nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00942-3

Moving towards genome-wide data integration for patient stratification with Integrate Any Omics

Received: 24 April 2024

Accepted: 31 October 2024

Published online: 23 January 2025

Check for updates

Shihao Ma[1,2,3], Andy G. X. Zeng[4,5], Benjamin Haibe-Kains[3,5,6], Anna Goldenberg[1,3,7,8,9], John E. Dick[4,5] & Bo Wang[1,2,3,8,9]

High-throughput omics profiling advancements have greatly enhanced cancer patient stratification. However, incomplete data in multi-omics integration present a substantial challenge, as traditional methods like sample exclusion or imputation often compromise biological diversity and dependencies. Furthermore, the critical task of accurately classifying new patients with partial omics data into existing subtypes is commonly overlooked. To address these issues, we introduce Integrate Any Omics (IntegrAO), an unsupervised framework for integrating incomplete multi-omics data and classifying new samples. IntegrAO first combines partially overlapping patient graphs from diverse omics sources and utilizes graph neural networks to produce unified patient embeddings. Our systematic evaluation across five cancer cohorts involving six omics modalities demonstrates IntegrAO's robustness to missing data and its accuracy in classifying new samples with partial profiles. An acute myeloid leukaemia case study further validates its capability to uncover biological and clinical heterogeneities in incomplete datasets. IntegrAO's ability to handle heterogeneous and incomplete data makes it an essential tool for precision oncology, offering a holistic approach to patient characterization.

Precision medicine, which tailors personalized treatment based on a patient's unique genetic profile, has been recognized as the future of cancer therapeutics[1]. The field is moving towards gathering multimodal data to address cancer's inherent heterogeneity[2], characterized by diverse genetic, transcriptomic and phenotypic variations[3,4]. Recent advancements in high-throughput technologies have enabled multi-dimensional profiling. Projects like The Cancer Genome Atlas (TCGA)[5] and the International Cancer Genome Consortium[6] have produced and collected thousands of tumour samples at different molecular levels. The rise of single-cell profiling, particularly single-cell transcriptomics, has deepened insights into tumour microenvironments by highlighting the distinct expression profiles of various cell types. Consequently, patient stratification based on genetic, transcriptomic and phenotypic data has become central to precision medicine, guiding the development of tailored treatments.

Integrating multi-omics data offers a more holistic understanding of cancer. Numerous methods have emerged over the past decade, including network-based methods[7–9], matrix-factorization-based approaches[10,11], Bayesian clustering techniques[12,13] and advanced deep learning approaches[14,15], which have shown successes in disease

[1]Department of Computer Science, University of Toronto, Toronto, Canada. [2]Peter Munk Cardiac Centre, University Health Network, Toronto, Canada. [3]Vector Institute for Artificial Intelligence, Toronto, Canada. [4]Department of Molecular Genetics, University of Toronto, Toronto, Canada. [5]Princess Margaret Cancer Centre, University Health Network, Toronto, Canada. [6]Department of Medical Biophysics, University of Toronto, Toronto, Canada. [7]Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Canada. [8]Canadian Institute for Advanced Research, Toronto, Canada. [9]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada. e-mail: bowang@vectorinstitute.ai

subtyping[16,17] and advancing precision medicine[18]. However, these methods often require complete data, which are rarely available due to experimental and financial constraints. For example, genotyping data may be complete for all patients, but gene expression or methylation data are frequently incomplete[19]. Analysing such incomplete omics data is challenging. Excluding samples with missing omics data reduces sample sizes, and imputing missing values can introduce bias and uncertainty[20,21]. This underscores the critical need for computational approaches that can directly handle incomplete datasets without discarding valuable information. Advanced integrative methods to address the missing data issue can be classified into two categories: joint imputation or optimization masking[22]. Although joint imputation approaches[23–25] predict missing values within the modelling framework, they often introduce potential biases and require large sample sizes. Optimization masking techniques[8,14,26,27], which work with processed data such as patient graphs, allow partial samples to contribute by masking missing data during the optimization process, but face challenges such as increased computational complexity, potential clustering inaccuracies as graph numbers grow and the need for at least one common data view, which is not always feasible.

Despite these limitations, multi-omics integration provides valuable diagnostic and prognostic insights. A substantial challenge remains in accurately classifying new patients into predefined subtypes, particularly when dealing with incomplete omics data from these individuals[28]. This limitation hinders the clinical adoption of molecular subtypes since many patients present with partial datasets. Addressing this gap by developing methods that can infer accurate subtypes from any available data is essential for advancing personalized patient care and fully realizing the potential of multi-omics integration in medicine.

To overcome these limitations, we present Integrate Any Omics (IntegrAO), an unsupervised framework for integrating incomplete multi-omics profiles and classifying new samples. IntegrAO first uses a unique partial graph fusion mechanism to integrate overlapping patient graphs from diverse omics sources, preserving data fidelity and minimizing noise. It then applies graph neural networks (GNNs) to extract and align patient embeddings into a unified space, enabling the accurate classification of new patients, even with incomplete data. To demonstrate the use of IntegrAO, we first show that IntegrAO exhibits robust integration across diverse missing data scenarios through the simulation of an omics dataset. A case study in acute myeloid leukaemia (AML) then illustrates IntegrAO's capacity to build a comprehensive view of heterogeneity from incomplete multi-omics. Systematic evaluations conducted on five cancer cohorts, covering six omics modalities, underscore IntegrAO's resilience to missing data and its effectiveness in integrating partial data and classifying new samples. Through its proficient handling of heterogeneous and incomplete datasets, IntegrAO emerges as an essential tool in precision oncology, facilitating an all-encompassing approach to patient characterization.

Results

Overview

We present IntegrAO, an unsupervised framework for integrating multi-omics datasets with partial overlap. As outlined in Fig. 1, IntegrAO has two key functionalities: transductive integration and inductive prediction.

Transductive integration consists of two core steps: (1) fusion of partially overlapping patient graphs; (2) extraction and alignment of patient embeddings across omics modalities (Fig. 1a). In step (1), IntegrAO accommodates samples with missing data types by constructing a patient graph for each omics data modality, where the nodes represent the patients and weighted edges denote the pairwise similarities (see 'Patient graph construction' in the Methods). IntegrAO then iteratively fuses these graphs (see 'Partial overlap graph fusion' in the Methods), using shared samples as bridges to propagate information across omics, even to patients with partial data. The degree of overlap between omics determines the extent of information fusion, with more shared patients enhancing integration. As patient overlap varies between omics modalities, IntegrAO performs pairwise fusion between graphs to maximize the information flow. Step (1) yields a fused graph for each omics data modality that incorporates information from others. Step (2) extracts low-dimensional patient embeddings from each omics data modality into a unified space (see 'Embedding extraction and alignment in transductive integration' in the Methods). Fused networks and omics data are passed through omics-specific GNN encoders and a shared projection head to generate embeddings. During training, the model ensures that the embeddings preserve similarity structures from the input graphs and aligning patient embeddings across different omics. Final embeddings are obtained by averaging across omics to construct the integrated graph.

Inductive prediction extends the unsupervised framework for supervised tasks (Fig. 1b). For example, after cancer subtypes are identified from the integrated graph, IntegrAO can be further fine-tuned to predict subtypes for new patients using any available omics data (see 'Model fine-tuning for subtype prediction' in the Methods). The prediction model builds on the unsupervised IntegrAO framework by adding a multilayer perceptron (MLP) head, which processes the averaged patient embeddings for accurate predictions. The pretrained model provides the initial weights, whereas the MLP head is initialized with random weights, ensuring robust feature extraction and adapting to the supervised task. Fine-tuning balances the objectives of embedding generation and subtype classification. This dual optimization enables the model to support subtype prediction in a modality-agnostic manner. During inference, new patients' omics data are fused into existing graphs, and the fine-tuned model predicts their cancer subtypes given the fused graphs along with the corresponding omics features (see 'Subtype prediction for new patient' in the Methods).

Robust integration across diverse simulated missing data scenarios

We evaluated IntegrAO using a simulated multi-omics dataset generated by the InterSim CRAN package[29], which produces data for three omics modalities (DNA methylation, mRNA expression and protein expression). We simulated a total of 500 samples distributed across 15 clusters, each of which had a variable, randomly determined size (see 'Data processing' in the Methods). We compared IntegrAO with two network-based methods, neighbourhood-based multi-omics clustering (NEMO)[8] and multiple similarity network embedding (MSNE)[26], using normalized mutual information (NMI) to assess clustering congruence with ground-truth labels. NEMO and MSNE were run using their default settings and hyperparameters.

We first tested the scenarios in which one omics modality remains intact and two other modalities undergo uniform random subsampling at ratios from 0.1 to 0.9, repeating this process ten times for each ratio (Fig. 2a). In integration scenarios with a partial overlap, two regimes emerge: low overlap, where minimizing noise from limited shared samples is crucial; and high overlap, where maximizing information flow between modalities improves integration. Across all the overlap ratios, IntegrAO consistently outperformed other methods and remained robust even when MSNE struggled under low-overlap conditions (Fig. 2a). In particular, MSNE's performance can decline when integrating more data under low overlap, sometimes falling below the baseline levels set by *K*-means clustering on the intact modality, highlighting its limitations with noisy, low-overlap data. Next, we evaluated a more complex setup with no intact omics modality. We selected a subset of common samples based on the specified overlapping ratio and then evenly distributed the remaining samples among the three modalities as unique entities. We observed enhanced clustering performance in all the three methods as the overlapping ratios increased. IntegrAO consistently outperformed other methods, maintaining effective clustering even at a minimal 10% overlap (Fig. 2b). IntegrAO's strength lies in its

https://doi.org/10.1038/s42256-024-00942-3

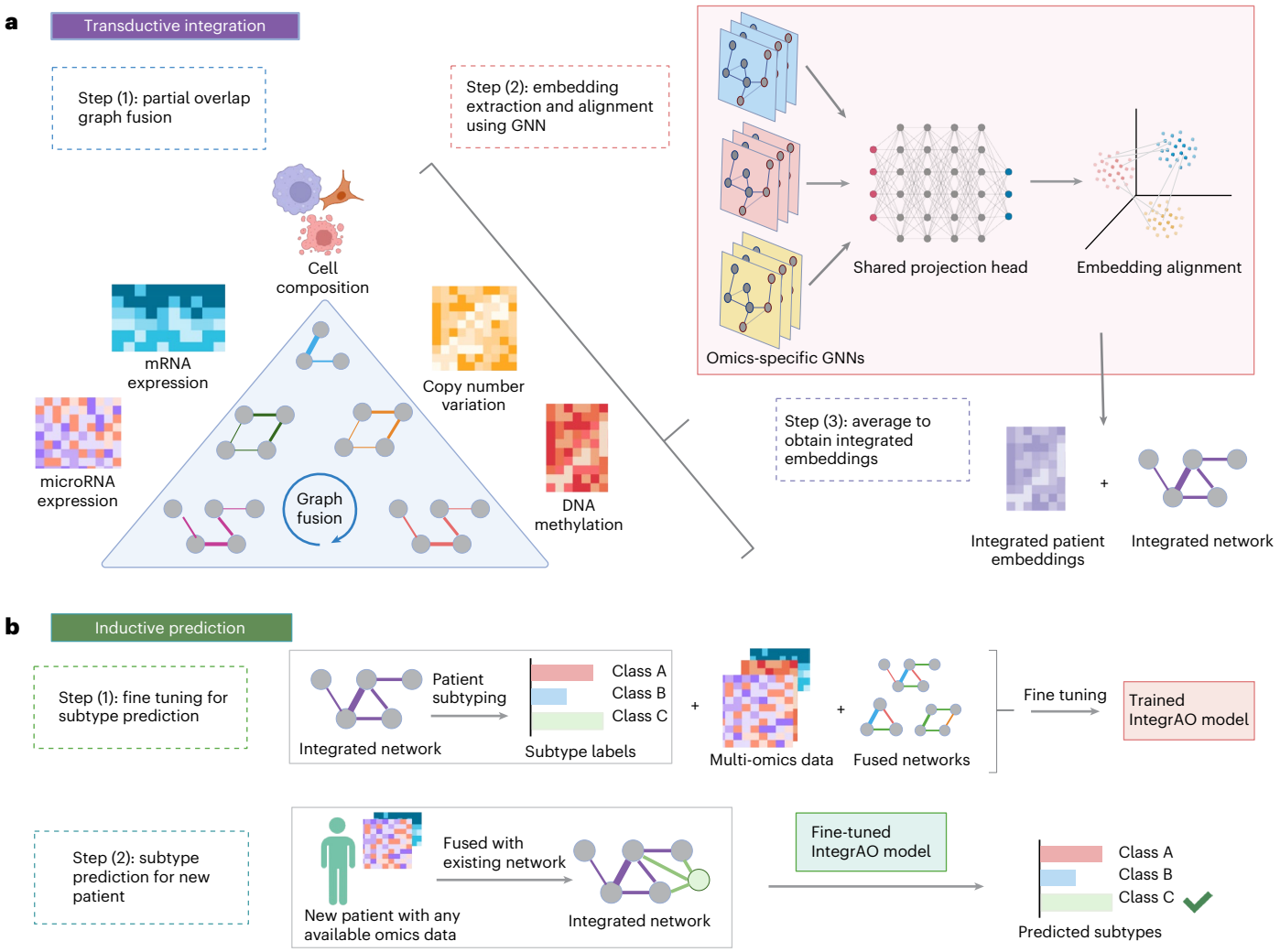


Fig. 1 | Overview of the IntegrAO framework. a, Step (1): example representation of cell composition, mRNA expression, microRNA expression, DNA methylation and copy number variation datasets are used to construct per-omics patient graphs. Patient data need not encompass all omics types. Subsequently, a fusion phase iteratively refines each graph with information gathered from other graphs, culminating in a unified graph for each type of omics. Step (2): both these unified graphs and their corresponding omics features are input into omics-specific GNNs to learn patient embeddings. These low-dimensional patient embeddings are optimized to retain similarity information from the individual unified graphs and minimize differences in embeddings for the same patients across different omics. Step (3): the conclusive embeddings are procured by averaging omics-specific embeddings and applied in the construction of the final integrated patient graph. **b**, Conversion of IntegrAO into a predictive framework. Utilizing the integrated graph, patient subtypes can be identified and leveraged to fine-tune the trained IntegrAO model. The fine-tuned IntegrAO model enables the classification of new patients with any accessible omics data. During the inference process, graph fusion is first conducted on new patients along with existing patients. The consequent fused graph and associated omics features are then input into the fine-tuned IntegrAO model, allowing for the prediction of patient subtypes.

ability to fuse unique samples even just with their modality, in contrast to NEMO, which requires samples to be observed in at least one common view with the others. In another challenging setup in which there were no common samples across the three modalities, we varied the pairwise overlap ratio of all the three pairs of modalities from 0.1 to 0.3 (maximum, 0.33) to evaluate how integration methods cope with varied pairwise overlap. Samples not shared between any two modalities were distributed evenly across the three. The results show that IntegrAO exhibited notable resilience (Fig. 2c). By contrast, NEMO and MSNE experienced substantial performance drops in the absence of common samples. This shows that IntegrAO effectively uses pairwise overlaps, eliminating the need for common samples across all modalities—a flexibility that suits real-world datasets with varying degrees of overlap.

To further investigate IntegrAO's integration effectiveness, we conducted a detailed visual analysis on a 70% overlap scenario with 350 shared samples and 50 unique samples per modality (Fig. 2b). We generated uniform manifold approximation and projection (UMAP) visualizations for each omics data modality before integration. In these visualizations, dots represent the shared samples, whereas diamonds, squares and triangles indicate unique samples of mRNA, protein and DNA methylation, respectively (Fig. 2d). Before integration, the embeddings displayed an entangled structure with randomly dispersed unique samples. By contrast, after IntegrAO integration, the UMAP shows clearly defined clustering of 15 clusters, with the coherent grouping of unique samples (Fig. 2e). This highlights IntegrAO's ability to disentangle complex mixed signals and uncover integrated structures through the joint analysis of distinct but partially overlapping datasets.

Identifying distinct clinical and biological AML subtypes

To elucidate heterogeneity in AML, a cancer marked by extensive interpatient and intrapatient heterogeneity, we applied IntegrAO to

https://doi.org/10.1038/s42256-024-00942-3

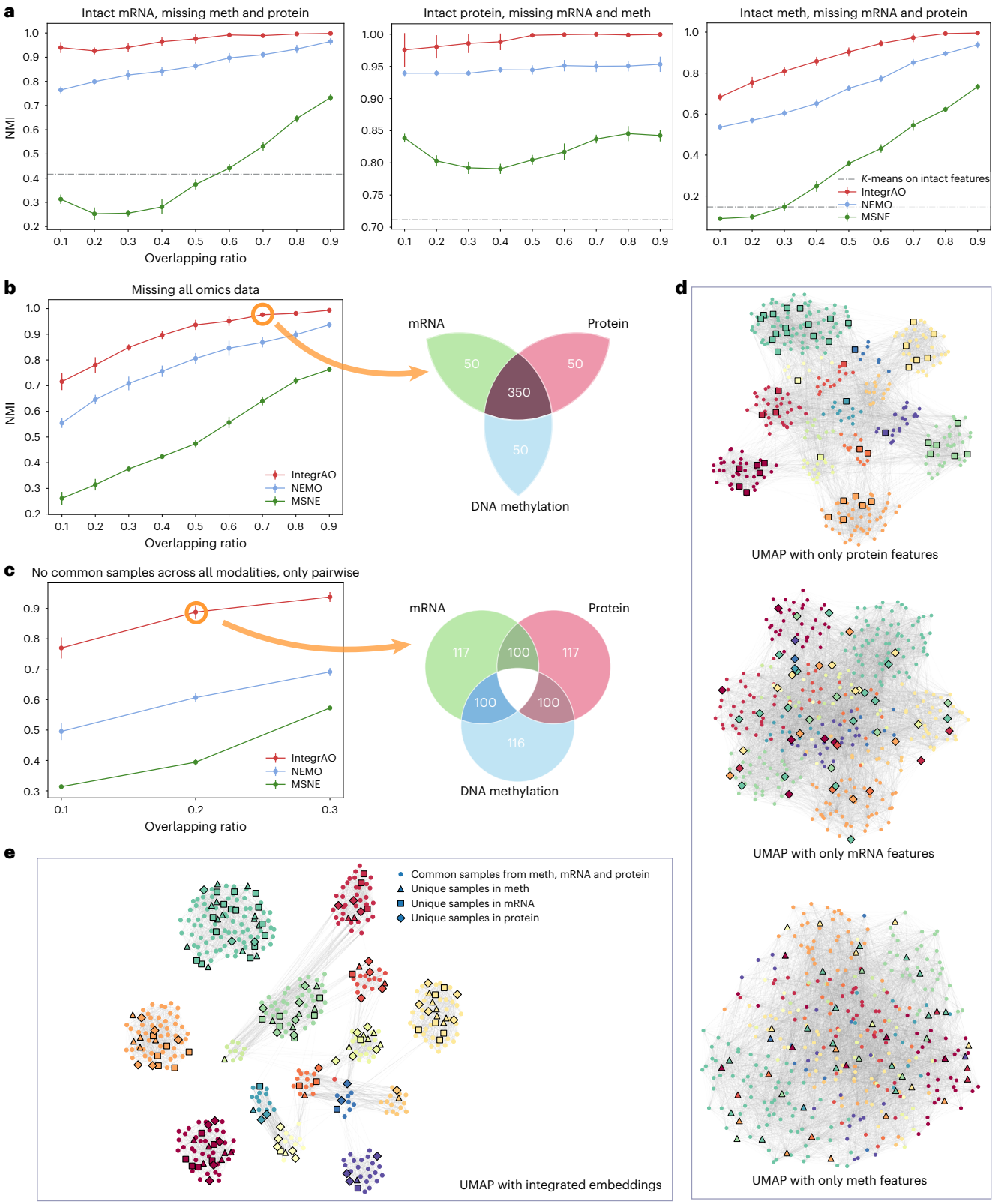


Fig. 2 | Benchmarking partial multi-omics integration with IntegrAO, NEMO and MSNE on a simulated cancer dataset using NMI. a, NMI versus overlapping data ratio across three missing scenarios ($n = 10$ experiments per ratio). From left to right, (1) random subsampling of DNA methylation and protein expression, (2) subsampling of mRNA expression and DNA methylation and (3) subsampling of mRNA and protein expression. IntegrAO shows superior performance in all scenarios. **b**, IntegrAO outperforms other methods when all the omics data are partially missing, illustrated with a 70% overlap (350 common and 50 unique samples per modality). **c**, IntegrAO excels even without common samples, demonstrated with 20% overlap (100 samples with pairwise overlap). For the line plots shown in **a**–**c**, the dots represent the mean value of ten experiments, and the error bars represent ±one standard deviation. **d**, UMAP visualizations of each modality before integration for the 70% overlap scenario, showing both common and unique samples. **e**, Post-integration UMAP using IntegrAO reveals enhanced clustering and alignment of unique samples across modalities.

an empirical AML dataset. Recently, a new dimension of heterogeneity has been identified in AML corresponding to the composition of each patient's leukaemia cell hierarchy[30], providing new insights into disease biology and drug response. We sought to utilize IntegrAO to integrate this new information with two other modalities, namely, mRNA expression and DNA methylation, to achieve an unprecedented multidimensional perspective on AML heterogeneity. Thus, we applied IntegrAO to three AML cohorts, namely, TCGA[31], BEAT-AML[32] and Leucegene[33], leveraging mRNA expression and hierarchy composition for 812 patients, and methylation profiles from 308 patients of those patients (see 'Data processing' in the Methods).

IntegrAO identified 12 biologically distinct AML subtypes (see 'Cluster number selection' in the Methods), refining previous groupings defined by cell hierarchy alone[30]. These subtypes showed clear patterns in cell composition, transcriptional profiles, methylation and genomic alterations (Fig. 3a,b). For example, 'Primitive' subtypes were enriched for primitive leukaemia stem and progenitor cells, 'Mature' subtypes for mono-like and conventional dendritic cells (cDC-like cells), whereas other subtypes had distinct cell type enrichments. Despite similar compositions, the two Primitive subtypes were distinguished by mutations: Primitive (NPM1) associated with NPM1/FLT3-ITD alterations and Primitive (canonical) with TP53/RUNX1. Further heterogeneity is observed in the four NPM1-driven subtypes with divergent cell hierarchies. A novel subtype dominated by erythroid progenitor cells emerged, potentially offering new research directions. Further analysis revealed clear differences in the enrichment of biological and metabolic pathways across subtypes (Supplementary Fig. 1), and gene regulatory analysis using VIPER analyses nominated transcriptional regulators that may drive disease biology in each subtype (Supplementary Fig. 2). These findings demonstrate IntegrAO's ability to uncover the deep biological heterogeneity of AML, informing potential therapeutic strategies.

We further assessed the clinical importance of the subtypes through survival analysis and drug sensitivity profiling. Kaplan–Meier survival curves for the combined TCGA and BEAT-AML cohorts showed significant differences (multigroup log-rank test P value = 1.21×10^{-7}) (Fig. 3c), with similar results in separate analyses (Supplementary Fig. 3a,b). Nested likelihood ratio tests showed that IntegrAO subtypes improved prognostic stratification beyond four established factors (age, cytogenetic risk, white blood cell count and NPM1 mutation), retaining independent significance in multivariable analysis (P value = 0.035), whereas subtypes from individual data types did not ($P > 0.05$) (Supplementary Fig. 3c). For drug sensitivity, analysis of variance tests were used to assess whether IntegrAO subtypes show differential responses to each of 122 anti-cancer agents in the BEAT-AML drug screening dataset (Fig. 3c). The analysis revealed that 47 out of 122 drugs showed differential responses across IntegrAO subtypes (analysis of variance, $P < 0.05$), supporting their clinical utility (Supplementary Fig. 4).

We hypothesized that the biological heterogeneity identified across our IntegrAO subtypes may reflect their disparate origins along normal haematopoietic differentiation. To test this, we evaluated subtype enrichment across stages of haematopoietic differentiation using a single-cell transcriptomic reference defined elsewhere[34] (Fig. 3d). Mapping the top 100 gene markers for each IntegrAO subtype to this reference revealed alignments, such as the 'Dendritic' subtype with plasmacytoid and conventional dendritic cells, Primitive (canonical) with haematopoietic stem cells and Mature Mono (NPM1) with monocytes. This validation further confirms that IntegrAO subtypes preserve distinct haematopoietic lineages, reflecting AML's intertumour heterogeneity.

In summary, the IntegrAO integration of complete and incomplete AML data effectively identifies distinct subtypes with biological and clinical relevance. By sharing information across modalities and preserving key distinctions, IntegrAO offers a comprehensive insight into cancer complexity, furthering biological discovery and precision medicine. Its ability to align patient stratification with clinical outcomes and disease biology demonstrates its potential for guiding personalized therapies, particularly for complex diseases like AML.

Pan-cancer evaluation on subtype identification

To further evaluate the efficacy of partial multi-omics integration, we compared IntegrAO with NEMO and MSNE across five distinct cancer datasets sourced from TCGA. For each cancer type, we leveraged the maximum number of patients in each of the five omics: mRNA expression, DNA methylation, miRNA expression, reverse-phase protein array and copy number variation. By utilizing all the possible samples from TCGA, this benchmark dataset encompasses rich, heterogeneous profiles without data waste. Additionally, we incorporated the cell type composition as an extra modality to enhance heterogeneity characterization (see 'Gene expression deconvolution' in the Methods). Details on data collection and preprocessing are available in the 'Data processing' section in the Methods, with patient and feature counts summarized in Supplementary Tables 1 and 2.

To evaluate the effectiveness of a given clustering solution, two specific metrics were used. First, age-adjusted differential survival between the resultant clusters was measured using the log-rank test. This method operates on the premise that clusters with significant differences in survival rates reflect biologically meaningful variations. Subsequently, we examined the enrichment of six clinical labels within the clusters, including gender, age at diagnosis, pathologic T (tumour progression), pathologic M (metastases), pathologic N (cancer in lymph nodes) and pathologic stage (total progression). Enrichment for discrete parameters was assessed using the χ^2 test for independence, whereas numeric parameters were evaluated using the Kruskal–Wallis test. Recognizing the absence of a definitive ground truth for the number of clusters pertaining to each cancer type, we executed clustering for a range of cluster numbers from 3 to 8. Figure 4 illustrates the comparative performance of IntegrAO against other methods across various cancer datasets.

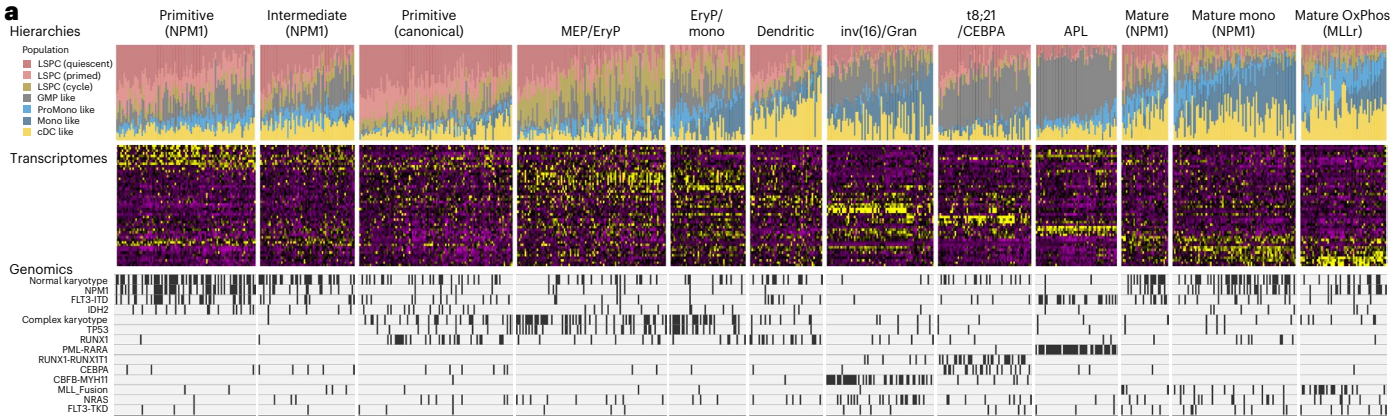
IntegrAO consistently identified subtypes with superior survival differentiation and clinical variable enrichment across cancer cohorts. For breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC) and lung adenocarcinoma (LUAD), IntegrAO solutions were clearly favourable considering both criteria. In skin cutaneous melanoma (SKCM), IntegrAO achieved better clinical enrichment and matching NEMO's performance on survival differentiation. In colon adenocarcinoma (COAD), IntegrAO outperformed the other methods in survival stratification despite suboptimal clinical enrichment results. By contrast, NEMO and MSNE showed inconsistent results. MSNE performed well in COAD but poorly in KIRC and SKCM, whereas NEMO excelled in BRCA but underperformed in COAD and LUAD. The inconsistency in MSNE's survival differentiation, effective in BRCA but not in KIRC or SKCM, as well as NEMO's varying success in identifying clinically enriched variables, further highlight their limitations. IntegrAO proficiently discerned both criteria, reflecting robust integration and patient stratification. This inconsistency among the other methods underscores the intricate challenge of integrating diverse partial multi-omics data, further emphasizing IntegrAO's value for translational applications requiring comprehensive patient characterization.

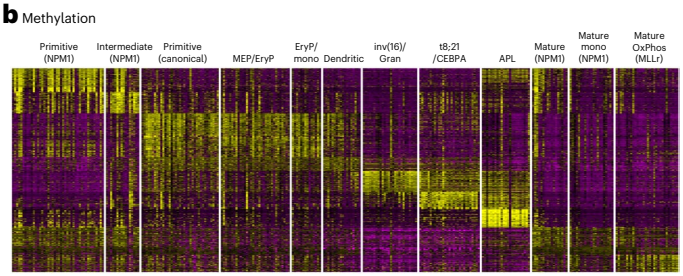
Robust new patient classification with incomplete omics data

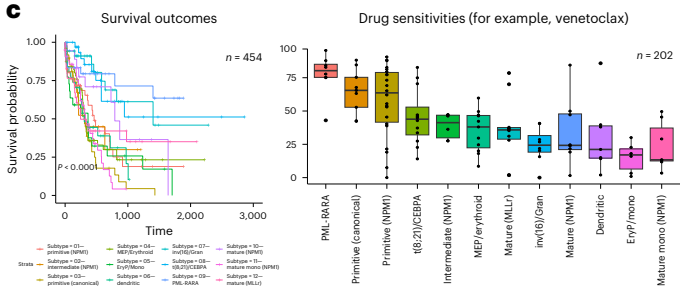
Classifying new patients into predefined subtypes is essential in clinical practice, especially when only partial omics data are available. IntegrAO addresses this challenge by enabling accurate classification using any available omics data.

To assess IntegrAO's classification ability, we designed an experiment mimicking real-world scenarios and compared it with five widely used classifiers: MLP, support vector machine, random forest, XGBoost and k-nearest neighbours (kNN). The dataset included

https://doi.org/10.1038/s42256-024-00942-3







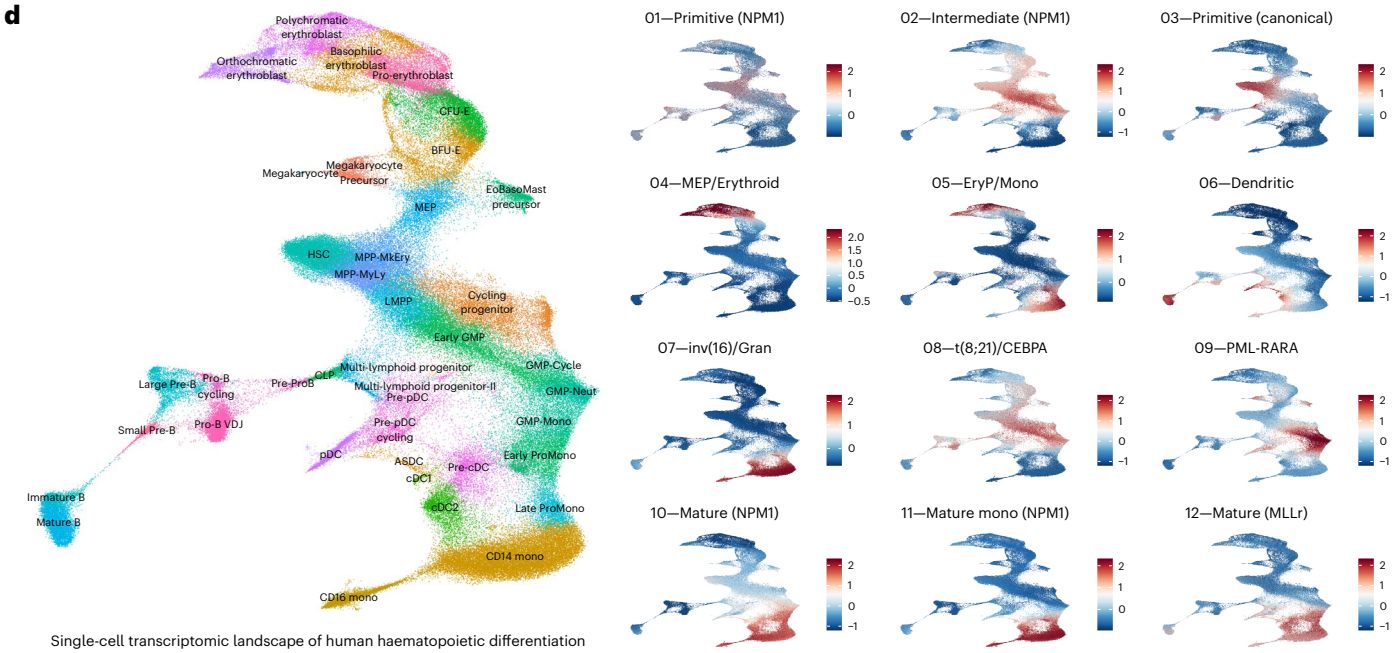


Fig. 3 | Multi-omics integrative analysis of AML elucidating intertumour heterogeneity. a, IntegrAO discerns 12 subtypes with distinct hierarchical composition, transcriptomic profiles and mutational patterns, preserving granular differentiations. **b**, The 12 subtypes identified by IntegrAO show unique DNA methylation profiles, highlighting the method's ability to differentiate based on epigenetic variations. **c**, Clinical relevance of IntegrAO subtypes is demonstrated through enhanced survival outcomes and drug sensitivities. The Kaplan–Meier plot reveals significant survival differences among subtypes (log-rank test P value = 1.21×10^{-7}, $n = 454$), confirming their distinct prognostic implications. Drug response to venetoclax (analysis of variance P value = 4.59×10^{-8}, $n = 202$) highlights distinct patterns across clusters. Each box shows the interquartile range of sensitivity, with the median as a horizontal line, individual dots as observations and whiskers extending to 1.5× the interquartile range. **d**, UMAP visualization of single-cell transcriptomic landscape of human haematopoietic differentiation[34], with haematopoietic lineage enrichment analysis validating the subtype differentiation and emphasizing the captured heterogeneity. HSC, hematopoietic stem cell; MPP-MkEry, multipotent progenitor megakaryocyte–erythroid; MPP-MyLy, multipotent progenitor myeloid–lymphoid; MEP, megakaryocyte–erythrocyte progenitor; BFU-E, burst-forming unit – erythroid; CFU-E, colony-forming unit – erythroid; EoBasoMast, eosinophil–basophil–mast cell; LMPP, lymphoid-primed multipotent progenitor; CLP, common lymphoid progenitor; pDC, plasmacytoid dendritic cell; GMP, granulocyte–monocyte progenitor; Neut, neutrophil; Mono, monocyte; ASDC, AXL+ SIGLEC6+ dendritic cell.

https://doi.org/10.1038/s42256-024-00942-3

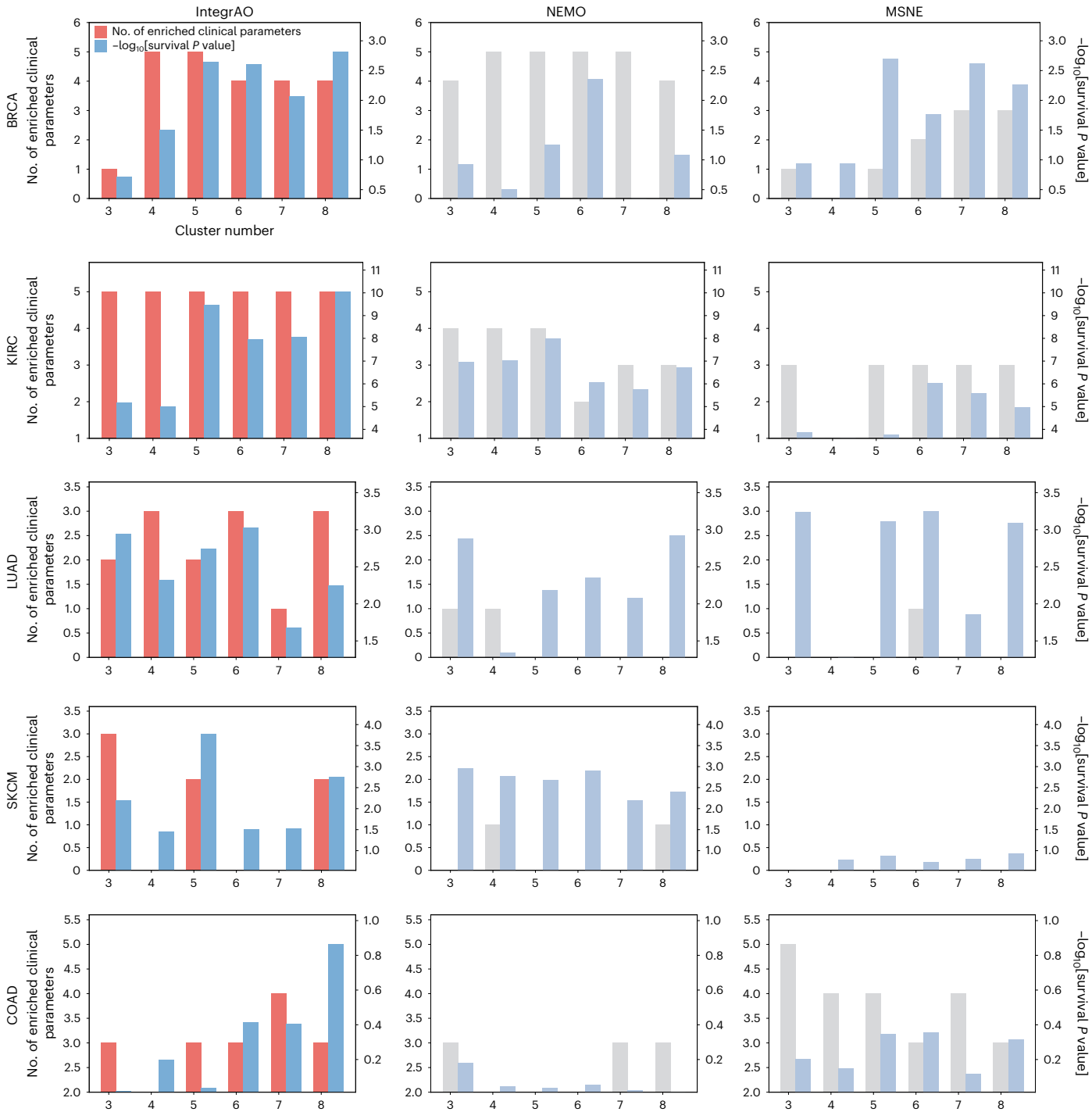


Fig. 4 | Performance comparison of IntegrAO, NEMO and MSNE in cancer cluster analysis across five cancer types. This series of plots demonstrates the differential survival between clusters and number of enriched clinical parameters within clusters across five cancer datasets with partial multi-omics data. The differential survival is quantified by $-\log_{10}$ of the *P* value from age-adjusted nested log-rank testing, where higher values indicate more pronounced survival differences. Similarly, clusters with a greater number of enriched clinical parameters indicate more meaningful clinical insights. Each plot evaluates the comparative performance of IntegrAO, NEMO and MSNE for varying cluster counts within a specific cancer type. Overall, IntegrAO more reliably identifies clusters with better survival differentiation and higher clinical enrichment than other methods.

miRNA, mRNA and DNA methylation profiles from the five TCGA cancer cohorts, focusing on patients with complete multi-omics data. IntegrAO integrated the full dataset to construct an integrated network, which was used to determine the optimal number of clusters and generate ground-truth cluster labels via spectral clustering (see 'Cluster number selection' in the Methods, Supplementary Table 4 and Supplementary Fig. 5). We performed stratified tenfold cross-validation, training the models on 90% of the samples and testing on the remaining 10%. Accuracy, F1-macro and F1-weighted metrics were used to evaluate multiclass prediction performance. For each dataset, IntegrAO first conducted unsupervised integration on the 90% training samples and then fine-tuned the model using the

https://doi.org/10.1038/s42256-024-00942-3

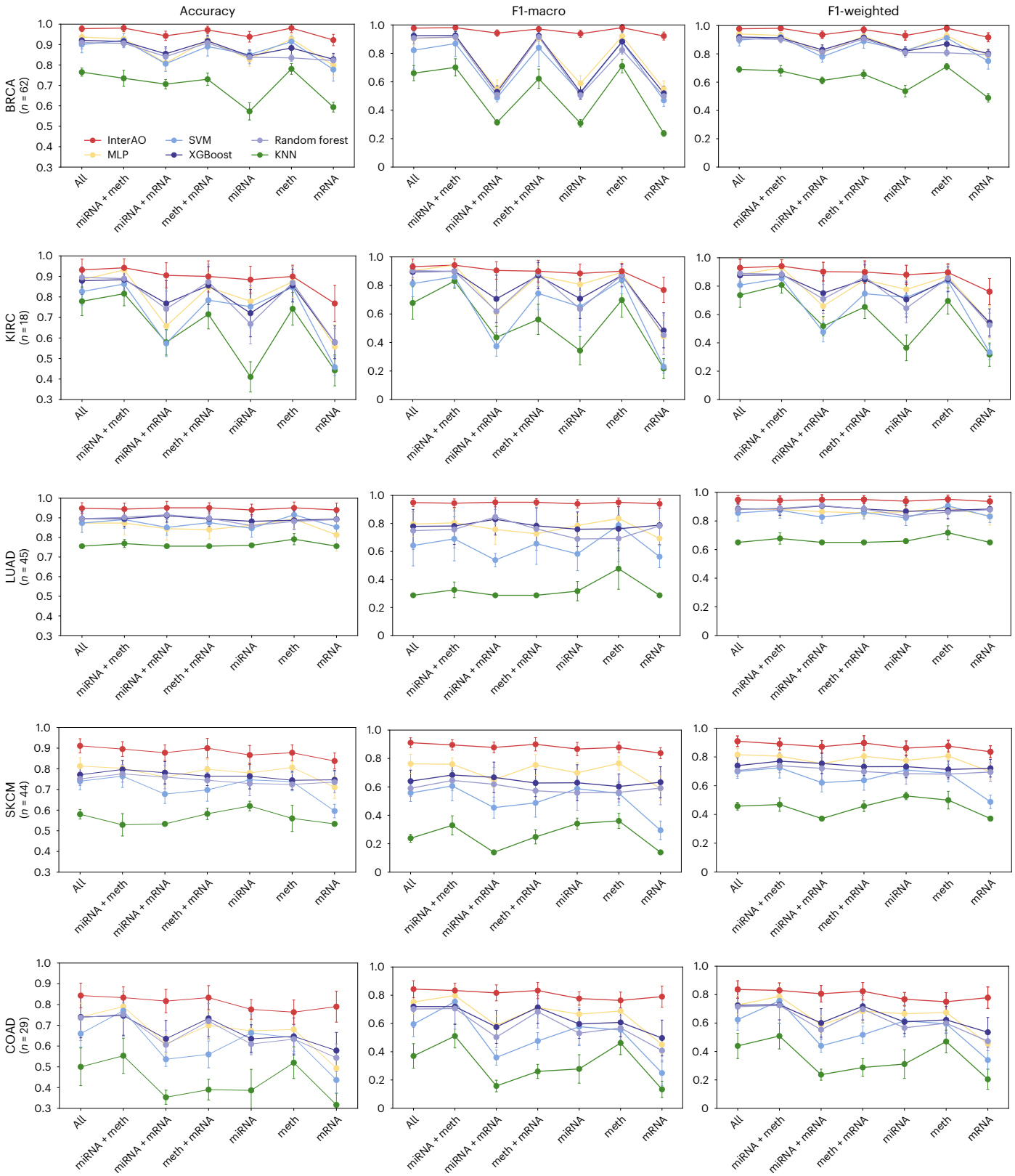


Fig. 5 | Performance comparison of new patient classification using IntegrAO versus MLP, support vector machine, XGBoost, random forest and *k*NN under different omics combinations. Accuracy, F1-macro and F1-weighted metrics were evaluated, with means and standard deviations from multiple experiments displayed (error bars denote ±one standard deviation). mRNA, meth and miRNA refer to single-omics classification using mRNA expression, DNA methylation and miRNA expression data, respectively. miRNA + meth, miRNA + mRNA and meth + mRNA indicate classification with the two corresponding omics, whereas 'all' used all the three data types. The number of biological samples is provided in the legend for each cancer type. Across all the metrics and inputs, IntegrAO substantially outperforms other methods, highlighting its ability to effectively leverage diverse omics for integrative patient classification.

ground-truth labels, and predicted the subtypes for the unseen test samples using any combination of omics data. By contrast, the other methods were trained on either single-omics data or direct concatenated multi-omics data from the 90% subset, and tested on the same 10% holdout set.

IntegrAO consistently outperformed all the other classifiers across every task, achieving superior results in accuracy, F1-macro and F1-weighted metrics (Fig. 5). IntegrAO demonstrated robust performance, while *k*NN was the least effective, and the remaining classifiers exhibited intermediate but noticeably lower accuracy. Further analysis revealed that IntegrAO's classification performance was highly robust across diverse omics combinations, whereas other methods displayed pronounced fluctuation and instability when missing certain data modalities, probably due to noisy or misleading omics data. We further tested IntegrAO's robustness with incomplete training data by varying the common sample ratio from 100% to 80%, 50% and 30%, and maintaining a consistent testing set across all the experiments for direct comparison against the top 2 performing methods, namely, MLP and XGBoost. The results (Supplementary Fig. 6) demonstrate IntegrAO's resilience: even with only 30% commonality, it can outperform MLP and XGBoost trained on fully complete datasets. This finding underscores IntegrAO's superior ability to leverage limited multi-omics data effectively.

Classifying new patients with incomplete data is challenging, but IntegrAO addresses this by embedding different omics features into a unified space, enabling accurate classification even with partial data. This feature holds critical clinical importance, as physicians frequently face the challenge of making diagnostic or treatment decisions with only partial omics information available. By bridging this gap, IntegrAO enhances the practical application of multi-omics in precision medicine, supporting better-informed clinical decisions.

Discussion

This study presents IntegrAO, an integrative framework designed to tackle key challenges in multi-omics analysis—handling incomplete heterogeneous data and projecting new samples using partial profiles. Simulated data tests demonstrate IntegrAO's resilience to noise at low data overlaps and effective integration at higher overlaps. In the AML case study, IntegrAO successfully combined the cell hierarchy composition, transcriptomics and DNA methylation, identifying 12 clinically and biologically distinct subtypes and illustrating AML's heterogeneity. Systematic evaluations across five cancer cohorts, encompassing six omics modalities, show IntegrAO's superiority in identifying distinct subtypes compared with other methods. Its consistent performance in projecting new samples, regardless of the number of available omics, highlights its potential in modality-agnostic inference and unified patient representation.

IntegrAO's adaptability to varied and incomplete datasets makes it a valuable tool for precision medicine. Its architecture maximizes the utility of diverse data types, addressing the inconsistency in clinical data availability. A core strength of IntegrAO lies in its dynamic scaling during graph fusion, which minimizes noise when a few shared samples are present and maximizes information flow with greater overlap. This ensures reliable signal capture from each modality, improving integration accuracy and robustness. IntegrAO's pairwise fusion strategy, which integrates data pairs rather than requiring a large common sample across all modalities, enhances its real-world applicability. During embedding extraction, the reconstruction loss preserves biological network structures, whereas the alignment loss aligns embeddings from different modalities. Together, these processes align patient data into a unified space, facilitating accurate subtype identification and reliable classification of new patients. These features establish IntegrAO as a transformative tool for creating comprehensive patient databases, advancing cancer understanding and seamlessly translating insights into clinical practice.

Future advancements could further enhance IntegrAO's scalability and applicability. Transforming the graph fusion process into an end-to-end neural network would improve computational efficiency, extending IntegrAO's utility beyond cancer to other diseases. Moreover, the integration of diverse and novel data types represents a substantial advancement opportunity. By developing sophisticated encoding techniques to incorporate data from histopathology images, clinical notes and real-time sensors into IntegrAO, the framework can achieve a more nuanced and comprehensive analysis of patient data. This holistic approach is expected to revolutionize personalized treatment plans and improve outcomes by leveraging the full spectrum of available medical data, thereby pushing the frontiers of precision medicine.

Despite advancements in IntegrAO, challenges remain for multi-omics research. A key issue is incomplete feature sets across omics platforms, which can lead to notable gaps in data analysis. Developing algorithms that can adapt to or reconstruct these missing components will be essential for maintaining robustness and accuracy in multi-omics integration. Data quality variability across devices calls for advanced normalization and calibration techniques to ensure reliable integration. Additionally, improving interpretability is crucial for clinical adoption. Incorporating advanced explainable artificial intelligence techniques directly into the data integration framework could enhance biomarker discovery processes, making the results more transparent and scientifically robust. Tackling these issues will not only refine the scientific methodologies but also enhance the practical applications of multi-omics research, ultimately leading to better patient care and health outcomes.

Methods

Data preprocessing

Simulated cancer omics datasets. We utilized the InterSim CRAN package[29] to simulate cancer omics datasets, generating a total of 500 samples distributed across 15 clusters of varying sizes, reflecting realistic clinical scenarios. For the hyperparameters, we set 'effect = 0.1' and 'p.DMP = 0.1', keeping the rest of the hyperparameters at their default values.

TCGA cancer datasets. For the cancer datasets, we leveraged multi-omics data across five tumour types from TCGA: BRCA, COAD, SKCM, KIRC and LUAD. Specifically, we obtained mRNA expression, DNA methylation, copy number variation and protein expression data directly from cBioPortal[35]. MicroRNA expression data were retrieved separately from the Broad Institute's Firehose source data. Relevant clinical information was also acquired for each patient. Before analysis, rigorous preprocessing was performed, including outlier removal, imputation of missing values via *k*NN and normalization by standard scaling to mean of 0 and standard deviation of 1. Patients with over 20% missing data for any data type and features with over 20% missing values across patients were excluded. We additionally selected the top 2,000 features exhibiting the greatest standard deviation from each data modality. For modalities with fewer than 2,000 total features, no feature filtering was performed.

AML cancer dataset. To construct an integrated AML dataset for heterogeneous analysis, we merged raw data from the TCGA, BEAT-AML and Leucegene cohorts. Gene expression data normalization was performed using a variance-stabilizing transformation for each each dataset. Batch effects were then corrected with the one-cell-at-a-time[36] algorithm, which also reduced the features to a 30-dimensional space. For cell composition, we used bulk gene expression deconvolution following another work[30], applying the one-cell-at-a-time algorithm for subsequent feature reduction. DNA methylation data, exclusive to the TCGA cohort, required no batch correction, and we selected 2,000 highly variable features based on dispersion. The final dataset

included 812 AML patients with cell hierarchy composition and mRNA expression data, and a subset of 308 patients with additional DNA methylation data.

Graph fusion in transductive integration

The first step of IntegrAO's transductive integration is the fusion of partially overlapping patient graphs (Fig. 1a, step 1). The subsequent section details the construction of these patient graphs and their partial overlap fusion. This graph fusion approach builds on our prior work, namely, similarity network fusion[7].

Patient graph construction. We first construct a patient graph for each type of omics analysis. Each graph can be represented as $G = (V, E)$, with vertices V correspond to the patients $\{x_1, x_2, \ldots, x_n\}$, each represented as a real-valued vector in R^d, and undirected weighted edges E denote the affinity between patients. It is important to note that the dimension d of these vectors can vary across different omics, reflecting the unique feature sets of each type of omics analysis. The weight of the edge is computed as

$$W(i,j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu\varepsilon_{i,j}}\right), \tag{1}$$

where $\rho(x_i, x_j)$ is the Euclidean distance between patients x_i and x_j, μ is a hyperparameter with the recommended setting in the range of [0.3, 0.8]. $\varepsilon_{i,j}$ is defined as

$$\varepsilon_{i,j} = \frac{1}{3} \times \left(\frac{1}{|N_i|}\sum_{k\in N_i}\rho(x_i, x_k) + \frac{1}{|N_j|}\sum_{l\in N_j}\rho(x_j, x_l) + \rho(x_i, x_j)\right), \tag{2}$$

where N_i is the set of x_i's neighbour including x_i in G. We then performed two operations on each graph to derive the transition probability matrix for the graph fusion stage. The first is normalizing the affinity matrix for numerical stability:

$$P(i,j) = \begin{cases} \dfrac{W(i,j)}{2\sum_{k\neq i}W(i,k)}, & i \neq j \\ 1/2, & i = j \end{cases} \tag{3}$$

The second is obtaining the local affinity matrix by considering only the K most similar patients per patient. This approach focuses on the strongest connections by selecting the top K connections based on the highest edge weights in the patient graph:

$$S(i,j) = \begin{cases} \dfrac{W(i,j)}{\sum_{k\in N_i}W(i,k)}, & j \in N_i \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Given v different data modalities, we can construct affinity matrices $W^{(m)}$ using equations (1) for the mth view, $m = 1, 2, \ldots, v$. $P^{(m)}$ and $S^{(m)}$ are obtained from equations (3) and (4), respectively.

Partial overlap graph fusion. After obtaining the patient graphs from each omics data modality, we can conduct the graph fusion step (Fig. 1a, step (1)). In the case of two modalities with partially overlapping patient sets, that is, $v = 2$, let a and b denote the total number of patients for each modality, respectively, and c be the number of common patients. Let C denote the set of common patients. The transition probability matrices $P^{(1)} \in \mathbb{R}^{a\times a}$ and $P^{(2)} \in \mathbb{R}^{b\times b}$ and local affinity matrices $S^{(1)} \in \mathbb{R}^{a\times a}$ and $S^{(2)} \in \mathbb{R}^{b\times b}$ are constructed as described previously. During fusion, each modality patient graph is initialized to its P matrix ($P^{(1)}_{t=0} = P^{(1)}$; $P^{(2)}_{t=0} = P^{(2)}$). The key concept for fusing such partially overlapped data is to leverage the common samples to propagate information across the graphs via graph fusion. IntegrAO iteratively updates the patient graph for each data modality as follows:

$$P^{(1)}_{t+1} = S^{(1)} \times P'^{(2\to1)}_t \times \left(S^{(1)}\right)^{\mathrm{T}}, \tag{5}$$

$$P^{(2)}_{t+1} = S^{(2)} \times P'^{(1\to2)}_t \times \left(S^{(2)}\right)^{\mathrm{T}}, \tag{6}$$

where the $P'^{(2\to1)}_t$ and $P'^{(1\to2)}_t$ are denoted as intermediate cross-modal transition matrices. They are constructed using affinity weights from common samples across the two modalities. Specifically, these matrices are populated by transitioning probabilities from one network's matrix to another but only for the indices corresponding to the common patients, as

$$W'^{(2\to1)}_{t\ (a\times a)}(i,j) = \begin{cases} P^{(2)}_t(i,j), & i,j \in C \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$

$$W'^{(1\to2)}_{t\ (b\times b)}(i,j) = \begin{cases} P^{(1)}_t(i,j), & i,j \in C \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

Then, we apply a novel scaling normalization:

$$P'_t(i,j) = \begin{cases} \dfrac{W'_t(i,j)}{2\sum_{k\neq i}W'_t(i,k)} \times \tau, & i \neq j \\ 1 - 1/2 \times \tau, & i = j \end{cases}, \tag{9}$$

$$\tau = \frac{c}{\text{number of samples in the to-fuse network}}.$$

The term 'to-fuse' refers to the network for which the cross-modal transition probabilities are being recalculated. For instance, when updating $P'^{(2\to1)}_t$, the 'to-fuse' is network 1, comprising a patients. Thus, the number of samples in the to-fuse network equals a. Conversely, when updating $P'^{(1\to2)}_t$, the 'to-fuse' is network 2, which consists of b patients, thereby the number of samples in the to-fuse network equals to b. This differentiation is essential as it affects the scaling normalization applied in the equation, ensuring that the transition probabilities are adjusted according to the specific size of each network and the proportion of common patients, thereby enabling accurate and effective information flow and fusion between the networks.

During iterative updates, each modality utilizes the shared patients' transition matrix from the other modality for fusion. The scaling normalization helps minimize the impact of the other modality when few patients are shared, and maximizing information flow when many patients are common. Not only the common patients' similarities can get updated through graph fusion but unique patients can also leverage the affinity information of the common patients from other modalities to learn more robust affinity for their own patient graph. This procedure updates the transition matrices each time, generating two parallel interchanging fusion processes. After each iteration, we performed normalization on $P^{(1)}_{t+1}$ and $P^{(2)}_{t+1}$ as in equation (3), for the following three reasons: (1) ensure a patient is always most similar to themselves than to other patients; (2) ensure the final graph is full rank; and (3) for quicker convergence of fusion. After t steps, we obtain the fused patient graph for each modality.

As our fusion approach leverages shared patients between modalities, the number of common patients may decrease when integrating more than two data types ($v > 2$). To address this, we perform pairwise fusion for multiple modalities:

$$P^{(m)} = \frac{\sum_{k\neq m}\left(S^{(m)} \times P'^{(k)} \times \left(S^{(m)}\right)^{\mathrm{T}}\right)}{v-1}, m = 1, 2, \ldots, v. \tag{10}$$

In this process, each modality pair is independently fused using equations (5) and (6), and the results are then averaged at the completion of each update cycle. Since the sample size differs across

modalities, the fused affinity matrices for each data type retain the original dimensionality. The subsequent step involves integrating these modal-specific graphs into a unified representation, which will be detailed in the following section.

Embedding extraction and alignment in transductive integration

The second step of IntegrAO's transductive integration is unsupervised extraction and alignment of patient embeddings across omics modalities (Fig. 1a, step (2)). This embedding step fulfils two critical goals: (1) deriving low-dimensional embeddings that maintain the affinity structure of the fused graphs for each data type and (2) aligning embeddings for the same patient across modalities.

Model architecture. The deep learning model in IntegrAO consists of two key components: (1) omics-specific graph encoders to extract patient embeddings within each data modality and (2) shared projection layers to map the embeddings from different omics into a common latent space. For each omics-specific GNN encoder (Fig. 1a, step (2)), inspired by GraphSAGE[37], instead of training individual embeddings for each node, we learn an aggregating function that generates embeddings by aggregating features from a node's local neighbourhood. This enables generating embeddings for unseen nodes using the learned functions given their local neighbourhood is provided.

Using the fused patient graphs, we obtain sparse affinity matrices per omics modality by considering only the K most strongly connected neighbours by the weight of each patient node, as defined in equation (4). The weighted graphs are converted into unweighted versions as inputs to the encoders. Formally, let $G = (V, E)$ denote the unweighted patient networks, where V is the node (patient), E is the edge connection (patient link) and L denotes the number of layers of the convolutional neural network encoders. The update rule for a node representation on the kth encoder layer is defined as

$$h_v^{(k)} = \sigma\left(W_1^{(k)} \times h_v^{(k-1)} + W_2^{(k)} \times \mathrm{MEAN}(\{h_u^{(k-1)} | u \in N(v)\})\right), \tag{11}$$

where $h_v^{(k)}$ is the representation of node v at the kth layer, $N(v)$ denotes the set of neighbours of node v and MEAN refers to the averaging operation. $W_1^{(k)}$ and $W_2^{(k)}$ are two learnable weight matrices. In particular, we use the original features from each type of omics analysis as the input to the first GNN layer. We set the number of layers to be 2 for each GNN encoder ($L = 2$). Last, the shared projection layers comprise stacked MLP layers that ingest the node representations from the final layer of each omics-specific encoder (Fig. 1a, step (2)). These MLP layers are designed to align these representations into a shared dimensionality, ultimately projecting them into a common latent space through training:

$$e_v = \mathrm{MLP}(h_v^{(L)}). \tag{12}$$

Learning objective. For better illustration, again consider the integration of two distinct data modalities: $\mathbf{X^{(1)}} \in \mathbb{R}^{n_1 \times d_1}$ and $\mathbf{X^{(2)}} \in \mathbb{R}^{n_2 \times d_2}$, where $\mathbf{X^{(1)}} \in \mathbb{R}^{n_1 \times d_1}$ represents the dataset from the first modality with n_1 samples and d_1 features, and $\mathbf{X^{(2)}} \in \mathbb{R}^{n_2 \times d_2}$ represents the dataset from the second modality with n_2 samples and d_2 features. In the IntegrAO embedding phase, our objective is to map these datasets into a unified embedding space of dimensionality q. The outputs from the shared MLP layers for the two modalities are represented as $\mathbf{X^{(1)'}}$ and $\mathbf{X^{(2)'}}$, where $\mathbf{X^{(1)'}} \in \mathbb{R}^{n_1 \times q}$ and $\mathbf{X^{(2)'}} \in \mathbb{R}^{n_2 \times q}$ are the lower-dimensional embeddings of the respective modalities after processing through the MLP. These embeddings are achieved by optimizing two distinct loss functions: the reconstruction loss L_{reconc} and the alignment loss L_{align}. The reconstruction loss L_{reconc} is conceptualized on the principles of t-distribution stochastic neighbour embedding[38] and can be formally defined as

$$L_{\mathrm{reconc}} = \mathrm{KL}(P^{(1)} || Q^{(1)}) + \mathrm{KL}(P^{(2)} || Q^{(2)}), \tag{13}$$

where $P^{(1)} \in \mathbb{R}^{n_1 \times n_1}$ and $P^{(2)} \in \mathbb{R}^{n_2 \times n_2}$ are the fused patient graphs obtained during the fusion stage with diagonal values set to 0, and $Q^{(1)} \in \mathbb{R}^{n_1 \times n_1}$ and $Q^{(2)} \in \mathbb{R}^{n_2 \times n_2}$, constrained to the t-distribution, are the sample-to-sample transition probability matrix calculated using the low-dimensional embedding $\mathbf{X^{(1)'}}$ and $\mathbf{X^{(2)'}}$. Specifically, the transition probabilities $Q_{ij}^{(m)}$ for modality m are computed as follows:

$$Q_{ij}^{(m)} = \frac{\left(1 + \| x_i^{(m)'} - x_j^{(m)'} \|^2\right)^{-1}}{\sum_{k \neq i}\left(1 + \| x_i^{(m)'} - x_k^{(m)'} \|^2\right)^{-1}}, \tag{14}$$

where $x_i^{(m)'}$ and $x_j^{(m)'}$ are the low-dimensional representations of samples i and j from modality m. Then, the Kullback–Leibler divergence is defined as

$$\mathrm{KL}(P||Q) = \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}}. \tag{15}$$

The alignment loss L_{align} is crucial for ensuring that embeddings of the same patients derived from different omics modalities are coherent and well aligned within a shared latent space. It quantifies the mean squared error between these embeddings, effectively minimizing the distance between them to foster a unified representation. This loss is formally defined as

$$L_{\mathrm{align}} = \frac{1}{c} \sum_{i=1}^{c} \mathbb{1}\,(i \in \mathbf{C}) \left(\mathbf{X}_i^{(1)'} - \mathbf{X}_i^{(2)'}\right)^2, \tag{16}$$

where C denotes the set of common samples between the two modalities. The final loss is the combination of reconstruction loss and alignment loss as

$$\mathrm{Loss} = L_{\mathrm{reconc}} + \beta \times L_{\mathrm{align}}, \tag{17}$$

where β is a trade-off parameter to balance the Kullback–Leibler terms and the embedding alignment term. We set $\beta = 1$ in all our experiments. The model can be readily extended to multiview data by adding additional Kullback–Leibler divergence terms to the reconstruction loss for each added view, and summing all the pairwise alignment losses between modalities for the matching loss. We solve the optimization problem using gradient descent with a fixed number of epochs. We set epoch = 1,000 in all our experiments.

Model output. After training, the final output is derived by averaging patient embeddings across modalities (Fig. 1a, step (3)). Let $M(i)$ denote the omics types available for patient i and then the final patient embeddings $E(i)$ for patient i are obtained by

$$E(i) = \frac{1}{|M(i)|} \sum_{m \in M(i)} e_i^{(m)}, \tag{18}$$

where $e_i^{(m)}$ is the embedding for patient i from modality m. The final integrated network is then computed using equation (1) followed by equation (3), taking the final patient embeddings as the input.

Inductive prediction

Model fine tuning for subtype prediction. After the unsupervised integration of multi-omics data, patient subtypes can be determined by clustering the final integrated network (Fig. 1b, step (1)). Given the defined subtype labels, IntegrAO can be further fine-tuned to predict the subtypes for new patients based on any combination of omics data. To enable this, we initialized with the unsupervised IntegrAO

model parameters and appended a prediction head that ingests the final patient embeddings to output a subtype prediction. We calculate the classification loss L_{clf} as

$$L_{\text{clf}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(\text{Pred}(E_i)) + (1-y_i)\log(1-\text{Pred}(E_i))\right), \quad (19)$$

where $\text{Pred}(\cdot)$ is the fully connected prediction head, and y_i is the defined subtype label. During fine tuning, we jointly optimize the total loss:

$$\text{Loss} = L_{\text{reconc}} + \beta \times L_{\text{align}} + \gamma \times L_{\text{clf}}. \quad (20)$$

The hyperparameter β and γ control the trade-off between the reconstruction loss, alignment loss and classification loss during optimization.

Subtype prediction for new patient. Fine-tuned IntegrAO can be used to predict the subtypes of new patient samples, accommodating any available omics modality (Fig. 1b, step (2)). During supervised fine tuning, for the omics data used in training, let $\{\mathbf{X}_{tr}^{(m)} | m = 1, 2, \ldots, v\}$ denote the input omics features for different modalities, and $\{\mathbf{P}_{tr}^{(m)} | m = 1, 2, \ldots, v\}$ denote the corresponding fused similarity matrix. The fine-tuned IntegrAO model can then be trained on $\{\mathbf{X}_{tr}^{(m)}\}$ and $\{\mathbf{P}_{tr}^{(m)}\}$, with training predictions represented as

$$\mathbf{Y}_{tr} = \text{IntegrAO}\left(\{\mathbf{X}_{tr}^{(m)}\}, \{\mathbf{P}_{tr}^{(m)}\}\right), m = 1, 2, \ldots, v, \quad (21)$$

where $\mathbf{Y}_{tr} \in \mathbb{R}^{n_{tr} \times c}$ contains the predicted subtype probabilities for each of the n_{tr} training samples, with c denoting the number of subtypes. For a new test sample $\{\mathbf{X}_{te}^{(m)} | m = 1, 2, \ldots, v\}$, to perform model inference, we extend the data matrix of the corresponding omics to $\{\mathbf{X}_{trte} = \begin{bmatrix} X_{tr} \\ X_{te} \end{bmatrix} | m = 1, 2, \ldots, v\}$, and generate the extended fusion matrix by performing the fusion step with the testing samples $\{\mathbf{P}_{trte}^{(m)} | m = 1, 2, \ldots, v\}$. Therefore, given $\{\mathbf{X}_{trte}\}$, $\{\mathbf{P}_{trte}\}$ and fine-tuned IntegrAO model, we have

$$\mathbf{Y}_{trte} = \text{IntegrAO}\left(\{\mathbf{X}_{trte}^{(m)}\}, \{\mathbf{P}_{trte}^{(m)}\}\right), m = 1, 2, \ldots, v, \quad (22)$$

where $\mathbf{Y}_{trte} \in \mathbb{R}^{n_{tr+1} \times c}$. The predicted subtype probability distribution for the testing sample is at the last row of $\mathbf{Y}_{trte}$.

Cluster number selection

To identify the optimal number of clusters for cancer datasets, we implemented a specific approach. First, after integrating the patient data, we conducted a tenfold train-test split. In each fold, a Gaussian mixture model was applied to 90% of the patient embeddings, and log-likelihood scores were calculated on the remaining 10%. This process was repeated for cluster numbers in a predefined range. We then computed the mean and standard deviation of the log-likelihood scores for each cluster number. The optimal cluster number was determined by the log-likelihood score, calculated by subtracting the mean from the standard deviation, and used to rank the suitability of each cluster number for the dataset.

In the new patient classification experiments, spectral clustering with this optimal cluster number was applied to the integrated network to obtain the clustering labels. For the AML case study, an initial identification of 18 clusters was refined by merging biologically similar clusters, resulting in 12 distinct AML subtypes.

Gene expression deconvolution

To generate the cell composition data for our cancer benchmarking experiments, we utilized BayesPrism[39] to deconvolute the raw gene expression counts from TCGA cancer cohorts. Our analyses were conducted exclusively through the the BayesPrism web portal, adhering to its default preprocessing steps. These steps included filtering outlier genes, selecting protein-coding genes and isolating signature genes for each cell type. For deconvolution job submissions, we used the portal's default settings. The resulting matrices—detailing fractions of patient-specific cell types—served as the cell composition modality for our integration benchmarking. The single-cell reference datasets utilized in the deconvolution process are detailed in Supplementary Table 5.

Implementation details

We used PyTorch[40] to implement the IntegrAO neural network model. We used snfpy 0.2.1 (ref. 7) to implement the patient graph construction and graph fusion. Other dependencies include Matplotlib 3.5.3, NetworkX 3.0, NumPy 1.24.1, pandas 1.3.5, scikit-learn 1.3.2, scipy 1.11.4, Seaborn 0.12.2, skunk 1.2.0, umap-learn 0.5.5 and torch-geometric 2.2.0.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets supporting the findings of this study are publicly available. The five TCGA cancer datasets, including the mRNA expression, DNA methylation, reverse-phase protein array and copy number variation data, were retrieved from cBioPortal (https://www.cbioportal.org/), and miRNA expression data were obtained from the Board GDAC Firehose (https://gdac.broadinstitute.org/). For AML datasets, the TCGA cohort was retrieved from cBioPortal, with the BEAT-AML dataset and Leucegene cohorts sourced from their respective original publications[32,33]. The single-cell reference datasets used to generate the cell composition data are reported in the 'Gene expression deconvolution' section in the Methods. The preprocessed data generated are available via Zenodo at https://doi.org/10.5281/zenodo.13989262 (ref. 41).

Code availability

The code to use IntegrAO is available via GitHub at https://github.com/bowang-lab/IntegrAO (ref. 42).

References

1. Shin, S. H., Bode, A. M. & Dong, Z. Precision medicine: the foundation of future cancer therapeutics. *NPJ Precis. Onc.* **1**, 12 (2017).
2. Steyaert, S. et al. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat. Mach. Intell.* **5**, 351–362 (2023).
3. Belizario, J. E. & Loggulo, A. F. Insights into breast cancer phenotying through molecular omics approaches and therapy response. *Cancer Drug Resist.* **2**, 527–538 (2019).
4. Lynch, H. T., Snyder, C. L., Shaw, T. G., Heinen, C. D. & Hitchins, M. P. Milestones of Lynch syndrome: 1895–2015. *Nat. Rev. Cancer* **15**, 181–194 (2015).
5. The Cancer Genome Atlas Research Network Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
6. Zhang, J. et al. International Cancer Genome Consortium Data portal—a one-stop shop for cancer genomics data. *Database* **2011**, bar026 (2011).
7. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
8. Rappoport, N. & Shamir, R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics* **35**, 3348–3356 (2019).
9. Nguyen, H., Shrestha, S., Draghici, S. & Nguyen, T. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics* **35**, 2843–2846 (2019).

10. Shen, R. et al. Integrative subtype discovery in glioblastoma using iCluster. *PLoS ONE* **7**, 35236 (2012).
11. Yang, Z. & Michailidis, G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32**, 1–8 (2016).
12. Vaske, C. J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, 237–245 (2010).
13. Wu, D., Wang, D., Zhang, M. Q. & Gu, J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genom.* **16**, 1022 (2015).
14. Lee, C. & van der Schaar, M. A variational information bottleneck approach to multi-omics data integration. In *International Conference on Artificial Intelligence and Statistics* 1513–1521 (PMLR, 2021).
15. Chen, L., Xu, J. & Li, S. C. DeepMF: deciphering the latent patterns in omics profiles with a deep learning method. *BMC Boinformatics* **20**, 648 (2019).
16. de Vega, W. C., Erdman, L., Vernon, S. D., Goldenberg, A. & McGowan, P. O. Integration of DNA methylation & health scores identifies subtypes in myalgic encephalomyelitis/chronic fatigue syndrome. *Epigenomics* **10**, 539–557 (2018).
17. Stefanik, L. et al. Brain-behavior participant similarity networks among youth and emerging adults with schizophrenia spectrum, autism spectrum, or bipolar disorder and matched controls. *Neuropsychopharmacology* **43**, 1180–1188 (2018).
18. Hamamoto, R., Komatsu, M., Takasawa, K., Asada, K. & Kaneko, S. Epigenetics analysis and integrated analysis of multiomics data, including epigenetic data, using artificial intelligence in the era of precision medicine. *Biomolecules* **10**, 62 (2019).
19. Martin, K. R. et al. The genomic landscape of tuberous sclerosis complex. *Nat. Commun.* **8**, 15816 (2017).
20. Little, R. J. & Rubin, D. B. *Statistical Analysis with Missing Data* Vol. 793 (John Wiley & Sons, 2019).
21. Henry, A. J., Hevelone, N. D., Lipsitz, S. & Nguyen, L. L. Comparative methods for handling missing data in large databases. *J. Vasc. Surg.* **58**, 1353–1359 (2013).
22. Flores, J. E. et al. Missing data in multi-omics integration: recent advances through artificial intelligence. *Front. Artif. Intell.* **6**, 1098308 (2023).
23. Fang, Z. et al. Bayesian integrative model for multi-omics data with missingness. *Bioinformatics* **34**, 3801–3808 (2018).
24. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
25. Lock, E. F., Park, J. Y. & Hoadley, K. A. Bidimensional linked matrix factorization for pan-omics pan-cancer analysis. *Ann. Appl. Stat.* **16**, 193 (2022).
26. Xu, H., Gao, L., Huang, M. & Duan, R. A network embedding based method for partial multi-omics integration in cancer subtyping. *Methods* **192**, 67–76 (2021).
27. Rappoport, N., Safra, R. & Shamir, R. MONET: multi-omic module discovery by omic selection. *PLoS Comput. Biol.* **16**, 1008182 (2020).
28. Hornung, R., Ludwigs, F., Hagenberg, J. & Boulesteix, A.-L. Prediction approaches for partly missing multi-omics covariate data: a literature review and an empirical comparison study. *WIREs Comput. Stat.* **16**, e1626 (2023).
29. Chalise, P., Raghavan, R. & Fridley, B. L. InterSIM: simulation tool for multiple integrative 'omic datasets'. *Comput. Methods Programs Biomed.* **128**, 69–74 (2016).
30. Zeng, A. G. et al. A cellular hierarchy framework for understanding heterogeneity and predicting drug response in acute myeloid leukemia. *Nat. Med.* **28**, 1212–1223 (2022).
31. The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
32. Tyner, J. W. et al. Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).
33. Marquis, M. et al. High expression of HMGA2 independently predicts poor clinical outcomes in acute myeloid leukemia. *Blood Cancer J.* **8**, 68 (2018).
34. Zeng, A. G. et al. Single-cell transcriptional mapping reveals genetic and non-genetic determinants of aberrant differentiation in AML. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.26.573390 (2024).
35. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
36. Wang, C. X., Zhang, L. & Wang, B. One Cell At a Time (OCAT): a unified framework to integrate and analyze single-cell RNA-seq data. *Genome Biol.* **23**, 102 (2022).
37. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst.* **30**, 1024–1034 (2017).
38. van der Maaten, L. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
39. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer* **3**, 505–517 (2022).
40. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
41. Ma, S. Dataset for 'Moving Towards Genome-wide Data Integration for Patient Stratification with Integrate Any Omics'. *Zenodo* https://doi.org/10.5281/zenodo.13989262 (2024).
42. Ma, S. bowang-lab/IntegrAO: IntegrAO 0.1.0. *Zenodo* https://doi.org/10.5281/zenodo.13760679 (2024).

Acknowledgements

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada, through the Canadian Institute for Advanced Research (CIFAR), and companies sponsoring the Vector Institute. This work was supported by funding from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-06189 and DGECR-2020-00294; B.W.), the CIFAR AI Chairs Program (B.W.) and the Peter Munk Cardiac Centre AI Fund at the University Health Network (B.W.). We acknowledge C. Wang, A. Young, V. Subasri, K. McKeen, J. Ma and V. Chu for feedback on the paper.

Author contributions

S.M. and B.W. conceived the project. S.M. contributed to the design and implementation of the algorithm and ran the experiments. A.G.X.Z. contributed to the analysis of computational experiments. S.M. and B.W. drafted the initial version of the paper. S.M., A.G.X.Z., B.H.-K., A.G., J.E.D. and B.W. contributed to the revision of the work. B.H.-K. and A.G. contributed to the design of the algorithm. J.E.D. and B.W. contributed to the conception and design of the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00942-3.

Correspondence and requests for materials should be addressed to Bo Wang.

Peer review information *Nature Machine Intelligence* thanks Kai Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Bo Wang

Last updated by author(s): Oct 25, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	snfpy 0.2.1, matplotlib 3.5.3, networkx 3.0, numpy 1.24.1, pandas 1.3.5, scikit-learn 1.3.2, scipy 1.11.4, seaborn 0.12.2, skunk 1.2.0, umap-learn 0.5.5, torch-geometric 2.2.0, torch 2.1.0, python 3.10, InterSIM 2.2.0; Github link for IntegrAO: https://github.com/bowang-lab/IntegrAO

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets supporting the findings of this study are publicly available. The five TCGA cancer datasets, including mRNA expression, DNA methylation, reverse-phase protein array, and copy-number variation data, were retrieved from cBioPortal (https://www.cbioportal.org/), while miRNA expression data was obtained from Board GDAC Firehose (https://gdac.broadinstitute.org/). For AML datasets, the TCGA cohort was retrieved from cBioPortal, with the BEAT-AML dataset and

Leucegene cohorts sourced from their respective original publications (https://doi.org/10.1038/s41586-018-0623-z, https://doi.org/10.1038/s41408-018-0103-6). The preprocessed data used to generated the figures are available on Zenodo at https://doi.org/10.5281/zenodo.13989262.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Power analysis was not performed. Experiments were generally conducted using all the samples available publicly.
Data exclusions	No data were excluded
Replication	For the simulation study, each experiment was repeated 10 times. For the AML case study and pan-cancer evalutation, each experiment was conducted 3 times. For the new patient classification study, 10-fold cross validation was performed for each experiment setup. All attempts at replication were successful.
Randomization	Uniform random subsampling was performed to select subset of samples for experiments.
Blinding	All the experiments in this study, including the simulation study, the AML case study, pan-cancer evalutation and new patient classification experiments were all blinded.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A