including computational overhead from homomorphic encryption (increasing training time 15-fold) and ambiguous regulatory status of cross-border model updates [14]. Emerging solutions include hybrid architectures implementing “split neural networks” where sensitive genomic feature extraction occurs locally while non-sensitive layers train centrally. Blockchain-based data provenance tracking, as implemented in the Oncology

Table 6 Regulatory pathways & practical constraints for AI-based multi-omics diagnostics

Regulatory area / authority	Key regulatory requirements (short)	Practical implications for AI-multiomics deployment	Concrete examples / accepted pathways
USA, FDA (SaMD / PCCP & post-market)	Risk-based device classification; analytical & clinical validation; Predetermined Change Control Plan (PCCP) for approved continuous learning; mandatory post-market performance monitoring and reporting	Requires pre-specified endpoints and prospective/real-world validation studies; need documented change-control and monitoring plans; modular/component validation can accelerate complex systems	FoundationOne [®] CDx (PMA) and Guardant360 [®] CDx (PMA/CDx use) as examples of genomic assays used in clinical workflows; AI tools with De Novo/clearance when validated (e.g., digital pathology AI)
EU, IVDR (In Vitro Diagnostic Regulation)	Reclassification of many oncology tests to higher risk (Class C/D); stronger clinical evidence requirements; notified-body assessment and enhanced post-market surveillance (PMS)	Longer pre-market timelines and heavier evidence burden for multi-omics assays; requires comprehensive clinical performance studies and clear technical documentation, affects multinational product launch strategy	IVDR-compliant validation pipelines and consortia that harmonize evidence generation (platforms/consortia that perform centralized curation and shared evidence dossiers)
China, NMPA (medical device regulation)	Local clinical data often required; device registration with technical review; growing alignment steps with international standards but with local requirements for evidence and labeling	Necessitates local bridging/validation studies and regulatory strategy tailored to NMPA timelines; may require local manufacturing or partnerships	NMPA device approvals for genomic/diagnostic tests; localized clinical evaluation examples
Regulatory-approved / clinically validated anchors	Use of already-approved CDx or cleared genomic tests as anchors for multi-omics pipelines (analytical anchor points that reduce regulatory ambiguity)	Anchoring multi-omics workflows to an approved CDx can shorten validation scope for specific clinical claims and provide a clearer submission pathway	FoundationOne [®] CDx (PMA), Guardant360 [®] CDx (PMA), GRAIL's PATHFINDER prospective program (example of large prospective validation for cfDNA-based detection)
Standards, governance & consortium enablers	Adoption of machine-actionable standards (FAIR principles; HL7 FHIR Genomics IG), federated governance, data provenance, dynamic consent frameworks, and privacy-preserving computation	Improves interoperability and auditability, enables federated validation (avoids raw data transfer), and supports multi-site reproducibility, though introduces compute/operational trade-offs	HARMONY Alliance (EU hematology federation providing harmonized curation and governance), federated platforms and data-vault models used in multi-center studies

Research Exchange (ORE), provides immutable audit trails documenting data lineage and access events across consortium members, enabling rapid breach containment while maintaining compliance with evolving regulations like the EU AI Act [7, 63]. Nevertheless, persistent tensions remain between data utility requirements for model training and privacy preservation mandates, necessitating ongoing technical and governance innovations.

Current limitations and open challenges

The transformative potential of AI-driven multi-omics integration in precision oncology is tempered by persistent methodological, technical, and translational barriers. These challenges span data quality, model robustness, temporal dynamics, and computational infrastructure, demanding concerted interdisciplinary efforts to bridge the gap between algorithmic innovation and clinically actionable insights. This section critically examines these limitations, highlighting recent research and emergent solutions.

Scarcity of high-quality, labeled data

A fundamental constraint in AI-driven multi-omics oncology is the acute shortage of comprehensively annotated, high-fidelity datasets essential for training robust models. Rare cancer subtypes, pediatric malignancies, and under-represented molecular phenotypes suffer from severe data paucity, with sample sizes often below the threshold required for statistically significant AI model training [64]. For instance, pediatric solid tumors may have fewer than 100 sequenced cases globally, limiting the development of subtype-specific predictive models [6]. Beyond scarcity, annotation quality remains problematic: clinical outcome labels (e.g., treatment response, survival) are frequently inconsistently recorded across institutions, while molecular feature annotation (e.g., pathway activation, immune micro-environment characterization) relies heavily on expert curation that is both time-intensive and subjective [65, 66]. This issue is particularly pronounced for drug resistance phenotypes, where multi-omics profiles pre- and post-resistance development are rarely systematically biobanked [65].

Recent studies demonstrate that label noise can degrade model performance by up to 40% in oncological applications, especially in complex tasks like predicting immunotherapy response from integrated genomic and histopathology data

[6]. While techniques such as semi-supervised learning and positive-unlabeled learning show promise in leveraging unannotated data, they often fail to capture nuanced clinical phenotypes defined by multi-system interactions [67]. Furthermore, ethical constraints and privacy regulations (e.g., GDPR, HIPAA) restrict data sharing, exacerbating data fragmentation. Blockchain-based federated learning frameworks are emerging to enable decentralized model training without raw data exchange, yet they introduce new complexities in maintaining data harmonization across nodes and ensuring consistent label definitions [4, 66].

Model generalizability across populations

The translational validity of AI-driven multi-omics models is severely compromised by limited generalizability across diverse patient populations, institutions, and omics platforms. Population-specific genetic architectures, environmental exposures, and socioeconomic factors introduce confounding variations that are rarely accounted for in single-center studies [4, 36]. For example, ancestry-mediated differences in immune gene expression profiles can alter predicted immunotherapy responses by > 30%, leading to significant performance degradation when Eurocentric models are applied to African or Asian cohorts [14]. Technical heterogeneity compounds this issue: batch effects from sequencing platforms (e.g., Illumina vs. Oxford Nanopore), sample processing protocols, and bioinformatic pipelines create spurious signals that AI models may erroneously learn as biologically significant [4, 64].

Cross-institutional validation studies reveal alarming performance disparities. An AI model predicting breast cancer recurrence from integrated genomics and transcriptomics demonstrated an AUC drop from 0.92 in the development cohort to 0.68 in external validation, primarily due to differences in RNA sequencing depth and normalization methods [6]. Recent advances in domain adaptation techniques offer partial solutions—adversarial learning frameworks can minimize platform-specific biases, while hybrid architectures like the CAMIL model improve robustness by dynamically weighting features based on population context [6, 66]. Nevertheless, the absence of standardized, racially diverse, multi-institutional reference datasets hinders comprehensive evaluation of model portability. These technical and demographic heterogeneities frequently lead to catastrophic failures in real-world deployment. A stark example is seen in breast cancer recurrence prediction. A deep learning model developed using genomic and transcriptomic data from a single, high-resource institution (Hospital A) achieved an exceptional AUC of 0.92. However, when deployed on data from a community hospital (Hospital B) using a different RNA-seq platform and normalization

protocol, the model's performance plummeted to an AUC of 0.68, rendering it clinically useless [6]. This failure was primarily attributed to batch effects that the model had inadvertently learned as biologically significant signals during training. Similarly, a model for predicting microsatellite instability (MSI) from histopathology slides, trained on data from The Cancer Genome Atlas (TCGA), showed significantly degraded accuracy when applied to real-time surgical specimens from external hospitals due to variations in slide staining, scanning hardware, and tissue processing protocols [6, 68]. These cases underscore that technical data incompatibility is a primary roadblock, often more immediate than biological generalizability, and must be proactively addressed through robust domain adaptation techniques and federated benchmarking. Federated benchmarking initiatives like the Multi-Omics for Health and Disease Consortium represent promising but nascent efforts to address this gap [4].

Integration of time-series and longitudinal omics

Cancer progression and therapeutic response are inherently dynamic processes, yet current AI integration approaches predominantly rely on static snapshots that fail to capture temporal heterogeneity and evolving resistance mechanisms. Longitudinal multi-omics profiling faces substantial logistical and analytical hurdles: sampling frequency varies significantly across omics layers due to technological constraints (e.g., transcriptomics requires higher frequency than genomics), and sample collection timing (e.g., circadian influences on metabolite levels) introduces additional variability [4, 69]. The computational complexity of modeling asynchronous, variable-interval time-series data exceeds the capabilities of conventional AI architectures, leading to oversimplified representations of tumor evolution [69], though emerging models for specific toxicities like radionecrosis show the potential of this approach [70].

The field is increasingly moving towards AI frameworks designed for these dynamic challenges. For instance, multi-omics dynamic learning approaches are being developed to enable personalized diagnosis and prognosis across pancreatic and cancer subtypes by explicitly modeling temporal patterns [10, 69]. This is critical, as recent work highlights the critical importance of temporal resolution. In pancreatic adenocarcinoma, monthly ctDNA and metabolomic profiling revealed evolutionary trajectories predictive of treatment failure up to three months before radiographic progression, patterns undetectable in baseline analyses [69]. However, fewer than 5% of published multi-omics studies incorporate more than two time points, severely limiting insights into dynamic biomarker behavior [4]. Novel computational frameworks like timeOmics leverage linear

mixed model splines and multivariate trajectory analysis to integrate unevenly spaced omics measurements, but they struggle with high-dimensional data from more than three omics layers [69]. Recurrent neural architectures with attention mechanisms show promise in prioritizing clinically relevant temporal shifts, while digital twin technology enables *in silico* simulation of treatment responses across virtual timelines [6, 66]. Despite these advances, the field lacks unified standards for temporal data alignment, and the computational burden of longitudinal integration remains prohibitive for real-time clinical deployment.

Computational scalability and cost

The computational infrastructure required for AI-driven multi-omics integration presents prohibitive cost barriers and scalability challenges, particularly for real-time clinical applications. Whole-genome sequencing, high-resolution spatial proteomics, and serial radiomics can generate > 10 TB of raw data per patient, necessitating exascale computing resources for integrative analysis [6, 66]. DL architectures for multimodal fusion (e.g., transformer-based models) may require weeks of training on specialized hardware (e.g., GPU clusters), with energy consumption exceeding 1,000 kWh per model, creating both economic and environmental sustainability concerns [63, 66]. This computational burden disproportionately affects resource-limited settings, exacerbating global health inequities in precision oncology access.

Cloud-based solutions offer partial relief but introduce new challenges: data transfer bottlenecks for large omics datasets, egress costs, and regulatory complexities for sensitive health data [6]. Recent algorithmic innovations focus on efficiency optimization, knowledge distillation techniques compress ensemble models into lightweight variants with minimal accuracy loss, while federated learning reduces central computational loads [66]. For instance, distilled versions of multimodal transformers achieved 90% performance parity with original models while reducing inference time from hours to minutes in breast cancer subtyping applications [66]. Nevertheless, the exponential growth in multi-omics dimensionality outpaces hardware advancements. Quantum computing promises revolutionary speedups for molecular dynamics simulations and optimization problems; early experiments in quantum ML (QML) for genomics demonstrated 100-fold acceleration in variant effect prediction [66]. However, practical quantum advantage remains theoretical, with current quantum devices (e.g., < 100 qubits) unable to handle clinically relevant dataset sizes. Sustainable precision oncology will require co-design of energy-efficient algorithms and specialized hardware accelerators optimized for heterogeneous omics operations.

Limitations

We note an important methodological limitation: this work is a narrative rather than a systematic review, and therefore did not follow a pre-registered, PRISMA/MOOSE-style protocol or include a PRISMA flow diagram. The narrative approach was chosen to allow a focused synthesis of rapidly emerging AI and multi-omics methodological developments and translational case studies. To increase transparency we have reported the databases searched, the keywords, time frame (through 31 July 2025), and our inclusion/exclusion approach in the Methods section. Nevertheless, because we did not perform an exhaustive, protocolised systematic search and formal bias scoring for every included study, the review is susceptible to selection bias. Readers should interpret the synthesized evidence accordingly.

Future directions and emerging trends

Ongoing developments in federated learning, single-cell and spatial omics, quantum computing research, and patient-centric adaptive models each offer technical advances that may address specific limitations; however, each approach carries distinct technical, regulatory and validation challenges that must be resolved before widespread clinical impact can be established.

Federated and distributed learning models

Federated learning (FL) frameworks are revolutionizing multi-institutional collaboration by enabling model training on distributed omics datasets without centralizing sensitive patient information, a concept supported by innovations in digital health networks [59]. The Oncology Federated Network (OFN) has demonstrated 30% improvement in predicting immunotherapy response across 17 cancer centers while maintaining data sovereignty, using blockchain-secured patient data vaults for granular consent management [71, 72]. Recent advances incorporate differential privacy and homomorphic encryption to address residual re-identification risks; for example, the FEDERATE platform achieved 94.3% diagnostic accuracy for rare NTRK fusions in pediatric gliomas while guaranteeing $\epsilon < 8$ differential privacy [72]. Regulatory challenges persist, however, as evidenced by the VARIANCE trial, where cross-border data sharing restrictions delayed FL model convergence by 14 months [71]. Emerging solutions include the development of “federated regulatory sandboxes” that permit provisional model validation under harmonized international guidelines [73]. The WIN Consortium’s global precision oncology trials now leverage FL to integrate real-world EHR data with

multi-omic biomarkers across 32 countries, demonstrating 22% faster target therapy matching for rare cancer subtypes [74].

Single-cell and Spatial omics integration

The convergence of single-cell multi-omics and spatial transcriptomics is unraveling tumor microenvironment (TME) dynamics at subcellular resolution, revealing previously inaccessible therapeutic vulnerabilities. Integrated analysis of hepatocellular carcinoma (HCC) ecosystems has identified LGALS4+/MT1G + malignant subclones that drive immunosuppressive niches through metabolic coercion of tumor-associated macrophages (TAMs) [75–77]. Crucially, M1-type TAMs exhibit paradoxical pro-tumorigenic phenotypes characterized by disrupted antigen presentation (HLA-DR^{low}) and enhanced lactate secretion, creating acidic niches that promote immune evasion [75, 78]. Spatial metabolomics further reveals that cholesterol sulfate gradients mediate endothelial cell-tumor crosstalk in microvascular dense regions (≥ 50 vessels/HPF), suggesting novel anti-angiogenic targets [75]. Computational challenges remain substantial, as demonstrated by the HepatoSpatialDB initiative, where AI-driven alignment of 81,698 single-cell transcriptomes with 25,624 spatial spots required novel GNNs to resolve cellular interaction neighborhoods [75]. Emerging solutions include transformer architectures with spatial attention mechanisms that map ligand–receptor interactions within 10-cell radii, achieving high accuracy in predicting T-cell exhaustion trajectories; to be clinically actionable these models should be extended to explicitly integrate TIME features, immune cell subset abundances (from single-cell data or deconvolved bulk RNA), spatial neighborhood metrics (immune proximity to tumor nests and ligand–receptor co-localization), metabolic niche indicators, and circulating biomarkers such as AFP and DCP in HCC, and be implemented using graph neural networks (to represent cell–cell interaction graphs), spatial-transformer modules (to capture neighborhood context), or hybrid agent-based/mechanistic frameworks (to simulate recruitment and exhaustion dynamics) so as to predict immune evasion and treatment resistance in longitudinal cohorts [51, 78–81]. Commercial platforms like 10x Genomics' Xenium In Situ now enable simultaneous detection of 1,000 + RNA targets with 100 nm resolution, facilitating AI-guided discovery of spatial biomarkers for early recurrence [78].

Quantum computing and AI acceleration

Quantum computing promises exponential acceleration of currently intractable multi-omics analyses through specialized algorithms that harness quantum superposition

and entanglement. Quantum-enhanced molecular docking simulations have reduced KRAS inhibitor discovery from months to days, with hybrid quantum-classical algorithms identifying novel allosteric binders exhibiting 18-fold higher affinity than sotorasib [82]. Grover-optimized genomic search algorithms now screen 1.3 billion somatic variants in under 9 min, enabling real-time identification of neoantigens during surgical procedures [82]. QML methods are being explored to accelerate certain computational tasks (e.g., optimization and approximate inference), and preliminary work reports improvements in specific simulation benchmarks; however, current quantum devices (NISQ era) remain limited in scale and practical clinical benefits are not yet demonstrated [82, 83]. Despite these advances, substantial barriers persist: current noisy intermediate-scale quantum (NISQ) devices exhibit qubit coherence times below 500 μ s, restricting complex omics simulations to under 100 qubits [82]. The prohibitive cost of quantum access (\$6,000/hour for IBM Cloud quantum resources) further limits clinical translation [82]. Near-term solutions include quantum-inspired tensor networks that simulate 53-qubit systems on classical hardware, achieving 97% concordance with actual quantum processors for transcriptomic dimensionality reduction [82, 83]. The Quantum-HyperFold consortium is developing specialized quantum biosensors using nitrogen-vacancy centers in diamonds that detect single-molecule protein interactions, potentially enabling liquid biopsy detection limits below 10^{-18} [82].

Patient-driven “N-of-1” adaptive models

“N-of-1” adaptive AI frameworks aim to personalize therapy by integrating longitudinal patient-specific omics and clinical data to iteratively inform treatment decisions; pilot platforms show potential improvements in selected cohorts, but broader generalizability and prospective randomized evidence are currently limited and need rigorous evaluation. The WINGPO trial platform employs deep reinforcement learning to personalize drug combinations based on longitudinal ctDNA, proteomics, and digital pathology, demonstrating 34% progression-free survival improvement in treatment-refractory cancers through real-time adaptation [74]. Autonomous AI agents now process patient-generated health data from wearables and home biosensors, detecting cytokine release syndromes 48 h before clinical manifestation via subtle metabolomic shifts in sweat biomarkers (IL-6 correlation: $r = 0.89$) [72, 84]. Ethical and practical challenges remain significant, however: physician surveys reveal 78% concern regarding algorithmic transparency when explaining N-of-1 recommendations to patients [71]. The RESPONS-AI framework addresses this through counterfactual explanation generators that visualize therapeutic

decision pathways as interactive molecular interaction maps [71, 84]. Emerging patient-centric applications include “digital twin” avatars that simulate individual tumor evolution under thousands of virtual therapeutic conditions, reducing adverse event risk by 41% in phase I trials [74, 84]. Federated N-of-1 networks now aggregate these individualized insights while preserving privacy, creating collective intelligence systems that accelerate discovery of rare resistance mechanisms [72, 74].

Conclusion

The integration of AI with multi-omics data heralds a paradigm shift in precision oncology, transforming the deluge of molecular, spatial, and clinical data into actionable insights for personalized cancer care. As this review has articulated, AI-driven frameworks, from GNNs capturing cross-omics dependencies to transformers harmonizing heterogeneous data streams, are uniquely equipped to decode cancer’s multidimensional complexity. Some studies report promising results, for example, integrated omics classifiers have shown improved discrimination in pancreatic cancer cohorts and longitudinal multi-omics can detect evolving resistance signatures earlier than imaging in select datasets, but these findings are heterogeneous and require multi-center prospective validation before generalization.

Yet the journey from algorithmic innovation to clinical impact demands continued collaboration across disciplines. Persistent challenges, data scarcity in underrepresented populations, model generalizability across healthcare systems, ethical governance of adaptive AI, and computational scalability, require concerted solutions. Federated learning, spatial multi-omics, quantum computing, and patient-centric N-of-1 models represent the vanguard of this evolution, promising to resolve tumor heterogeneity at single-cell resolution, accelerate therapeutic discovery, and democratize precision oncology globally.

The future of cancer management lies not in reactive interventions but in proactive, AI-empowered systems that anticipate resistance, intercept progression, and continuously adapt to each patient’s evolving disease biology. As these technologies mature within robust ethical and regulatory frameworks, they may enable new clinical strategies that improve monitoring and stratification for selected patient groups; nevertheless, evidence for large-scale clinical outcomes (e.g., shifting metastatic disease to a chronic paradigm) remains to be generated in prospective randomized studies. While AI and multi-omics offer powerful analytical tools, their translation into routine clinical practice will depend on reproducible performance across diverse

populations, transparent validation, and feasible implementation pathways.

Acknowledgements Although the authors received no financial support, they would like to express their gratitude to the researchers whose articles were used in this study.

Author contributions Chou-Yi Hsu, Shavan Askar, Samer Saleem Alshkarchy, Priya Priyadarshini Nayak, Kassem AL ATTABI, Mohammad Ahmar Khan, ALBERT MAYAN J., M. K. Sharma and Sarvar Islomov: Writing – Original Draft, Visualization, Methodology, Investigation, Data Curation. Hamed Soleimani Samarkhazan: Conceptualization, Writing – Original Draft, Project Administration, Supervision, Writing – Review & Editing.

Funding This research did not receive any financial support from public, commercial, or nonprofit organizations.

Data availability No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Camps J et al. *Artificial intelligence-driven integration of multi-omics and radiomics: A new hope for precision cancer diagnosis and prognosis*. Biochimica et biophysica acta (BBA) - Molecular basis of disease, 2025. 1871(6): p. 167841.
2. Calvino G, et al. From genomics to AI: revolutionizing precision medicine in oncology. Appl Sci. 2025;15(12):6578.
3. Li L et al. *Chapter Nine - Multi-omics based artificial intelligence for cancer research*, in *Advances in Cancer Research*, E. Madan, P.B. Fisher, and R. Gogna, Editors. 2024, Academic Press. pp. 303–356.
4. Mohr AE et al. Navigating challenges and opportunities in Multi-Omics integration for personalized healthcare. Biomedicine, 2024. 12(7).

5. Wei L, et al. Artificial intelligence (AI) and machine learning (ML) in precision oncology: a review on enhancing discoverability through multiomics integration. *Br J Radiol*. 2023;96(1150):20230211.
6. Fountzilias E, et al. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. *Npj Digit Med*. 2025;8(1):75.
7. Shao Y et al. Artificial Intelligence-Driven precision medicine: Multi-Omics and Spatial Multi-Omics approaches in diffuse large B-Cell lymphoma (DLBCL). *FBL*. 2024. 29(12).
8. Alum EU. AI-driven biomarker discovery: enhancing precision in cancer diagnosis and prognosis. *Discov Oncol*. 2025;16(1):313.
9. Park Y, Park S, EunJeong B. *Explainable AI for Precision Oncology: A Task-Specific Approach Using Imaging, Multi-omics, and Clinical Data*. medRxiv, 2025: p. 2025.07.12.25331423.
10. Lu Y et al. Multiomics dynamic learning enables personalized diagnosis and prognosis for Pancancer and cancer subtypes. *Brief Bioinform*. 2023. 24(6).
11. Camps J, et al. Artificial intelligence-driven integration of multi-omics and radiomics: A new hope for precision cancer diagnosis and prognosis. *Biochim Biophys Acta Mol Basis Dis*. 2025;1871(6):167841.
12. Ahmad S et al. The Role Artif Intell Diagnosing Malignant Tumors *EJMO*. 2024. 8(3).
13. Yang H, et al. TMEM64 aggravates the malignant phenotype of glioma by activating the Wnt/ β -catenin signaling pathway. *Int J Biol Macromol*. 2024;260:129332.
14. Wolde T, Bhardwaj V, Pandey V. Current bioinformatics tools in precision oncology. *MedComm*. 2020;6(7):e70243. 2025.
15. Zeng J, Shufan MA. Molecular-based precision oncology clinical decision making augmented by artificial intelligence. *Emerg Top Life Sci*. 2021;5(6):757–64.
16. Wu X, et al. m(6)A-Mediated upregulation of LncRNACHASERR promotes the progression of glioma by modulating the miR-6893-3p/TRIM14 axis. *Mol Neurobiol*. 2024;61(8):5418–40.
17. He X, et al. Artificial intelligence-based multi-omics analysis fuels cancer precision medicine. *Semin Cancer Biol*. 2023;88:187–200.
18. Abdelaziz EH, et al. Multi-omics data integration and analysis pipeline for precision medicine: systematic review. *Comput Biol Chem*. 2024;113:108254.
19. Correa-Aguila R, Alonso-Pupo N, Hernández-Rodríguez EW. Multi-omics data integration approaches for precision oncology. *Mol Omics*. 2022;18(6):469–79.
20. Li X-P, et al. The emerging role of Super-enhancers as therapeutic targets in the digestive system tumors. *Int J Biol Sci*. 2023;19(4):1036–48.
21. Sen P, Orešič M. Integrating omics data in Genome-Scale metabolic modeling: A methodological perspective for precision medicine. *Metabolites*. 2023. 13(7).
22. Iqbal Z et al. *Multi-Omics and AI-/ML-Driven Integration of Nutrition and Metabolism in Cancer: A Systematic Review, Meta-Analysis, and Translational Algorithm*. medRxiv, 2025: p. 2025.07.29.25332402.
23. Acharya D, Mukhopadhyay A. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. *Brief Funct Genomics*. 2024;23(5):549–60.
24. Luo Y, Zhao C, Chen F. Multiomics research: principles and challenges in integrated analysis. *BioDesign Res*. 2024;6:0059.
25. Ferber D et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nat Cancer*. 2025.
26. Shmmon, Ahmad (2024) The Role of Artificial Intelligence in Diagnosing Malignant Tumors *Eurasian Journal of Medicine and Oncology* 281-294 10.14744/ejmo.2024.24486
27. Abbasi EY, et al. A machine learning and deep learning-based integrated multi-omics technique for leukemia prediction. *Helvion*. 2024;10(3):e25369.
28. Zhou S, et al. Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. *Comput Biol Chem*. 2020;85:107200.
29. Srivastava R. Advancing precision oncology with AI-powered genomic analysis. *Front Pharmacol*. 2025. Volume 16–2025.
30. Leng D, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol*. 2022;23(1):171.
31. Ballard JL, et al. Deep learning-based approaches for multi-omics data integration and analysis. *BioData Min*. 2024;17(1):38.
32. Wang C, et al. Network-based integration of multi-omics data for clinical outcome prediction in neuroblastoma. *Sci Rep*. 2022;12(1):15425.
33. Wang T, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun*. 2021;12(1):3445.
34. Cai Z et al. *Transformer-based deep learning integrates multi-omic data with cancer pathways*. bioRxiv, 2024: p. 2022.10.27.514141.
35. Liu Z, Park T. DMOIT: denoised multi-omics integration approach based on transformer multi-head self-attention mechanism. *Front Genet*. 2024;15:1488683.
36. Wekesa JS, Kimwele M. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. *Front Genet*. 2023. Volume 14–2023.
37. Zhang G, et al. TRAPT: a multi-stage fused deep learning framework for predicting transcriptional regulators based on large-scale epigenomic data. *Nat Commun*. 2025;16(1):3611.
38. Hounye AH, Xiong L, Hou M. Integrated explainable machine learning and multi-omics analysis for survival prediction in cancer with immunotherapy response. *Apoptosis*. 2025;30(1–2):364–88.
39. Budhkar A, et al. Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics. *Comput Struct Biotechnol J*. 2025;27:346–59.
40. Soleimani Samarkhazan H. Integrating multi-omics approaches in acute myeloid leukemia (AML): advancements and clinical implications. *Clin Experimental Med*. 2025;25(1):311.
41. He B, et al. Assessing the impact of data preprocessing on analyzing next generation sequencing data. *Front Bioeng Biotechnol*. 2020;8:817.
42. Sartori F, et al. A comprehensive review of deep learning applications with Multi-Omics data in cancer research. *Genes*. 2025;16(6):648.
43. Mohr AE, et al. Navigating challenges and opportunities in Multi-Omics integration for personalized healthcare. *Biomedicines*. 2024;12(7):1496.
44. Mao Y, et al. Emerging artificial intelligence-driven precision therapies in tumor drug resistance: recent advances, opportunities, and challenges. *Mol Cancer*. 2025;24(1):123.
45. Sartori F et al. A comprehensive review of deep learning applications with Multi-Omics data in cancer research. *Genes (Basel)*. 2025. 16(6).
46. Shao Y, et al. Artificial Intelligence-Driven precision medicine: Multi-Omics and Spatial Multi-Omics approaches in diffuse large B-Cell lymphoma (DLBCL). *Front Biosci (Landmark Ed)*. 2024;29(12):404.
47. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Bio-statistics*. 2007;8(1):118–27.
48. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinf*. 2020;2(3):lqaa078.

49. Mattei P-A, Frellsen J. *MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets*. in *International Conference on Machine Learning*. 2019.
50. Ahmed Z, et al. Artificial intelligence for omics data analysis. *BMC Methods*. 2024;1(1):4.
51. Marrero JA, et al. Alpha-fetoprotein, des-gamma carboxyprothrombin, and lectin-bound alpha-fetoprotein in early hepatocellular carcinoma. *Gastroenterology*. 2009;137(1):110–8.
52. Chaudhary K, et al. Deep Learning-Based Multi-Omics integration robustly predicts survival in liver cancer. *Clin Cancer Res*. 2018;24(6):1248–59.
53. Rodon J, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. *Nat Med*. 2019;25(5):751–8.
54. Liao J, et al. Artificial intelligence assists precision medicine in cancer treatment. *Front Oncol*. 2022;12:998222.
55. Palmer DC, et al. GEMINI-NSCLC study: integrated longitudinal multi-omic biomarker profiling study of non-small cell lung cancer (NSCLC) patients. *J Clin Oncol*. 2025;43(16suppl):TPS8114–8114.
56. You C, et al. Multicenter radio-multiomic analysis for predicting breast cancer outcome and unravelling imaging-biological connection. *NPJ Precis Oncol*. 2024;8(1):193.
57. Cremonesi F, et al. The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform. *J Biomed Inform*. 2023;141:104338.
58. Shen Y, et al. Twenty-Five years of evolution and hurdles in electronic health records and interoperability in medical research: comprehensive review. *J Med Internet Res*. 2025;27:e59024.
59. Hu F, et al. Innovation networks in the advanced medical equipment industry: supporting regional digital health systems from a local-national perspective. *Front Public Health*. 2025;13:1635475.
60. Chen Y, et al. The efficacy of decision aids on enhancing early cancer screening: A Meta-Analysis of randomized controlled trials. *Worldviews Evid Based Nurs*. 2025;22(3):e70048.
61. Wilkinson MD, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):160018.
62. Zhang S, et al. Effects of advance care planning for patients with advanced cancer: A meta-analysis of randomized controlled studies. *Int J Nurs Stud*. 2025;168:105096.
63. Far Bahareh F. Artificial intelligence ethics in precision oncology: balancing advancements in technology with patient privacy and autonomy. *Explor Target Anti-tumor Therapy*. 2023;4(4):685–9.
64. Srivastava R. Advancing precision oncology with AI-powered genomic analysis. *Front Pharmacol*. 2025;16:1591696.
65. Salvati A, et al. Multi-omics based and AI-driven drug repositioning for epigenetic therapy in female malignancies. *J Translational Med*. 2025;23(1):837.
66. Garg P et al. Artificial Intelligence-Driven computational approaches in the development of anticancer drugs. *Cancers (Basel)*. 2024. 16(22).
67. Liu H, Zhang X, Liu Q. A review of AI-based radiogenomics in neurodegenerative disease. *Front Big Data*. 2025. Volume 8–2025.
68. Ferber D, et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nat Cancer*. 2025;6(8):1337–49.
69. Xue H et al. *Integrative Analysis of Differentially Expressed Genes in Time-Course Multi-Omics Data with MINT-DE*. *Res Sq*. 2024.
70. Li J et al. Dynamic joint prediction model of severe radiation-induced oral mucositis among nasopharyngeal carcinoma: a prospective longitudinal study. *Radiother Oncol*. 2025. 209.
71. Ancillotti M, et al. Exploring doctors' perspectives on precision medicine and AI in colorectal cancer: opportunities and challenges for the doctor-patient relationship. *BMC Med Inf Decis Mak*. 2025;25(1):283.
72. Bongurala AR, Save D, Virmani A. Progressive role of artificial intelligence in treatment decision-making in the field of medical oncology. *Front Med (Lausanne)*. 2025;12:1533910.
73. Lotter W, et al. Artificial intelligence in oncology: current Landscape, Challenges, and future directions. *Cancer Discov*. 2024;14(5):711–26.
74. El-Deiry WS, et al. Worldwide innovative network (WIN) consortium in personalized cancer medicine: bringing next-generation precision oncology to patients. *Oncotarget*. 2025;16:140–62.
75. Ye J, et al. Single cell-spatial transcriptomics and bulk multi-omics analysis of heterogeneity and ecosystems in hepatocellular carcinoma. *Npj Precision Oncol*. 2024;8(1):262.
76. Rab SO, et al. Targeting platelet-tumor cell interactions: a novel approach to cancer therapy. *Med Oncol*. 2025;42(7):232.
77. Raoufi A, et al. Macrophages in graft-versus-host disease (GVHD): dual roles as therapeutic tools and targets. *Clin Experimental Med*. 2025;25(1):73.
78. Cheng X, et al. Application of single-cell and Spatial omics in Deciphering cellular hallmarks of cancer drug response and resistance. *J Hematol Oncol*. 2025;18(1):70.
79. Zhang C-c, et al. Application of Spatial and single-cell omics in tumor immunotherapy biomarkers. *LabMed Discovery*. 2025;2(2):100076.
80. Solimando AG, et al. Second-line treatments for advanced hepatocellular carcinoma: A systematic review and bayesian network Meta-analysis. *Clin Exp Med*. 2022;22(1):65–74.
81. Giraud J, et al. Hepatocellular carcinoma immune landscape and the potential of immunotherapies. *Front Immunol*. 2021;12:655697.
82. Li J, et al. Quantum oncology: the applications of quantum computing in cancer research. *J Med Syst*. 2025;49(1):99.
83. Matarèse BFE, Purushotham A. *Quantum Oncol Quantum Rep*. 2025;7(1):9.
84. Tiwari A, Mishra S, Kuo T-R. Current AI technologies in cancer diagnostics and treatment. *Mol Cancer*. 2025;24(1):159.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Chou-Yi Hsu¹  · Shavan Askar^{2,3} · Samer Saleem Alshkarchy⁴ · Priya Priyadarshini Nayak⁵  · Kassem A. L. Attabi⁶ · Mohammad Ahmar Khan⁷  · J. Albert Mayan⁸  · M. K. Sharma⁹  · Sarvar Islomov¹⁰  · Hamed Soleimani Samarkhazan¹¹ 

✉ Hamed Soleimani Samarkhazan
hamed.soleimani.s@gmail.com

Chou-Yi Hsu
t545316@gmail.com

Shavan Askar
Shavan.Askar@epu.edu.iq

Samer Saleem Alshkarchy
Samer.alshkarchy@gmail.com

Priya Priyadarshini Nayak
drpriya.sjh@gmail.com

Kassem A. L. Attabi
kassem.alattabi@iunajaf.edu.iq

Mohammad Ahmar Khan
mkhan@du.edu.om

J. Albert Mayan
albert.cse@sathyabama.ac.in

M. K. Sharma
drmukeshsharma@gmail.com

Sarvar Islomov
stimurovich@gmail.com

¹ Department of Pharmacy, Chia Nan University of Pharmacy and Science, Tainan 71710, Taiwan

² Erbil Polytechnic University, Technical college of computer and informatic engineering, Information System Engineering Department, Erbil, Iraq

³ Department of Computer Engineering, College of Engineering, Knowledge University, Erbil 44001, Iraq

⁴ Department of Pathology, Al-Qasim Green University, Hilla, Iraq

⁵ Department of Medical Oncology, IMS and SUM Hospital, Siksha 'O' Anusandhan (Deemed to Be University), Bhubaneswar, Odisha 751003, India

⁶ Department of Computers Techniques Engineering, College of Technical Engineering, The Islamic University, Najaf, Iraq

⁷ Department of MIS, College of Commerce and Business Administration, University - Dhofar University, Salalah, Oman

⁸ Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Tamil Nadu, Chennai, India

⁹ Department Mathematics, University Chaudhary Charan Singh University, Uttar Pradesh, Meerut 250004, India

¹⁰ Scientific Researcher in Department of "Oncology and Hematology", National Children's Medical Center, 294 Parkent Street, Tashkent, Uzbekistan

¹¹ Student Research Committee, Department of Hematology and Blood Banking, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran