nature communications

Article

https://doi.org/10.1038/s41467-025-60822-1

GAUDI: interpretable multi-omics integration with UMAP embeddings and density-based clustering

Received: 9 October 2024

Accepted: 3 June 2025

Published online: 01 July 2025

Check for updates

Pol Castellano-Escuder[1], Derek K. Zachman[1,2], Kevin Han[1] & Matthey D. Hirschey[1,3,4,5]

Integrating high-dimensional cellular multi-omics data is crucial for understanding various layers of biological control. Single ‘omic methods provide important insights, but often fall short in handling the complex relationships between genes, proteins, metabolites and beyond. Here, we present a novel, non-linear, and unsupervised method called GAUDI (Group Aggregation via UMAP Data Integration) that leverages independent UMAP embeddings for the concurrent analysis of multiple data types. GAUDI uncovers non-linear relationships among different omics data better than several state-of-the-art methods. This approach not only clusters samples by their multi-omic profiles but also identifies latent factors across each omics dataset, thereby enabling interpretation of the underlying features contributing to each cluster. Consequently, GAUDI facilitates more intuitive, interpretable visualizations to identify novel insights and potential biomarkers from a wide range of experimental designs.

Multi-omic analyses integrate diverse data types such as genomics, proteomics, and metabolomics. Combining multiple omics modalities has the potential to uncover novel insights and biomarkers more than when each data type is analyzed alone[1,2]. The growth in high-throughput technologies has precipitated an exponential increase in omics data, underscoring the urgent need for new integration methods.

Traditional approaches to multi-omics integration have primarily focused on dimension reduction techniques. For example, methods based on Canonical Correlation Analysis (CCA) are used in RGCCA[3], while Co-Inertia Analysis is used in MCIA[4]. Similarly, Bayesian Factor Analysis underpins methods such as MOFA+[5], Negative Matrix Factorization is central to intNMF[6], Principal Components Analysis to JIVE[7], and Independent Components Analysis forms the basis of tICA[8]. Although these methods have been applied across various omics datasets and biological contexts, their effectiveness and limitations vary, highlighting the need for careful consideration in their application[9].

A central limitation shared by these methods is their reliance on linear assumptions. While appropriate in some cases, this assumption can be inadequate for accurately capturing the complex, often non-linear interplay among different omics layers[10,11]. Moreover, their computational intensity poses challenges, particularly for large-scale datasets. In response to these challenges, recent advancements have shifted towards non-linear integration approaches[9,10].

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can reveal the underlying structure in complex datasets[12]. By combining manifold learning with topological data analysis, it effectively visualizes high-dimensional data in lower-dimensional spaces. UMAP stands out from other methods like PCA and t-SNE by efficiently preserving global data structures, making UMAP particularly effective in visualizing large datasets and detecting patterns that might remain hidden using other techniques[11]. Recent studies have also explored UMAP for multi-omics integration in specific contexts such as cancer classification[13], though

[1]Duke Molecular Physiology Institute, Duke University School of Medicine, Durham, NC, USA. [2]Duke Department of Pediatrics, Division of Hematology-Oncology, Duke University School of Medicine, Durham, NC, USA. [3]Department of Medicine, Duke University School of Medicine, Durham, NC, USA. [4]Department of Pharmacology & Cancer Biology, Duke University School of Medicine, Durham, NC, USA. [5]Cardiovascular and Metabolic Disorders Program, Duke-National University of Singapore (NUS) Medical School, Singapore, Singapore. e-mail: matthew.hirschey@duke.edu

these applications have been limited in scope due to constraints such as RGB-based visualization requirements restricting integration to three omics layers, dependence on gene similarity networks, and evaluation primarily focused on cancer-type classification tasks.

One of its key strengths is handling non-linear data relationships, which is crucial in multi-omics, where data types often interact in complex, non-linear ways[11]. Thus, UMAP can greatly improve the interpretation and understanding of multi-omic interactions.

Despite UMAP's strengths in addressing some of these challenges, gaps remain that require further development. For example, integrating disparate feature spaces from different omics layers can complicate the analysis and interpretation of results[14]. Additionally, UMAP is often sensitive to parameters and pre-processing choices, leading to potentially misleading outputs that may not accurately reflect biology[15]. The need for more robust methods that enhance UMAP's adaptability across various data modalities and settings remains crucial, particularly for addressing the complexities of biological systems[15]. Therefore, we set out to develop a model that leverages UMAP's inherent advantages, to address the specific needs of multi-omics integration, and to benchmark its performance against six prominent joint dimension reduction multi-omics integration methods using a variety of real-world and simulated datasets.

Results

Group Aggregation via UMAP Data Integration (GAUDI) is a method developed for the unsupervised integration of multi-omics data, capitalizing on the strengths of Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction (Fig. 1). GAUDI begins by applying UMAP independently to each omics dataset, preserving the unique characteristics of each data type.

After processing each dataset, GAUDI concatenates the individual UMAP embeddings into a unified dataset (see Methods) and then applies a second UMAP to this concatenated dataset. This step combines the distinct omics layers into a single, lower-dimensional representation, allowing for the identification of underlying biological patterns that may not be evident in higher dimensions.

GAUDI then employs Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) for clustering[16]. Given the non-linear nature of the latent space, HDBSCAN is effective as it handles clusters of varying densities and irregular shapes without assuming a predefined number of clusters. It identifies clusters based on data density and is robust against noise and outliers, making it suitable for multi-omics datasets.

Finally, GAUDI computes metagenes using a XGBoost model to synthesize molecular features[17]. This involves using XGBoost to predict UMAP embedding coordinates from molecular features (e.g., gene

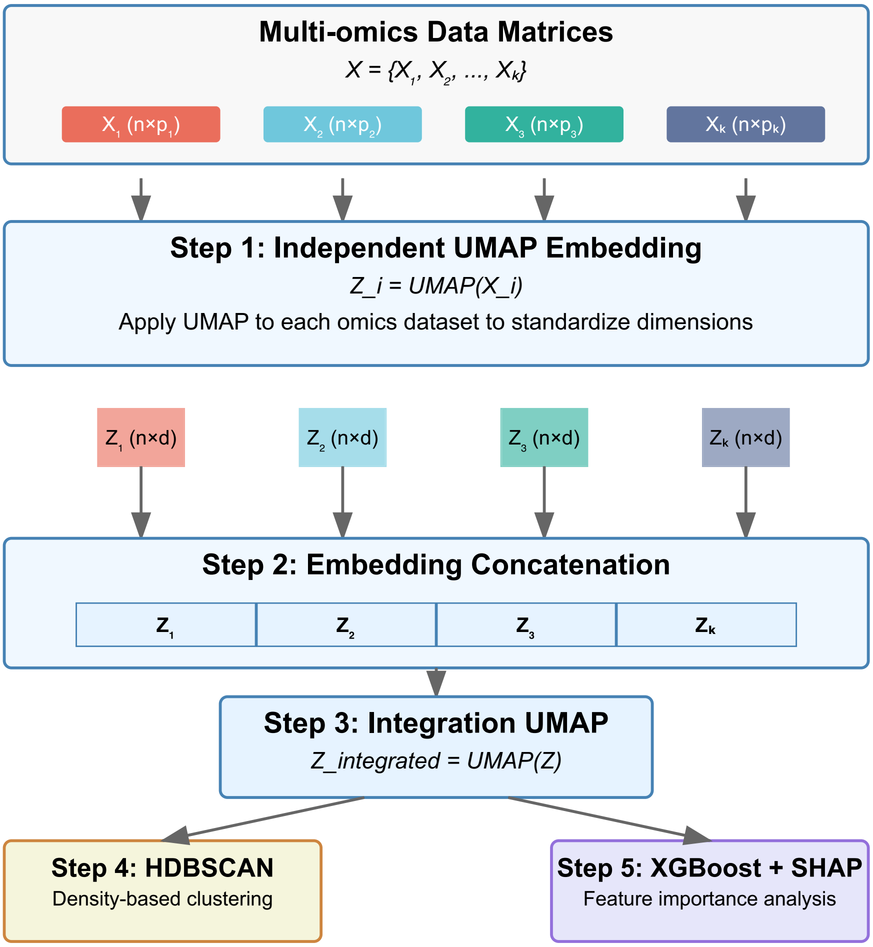

Fig. 1 | GAUDI method overview and workflow. The GAUDI (Group Aggregation via UMAP Data Integration) method integrates multi-omics data through a non-linear unsupervised approach. The workflow consists of five steps: (1) Independent UMAP embedding is applied to each omics dataset, transforming matrices of different dimensionalities ($n×p_1$, $n×p_2$, etc.) into standardized representations of equal dimensions (n×d); (2) These embeddings are concatenated to create a combined data matrix; (3) A second UMAP transformation is applied to the concatenated matrix to create an integrated representation; (4) Hierarchical Density-Based Spatial Clustering (HDBSCAN) is used to identify sample groups with similar multi-omic profiles; and (5) XGBoost combined with SHAP values is used to calculate feature importance, enabling biological interpretation of the factors driving sample clustering. This approach mitigates bias from high-dimensional omics types, captures non-linear relationships between variables, and provides interpretable results that can reveal novel biological insights across diverse experimental designs.

expressions) and identifying key features that influence the positioning of samples in the integrated latent space. Feature importance scores are then extracted using SHapley Additive exPlanations (SHAP) values to determine each feature's contribution to the embeddings[18].

GAUDI computes metagenes for both individual omics datasets and the combined omics, integrating feature importance scores from all datasets to provide a comprehensive view of molecular features across different biological layers.

Multi-omics integration methods comparison on artificial datasets

Next, we assessed GAUDI's performance. In the initial phase of our validation and comparative analysis, we generated three artificial omics datasets using the *InterSIM* R package[19], following the approach described by Cantini et al. These datasets were designed with predefined reference clusters, enabling a direct comparison between the clusters generated by each integration method and the dataset's intrinsic 'real' clusters. We simulated these datasets with varying cluster complexities, creating sets with five, ten, and fifteen clusters. Each cluster set was further diversified by simulating two distributions: a homogeneous distribution ('_EQ') and a heterogeneous distribution ('_HET'), where the latter featured a varied number of samples per cluster.

Using these synthetic datasets, we compared GAUDI to six other leading multi-omics integration methods[9]. Each method decomposed the omic matrices into factors corresponding to the artificially generated cluster counts—5, 10, and 15. Notably, intNMF and GAUDI are unique in providing default clustering outputs. For the other five methods, we derived clusters by implementing k-means consensus clustering on the factor matrices[9].

We quantified the alignment between the clusters obtained from these integration methods and the ground-truth clusters using the Jaccard index (JI), with a score of 1 indicating perfect agreement and 0 indicating complete discrepancy. Generally, all tested algorithms demonstrated strong performance across different scenarios, with most JIs exceeding 0.6 (Fig. 2). intNMF, primarily designed for clustering, consistently outperformed the other five methods in almost all cases. However, GAUDI distinguished itself by achieving a JI of 1 in every scenario, irrespective of cluster count or sample distribution heterogeneity.

Notably, intNMF displayed heightened variability in its performance as the number of clusters expanded and with the increase in sample size (using $N = 500$ in our study compared to $N = 100$ in Cantini et al.'s benchmark). In contrast, GAUDI showed remarkable consistency in clustering accuracy, unaffected by changes in cluster numbers or sample size variations (Fig. 2).

These findings position GAUDI as a superior method for clustering purposes, demonstrating high accuracy and generating more condensed and differentiated clusters across all tested scenarios (Supplementary Fig. 1).

Multi-omics integration methods comparison on TCGA cancer datasets

To determine the efficacy of GAUDI to cluster multi-omics from real-world data, we used TCGA multi-omic data from eight different cancer types[20]. We included gene expression, DNA methylation, and miRNA expression data, with sample sizes ranging from 170 for acute myeloid leukemia (AML) to 621 for breast cancer.

Following the benchmarking strategy outlined by Cantini et al., we decomposed the omic matrices into five latent factors and applied a Cox proportional hazards regression model to evaluate the association of these factors with survival. Consistent with previous findings, the number of factors linked to survival varied more with cancer type than with the integration method used (Fig. 3A), suggesting that intrinsic biological differences across cancer types significantly influence survival outcomes[9]. This underscores the importance of context-specific analysis when interpreting multi-omics data, revealing that factors such as tumor microenvironment, genetic heterogeneity, and molecular pathways play crucial roles in determining survival.

To extend our analysis beyond survival to clinical annotations, we tested for significant associations for the clinical annotations "age", "days to new tumor", "gender", and "neo-adjuvant therapy administration" with the factors generated for each cancer type[9]. The average selectivity value across all cancers and methods was 0.53 (see Methods). Methods excelling at producing clinically interpretable latent factors in each cancer type were those with the highest number of factors associated with clinical annotations and selectivity values above this average (Fig. 3B).

Each method exhibited similar performance in terms of the number of generated latent factors associated with clinical annotations and selectivity scores across all cancer types, suggesting that while they identified similar numbers of associations, the relevance and impact of these associations differed depending on the specific cancer type and the underlying biological context.

To enhance the interpretability of these results and further compare the clinical applicability of each method, we examined the relationship between the clusters generated by each method and patient survival. We identified clusters with the highest and lowest median survival in each cancer type and compared their survival curves. This approach aimed to maximize survival differences based on distinct multi-omic profiles, thereby identifying molecular features predictive of extreme survival groups.

While the association between factors and survival generally depended more on cancer type than on the method used, a similar pattern emerged in the relationship between clusters and survival. Notably, none of the methods revealed significant differences in breast, colon, and ovarian cancers, but they showed comparable results for the other five cancer types (Supplementary Fig. 3).

However, GAUDI stood out for its ability to detect sample groups with multi-omic profiles linked to markedly lower overall survival. This was most pronounced in acute myeloid leukemia (AML), where GAUDI identified the greatest median survival difference between the two most distant clusters (718 days, $p = 1.31\text{e-}08$), pinpointing a small high-risk group with a median survival of only 89 days—a threshold not reached by other methods (Fig. 3C). For AML, intNMF produced clusters with identical survival medians and classified varying multi-omic profiles into similar survival groups (Fig. 3C), indicating limited sensitivity at the extremes.

Using human multi-omic data, GAUDI stood out from other methods by identifying profiles associated with significantly lower overall survival times, especially in AML. These analyses demonstrate GAUDI's enhanced sensitivity in detecting critical survival differences, which has the potential to offer valuable insights for precision medicine. In some cancers like AML, GAUDI was uniquely effective in isolating high-risk groups, an important capability that is urgently needed.

Multi-omics integration methods comparison on single-cell datasets

Single-cell analysis has been essential for studying the biology and behavior of single cells from cell populations. As single-cell analytic techniques improve, including the ability to apply next-generation sequencing to individual cells, we have learned that cell populations previously thought to be homogeneous display considerable heterogeneity[21,22]. Thus, we tested the ability of GAUDI to detect multi-omic heterogeneity within scRNA-seq and scATAC-seq datasets. We obtained datasets containing gene expression and chromatin accessibility information across three cancer cell lines (HCT, Hela, and K562), covering 206 cells[9,23]. For this analysis, we reduced the omic

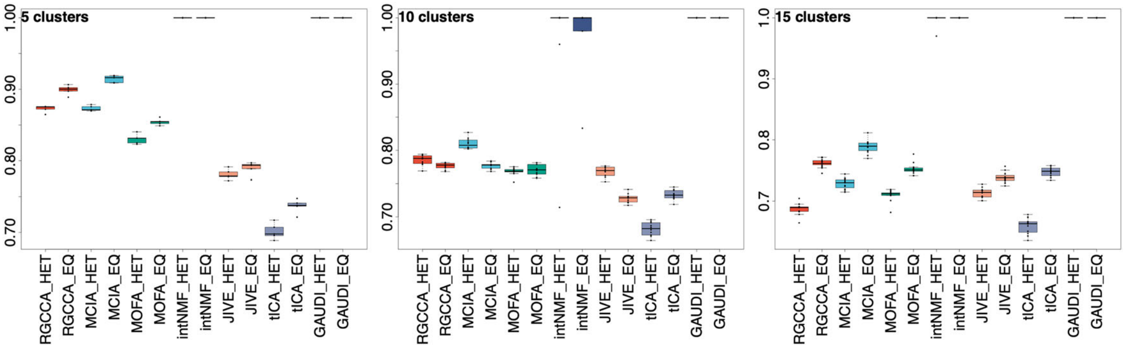

Fig. 2 | Clustering Accuracy on Simulated Multi-Omics Datasets. This figure displays boxplots illustrating the Jaccard Index (JI) scores for clusters identified by various multi-omics integration methods compared to the ground-truth clusters in the simulated data (in these boxplots, the center line represents the median, box boundaries show the interquartile range (25th–75th percentiles), whiskers extend to points within 1.5 × IQR, and points beyond the whiskers indicate outliers). The analysis encompasses scenarios with 5, 10, and 15 pre-defined clusters. For each method, we present results for both heterogeneous (HET) and homogeneous (EQ) cluster distributions. The analysis is based on datasets comprising 500 samples, and the depicted results are aggregated over 1000 independent iterations of k-means clustering, ensuring robust and reliable performance evaluation. Source data are provided as a Source Data file.

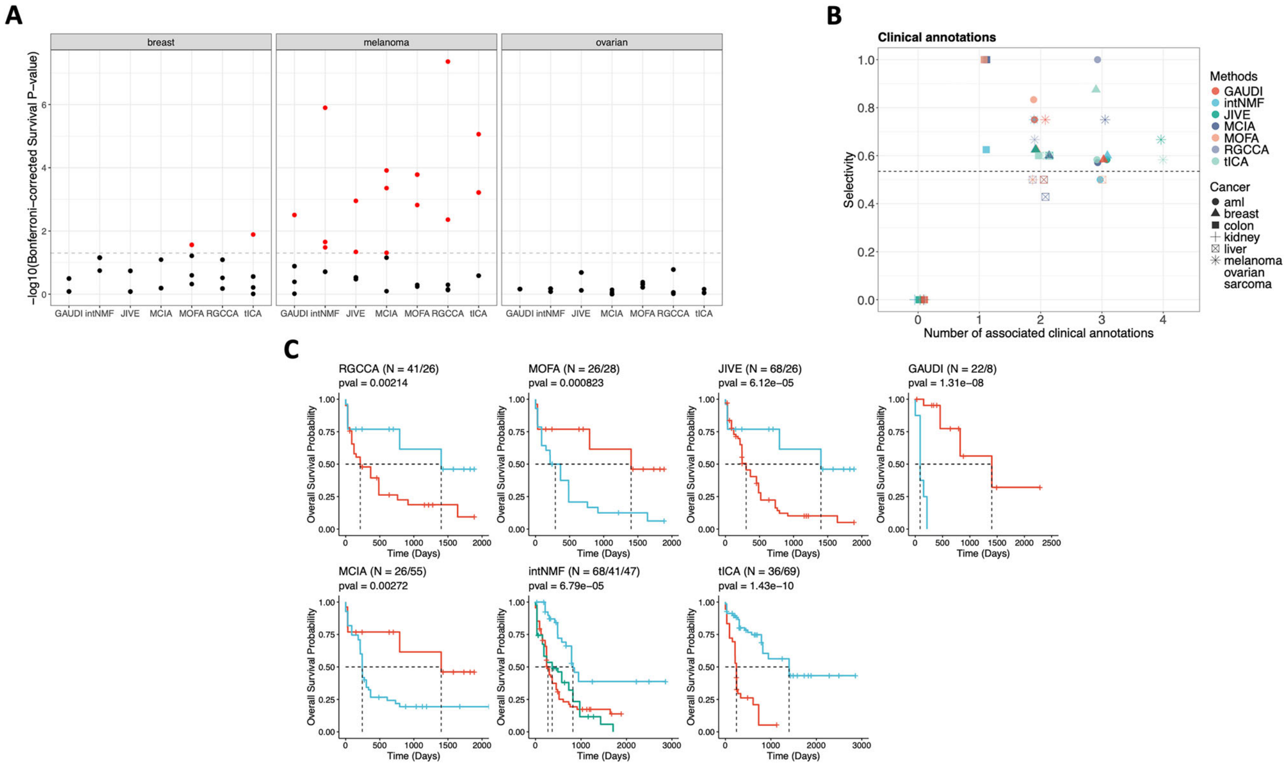

Fig. 3 | Comprehensive analysis of survival prediction and clinical associations using TCGA multi-omics data. A Bonferroni-corrected *p*-values for each of the 5 factors identified by multi-omics integration methods as predictive of survival, based on a Cox regression analysis. Dotted lines indicate the threshold for a corrected *p*-value of 0.05. Results for the remaining five cancer types are available in Supplementary Fig. 2. **B** Number of factors associated with clinical annotations in each method and the selectivity score of the factor-annotation associations. The dashed horizontal line represents the mean selectivity score. **C** Kaplan–Meier survival curves for AML clusters with the highest and lowest median survival times identified by each method. The log-rank test is used for *p*-value calculations, and the sample size for each cluster is provided (*N*). Source data are provided as a Source Data file.

datasets to two dimensions, aiming for precise classification of each cell according to its cell line.

We performed a comparative analysis including seven of the multi-omics integration methods, selected based on their C-index performance (see Methods). All methods displayed a general ability to distinguish between the three cell line types; however, MOFA+ and JIVE faced challenges in creating a clear classification in two dimensions. While the rest varied in the dispersion of their clusters, GAUDI had a misclassification rate of 1.46% (misclassifying 3 out of 206 cells), generating the most compact clusters and achieving a clear cell line identification (Fig. 4A).

Beyond accurate cell line classification, GAUDI uniquely identified and categorized distinct multi-omic profiles within the same cell line (K562) into separate clusters (Fig. 4B). We performed a differential expression analysis comparing clusters 2 and 3 within the K562 cell line, uncovering 190 differentially expressed genes with an FDR < 0.05

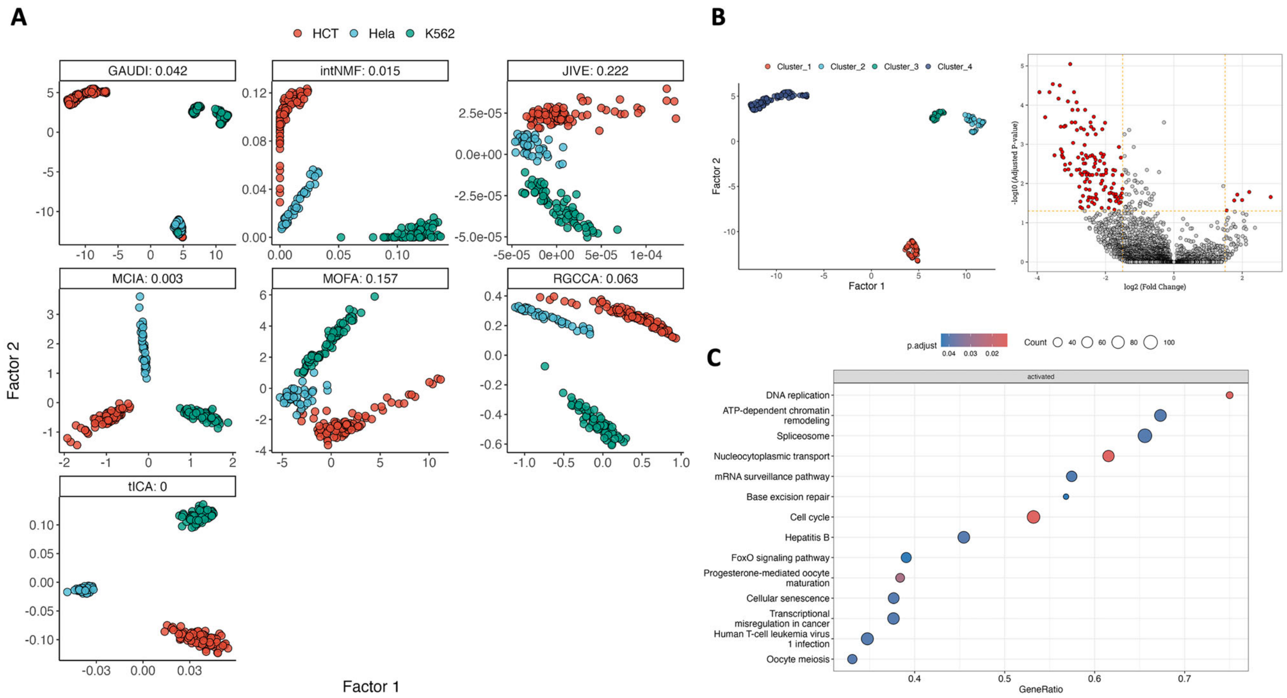

Fig. 4 | Multi-omics integration and sub-structure analysis in single-cell cancer data. A Scatter plots of the two factors for each integration method (along with its C-index), color-coded by the origin of cancer cell lines. **B** Density-based clusters by GAUDI reveal underlying sub-structures within the K562 cell line. Adjacent is a volcano plot detailing differentially expressed genes between clusters 2 and 3, highlighting genes with significant differential fold changes and FDR values. **C** GSEA results using KEGG pathway annotations demonstrate the biological pathways differentially enriched between GAUDI-derived clusters 2 and 3 in the K562 cell line, providing insights into the functional implications of the identified substructures. Source data are provided as a Source Data file.

(Fig. 4B)[24]. Subsequent Gene Set Enrichment Analysis (GSEA) using KEGG pathways highlighted significant variations in DNA replication, ATP-dependent chromatin remodeling, spliceosome, and cell cycle processes between the sub-clusters (Fig. 4C), mirroring the complexity of the integrated datasets[25].

These findings underscore GAUDI's effectiveness as a clustering tool, highlighting its ability to dissect both bulk and single-cell multi-omics datasets. GAUDI's capacity to detect and differentiate biologically significant substructures within these datasets sets it apart from other methods that fail to capture such details.

Multi-omics integration methods comparison on functional genomics datasets

In the final of four benchmarks in this study, we applied the multi-omics integration methods to the Cancer Dependency Map (DepMap) Project datasets, which include a wide array of cancer cell lines characterized by multiple omic layers. This benchmark aimed to evaluate the ability of each method to integrate and interpret complex, large-scale multi-omics data with a potential impact on cancer research and therapeutic targeting.

For consistency and comparability with the TCGA benchmark (Fig. 3), the DepMap evaluation focused on 258 cell lines, representing seven of the eight cancer types evaluated in the TCGA benchmark. These include AML, breast, colorectal, kidney, liver, ovarian, and melanoma lineages. We integrated four distinct omics layers—gene expression, DNA methylation, miRNA expression, and metabolomics—reducing this multifaceted data into two latent factors to test precise lineage classification for each cell.

The performance of each method was gauged by the lineage accuracy of the resulting clusters. For methods not intrinsically producing clusters, we employed k-means consensus clustering on the factor matrices. We then compared each cell line cluster against its known lineage, computing a composite score that combined the Adjusted Rand Index (ARI), which measures the similarity between the true classification and the clustering outcome, and cluster purity, an indicator of the homogeneity of the clusters (see Methods). The composite score ranged from 0 to 1, with 1 representing perfect ARI and purity, indicative of optimal clustering.

GAUDI excelled in this comprehensive test, achieving the highest composite score of 0.656, indicating a superior performance in creating condensed and pure clusters (Fig. 5A). This score surpassed that of MOFA+, which followed at 0.561, showing approximately a 15% lower performance compared to GAUDI (Fig. 5B).

The functional genomics benchmark demonstrates GAUDI's strengths in multi-omics data integration. It demonstrated the method's precision and efficiency, even as the complexity of the data increased.

Integration of single-cell chromatin conformation and gene expression data

As an additional use case, we applied GAUDI to recently published large-scale datasets. Single-cell analysis presents unique computational challenges due to data set size, sparsity, technical noise, and the need to identify rare cell populations. These challenges are magnified when integrating different data modalities at single-cell resolution. To further validate GAUDI's capabilities on emerging large-scale single-cell data, we applied it to mouse embryo data with single-cell Hi-C (scHi-C) and RNA-seq (scRNA-seq) profiles from E9.5 embryos as described in Rappoport et al.[26]. After standard normalization and filtering of the scHi-C and scRNA-seq datasets (see Methods), we applied GAUDI using default parameters.

GAUDI integration generated a low-dimensional representation that revealed distinct cell clusters corresponding to major embryonic cell populations (Fig. 6A). Through density-based clustering, GAUDI identified 8 cell populations after removing noise samples, with cluster sizes ranging from 19 to over 2000 cells.

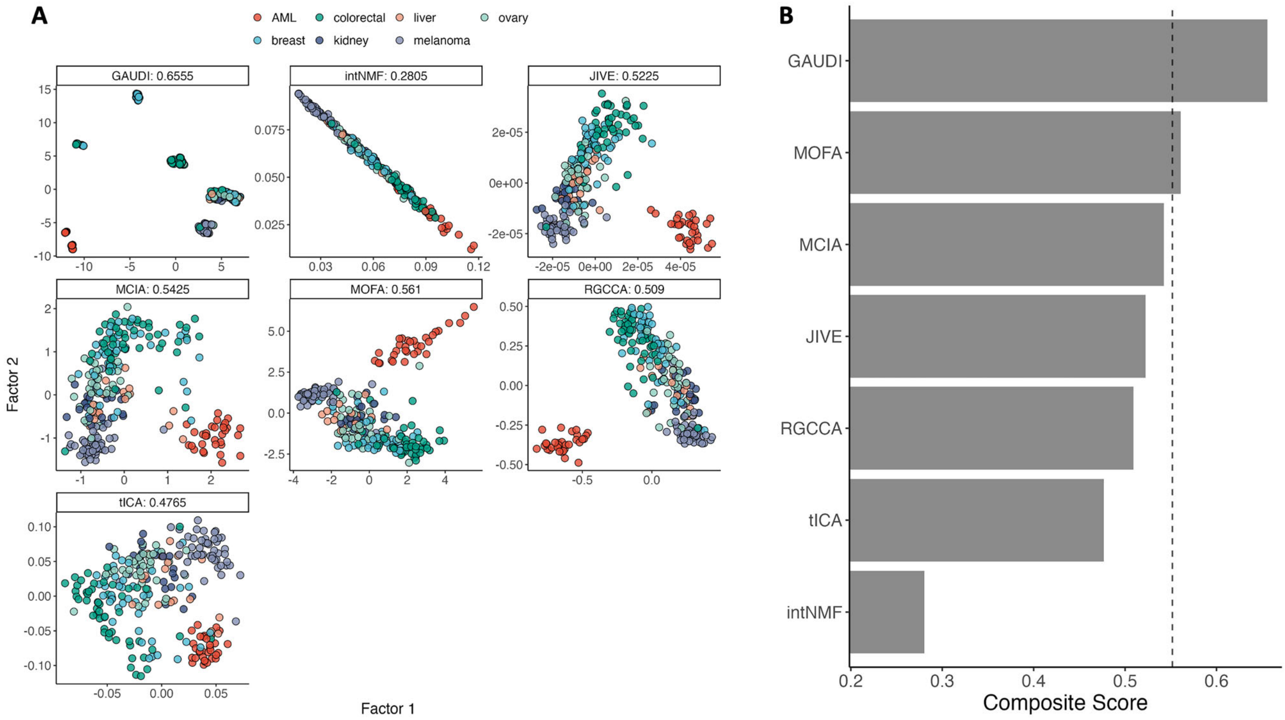

Fig. 5 | Lineage discrimination analysis in DepMap multi-omics data. **A** Scatterplots of the two factors derived from multi-omics integration across each method. The plots are annotated with composite scores, calculated as the average of Adjusted Rand Index (ARI) and cluster purity, which collectively assess the accuracy and homogeneity of the lineage-specific clustering. A higher composite score indicates superior performance in accurately classifying cell lineages based on the integrated multi-omics data. The cells are color-coded by its lineage. **B** Composite scores are presented across the methods, with the vertical dashed line indicating the third quantile (75%). Source data are provided as a Source Data file.

GAUDI's Cluster 0 exhibited highest expression of hemoglobin genes (Hba-a2, Hba-x, Hbb-b1, Hbb-bh1, Hbb-y) and Cpox. This corresponds to the primitive erythrocyte cluster (C3) from the original study. The spatial segregation of Cluster 0 in the manifold (Fig. 6A) aligns with its distinctive gene expression profile. This suggests primitive erythrocytes maintain a unique transcriptional program clearly demarcated from other lineages.

For pluripotency markers, we observed a mixed pattern across multiple clusters. Cluster 0 showed elevated expression of Dppa4. Cluster 1 exhibited the highest expression of Tet2. Cluster 2 showed the highest expression of Zfp42. Clusters 0, 1, and 2 all appear at the periphery of the UMAP embedding (Fig. 6A). This suggests they may share developmental relationships in the higher-dimensional manifold despite appearing distant in the 2D representation. This spatial arrangement, combined with the distributed expression of pluripotency markers across these peripheral clusters, suggests a complex differentiation trajectory. This differs from the discrete ESC population originally proposed by Rappoport et al., possibly reflecting different stages of stem cell characteristic loss during lineage commitment.

The ectoderm markers show a distinct expression pattern with Cluster 3 displaying the most prominent activity. Specifically, Cluster 3 exhibits dramatic upregulation of Crabp1 compared to all other clusters. Cluster 3 also shows moderate peaks for Sox2, Neurog1, Pou3f1, Pou3f2, Sox11, and Crabp2. Six family and Lhx family genes are recognized ectoderm markers but don't show particularly high expression in Cluster 3. Notably, Cluster 6 shows a distinct peak for Lhx6, suggesting potential subpopulation specialization within the ectoderm lineage. The overall expression profile strongly supports Cluster 3 (147 cells) corresponding to the C2.1 ectoderm cluster from Rappoport et al., with Crabp1 serving as the most definitive marker for this population.

For mesoderm/endoderm markers, we observe distinct expression patterns across several clusters. Cluster 1 shows the most pronounced expression profile with a dramatic peak in Ankrd1 expression and moderate upregulation of Igf2, Hand1, Vim, Dlk1, and Tbx5. This suggests correspondence to the C2.3 mesoderm/endoderm cluster from the original study. Cluster 0 displays a unique expression pattern with a distinct peak for Ftl1, suggesting a different mesoderm subpopulation with specialized characteristics. Cluster 5 (largest cluster with 2093 cells) maintains consistent moderate expression across several mesoderm/endoderm markers, indicating a stable, broadly defined mesoderm lineage population. The remaining clusters show more moderate expression patterns across the marker genes, suggesting they may represent intermediate states with less defined identities.

GAUDI's SHAP-based feature importance analysis identified many of the same marker genes highlighted in Rappoport et al. as the most significant contributors to GAUDI's latent space. This indicates that GAUDI's unsupervised approach can recover key biological signals without prior knowledge. GAUDI also identified several ribosomal protein genes (such as Rpl29 and Rps3a) as significant contributors to the latent space, capturing cell cycle or translational activity variation that was filtered out in Rappoport et al.'s analysis, as the authors explicitly modeled and removed cell cycle effects[26]. Our functional analysis of top UMAP coordinate-contributing genes revealed significant ($FDR < 0.05$) enrichment patterns between dimensions (Fig. 6C). The first UMAP dimension showed enrichment for embryonic vascular/leptomeningeal cells, endothelial cells, neural progenitors, and stem cell lines. The second UMAP dimension displayed stronger enrichment for embryonic vascular/leptomeningeal cells, neural progenitors, and pericyte lineage cells, including post-natal vascular signatures. These neurovascular developmental patterns align with the clear primitive erythrocyte separation observed in Fig. 6A. While

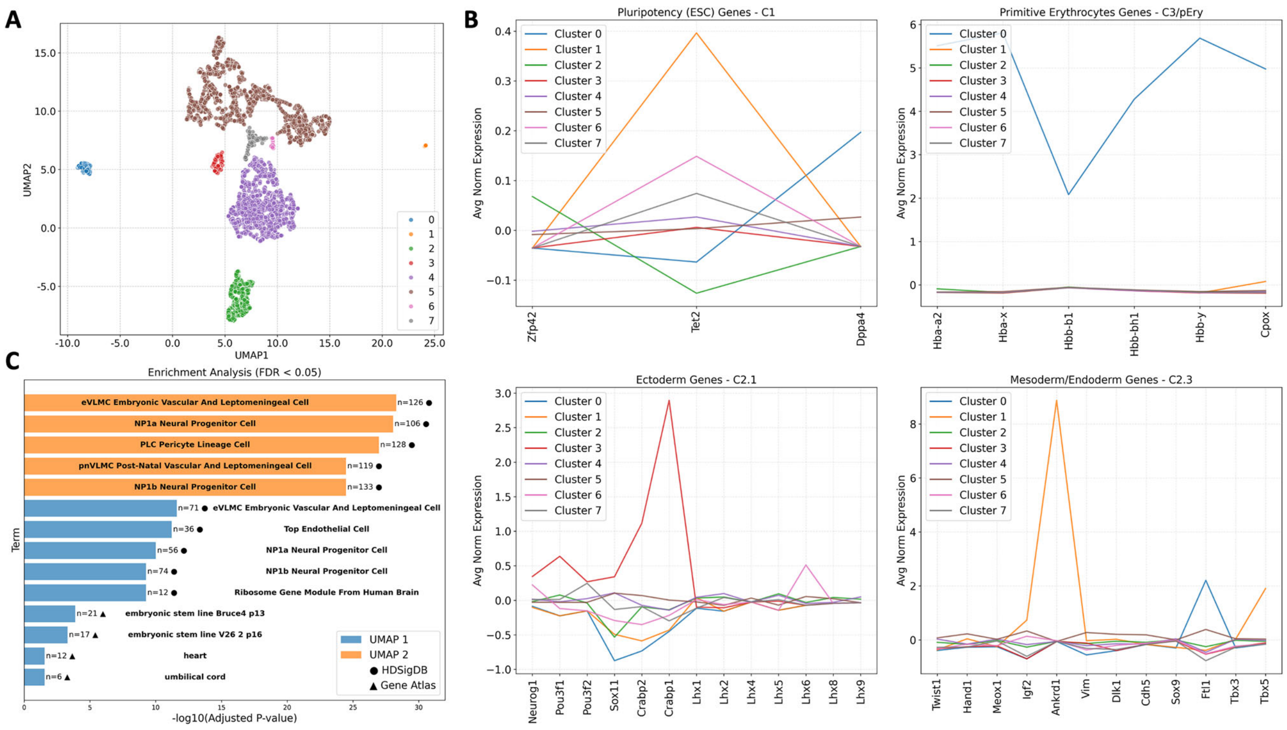

Fig. 6 | Integration of single-cell Hi-C and RNA-seq data from mouse embryos. **A** UMAP visualization of the integrated embedding, with cells colored by GAUDI-identified clusters, showing clear separation of major embryonic cell populations. **B** Expression patterns of key marker genes across GAUDI clusters, organized by gene categories: pluripotency genes associated with ESCs (C1), primitive erythrocyte markers (C3/pEry), ectoderm markers (C2.1), and mesoderm/endoderm markers (C2.3), as defined by Rappoport et al. **C** Functional enrichment analysis of top genes contributing to GAUDI's latent space dimensions, showing significant enrichment of developmental pathways including embryonic vascular and leptomeningeal cell development, neural progenitor cell markers, and pericyte developmental programs. Source data are provided as a Source Data file.

Rappoport et al. employed complex analytical approaches to deconvolve cell cycle effects from lineage-specific chromatin conformations, GAUDI's integrative approach effectively captured both cell cycle and lineage-specific variation in a unified framework. The original study primarily identified four main clusters. GAUDI recapitulated these major divisions and identified additional substructures within the mesoderm and ectoderm lineages. This enhanced resolution can be attributed to GAUDI's ability to simultaneously leverage complementary information from both data modalities.

Discussion

In this study, we developed a novel approach to multi-omics integration and evaluated it against six leading methods benchmarked and found it to be the best joint dimension reduction methods for multi-omics integration[9]. This comprehensive analysis has not only positioned GAUDI as a useful tool for multi-omics integration methods but also demonstrated that, in some instances, GAUDI performs superior to existing techniques, showcasing its proficiency in integrating complex biological data. These advancements are particularly notable in identifying survival outcomes and potential biomarkers, highlighting GAUDI's potential to contribute to the advancement of precision medicine.

The initial benchmark using simulated datasets illustrated GAUDI's superior capability in clustering accuracy. It achieved perfect Jaccard index scores across various scenarios, demonstrating its robustness in the face of both homogeneous and heterogeneous cluster distributions.

In the TCGA benchmark, GAUDI consistently detected sample groups with multi-omic profiles associated with markedly lower overall survival times. Notably, in acute myeloid leukemia, GAUDI identified high-risk patient groups that other methods failed to distinguish, thereby highlighting its utility in clinical prognostication. It is important to note that in this benchmark, the methods used to test the significance of clinical annotations were linear, while GAUDI's factors are non-linear, unlike the other methods. Given this, we would expect poorer performance from GAUDI in this analysis. However, not only did GAUDI perform comparably to other methods, but it also achieved the best results in sarcoma, with a selectivity of 0.83 and two associations. These results highlight GAUDI's capabilities not just in predicting survival but also in its association with a range of clinical annotations.

Our single-cell analysis benchmark further reinforced GAUDI's strength. It was the only method that could detect significant substructures within the same cell line types, which was substantiated by differential gene expression analysis. This capability to discern finer biological details holds promise for advancing our understanding of cellular heterogeneity in cancer and other complex diseases. Additionally, these findings indicate that as the capability to analyze multi-omics in single-cell data advances, GAUDI precisely classifies and manages such data. This capability offers an enhanced level of detail and interpretation, enriching our understanding of complex biological interactions at the single-cell level.

The functional genomics benchmark was perhaps the most challenging, given the scale and diversity of the datasets. Nonetheless, GAUDI demonstrated strong performance, showcasing its efficiency and accuracy even as the dimensionality of the integration increased. The method's ability to generate precise and biologically informative clusters, even with the addition of more 'omic layers, emphasizes its applicability to current high-throughput multi-omics datasets in cancer biology.

Finally, GAUDI successfully integrated mouse embryo single-cell Hi-C and RNA-seq data. It identified eight distinct embryonic cell populations that extended beyond the four clusters in the original study[26]. SHAP feature importance analysis highlighted key developmental marker genes. Functional enrichment identified significant biological pathways including neural progenitor and vascular development. These results demonstrate GAUDI's ability to capture meaningful biological signals without prior knowledge. GAUDI revealed additional substructure within mesoderm and ectoderm lineages that enhances our understanding of embryonic development at the single-cell level.

One limitation that emerged across the benchmarks is the computational intensity inherent in processing large-scale multi-omics data. Although GAUDI is relatively efficient, further advancements in computational strategies are needed to enhance its scalability and speed. Additionally, while GAUDI showed promising performance on synthetic data, this success does not always translate seamlessly to real-world datasets, which tend to exhibit more noise and variability. Another challenge lies in the nature of clustering outcomes. Clustering cell types with high contrast, such as those across three different single-cell datasets, can be more straightforward than distinguishing biologically similar cells, where differences are more nuanced. This subtlety affects the clustering's applicability and raises the question of when clustering or classification is the appropriate tool. In some research scenarios, clustering may not yield useful outcomes, making alternative methods, such as network analysis or pathway enrichment, more suitable for uncovering complex biological interactions.

Moreover, while GAUDI effectively clusters complex biological data, there are inherent limitations concerning overfitting. This concern persists even in unsupervised learning, as models might capture noise or dataset-specific patterns instead of underlying general structures. This issue is particularly pronounced in high-dimensional, noisy multi-omics data, where distinguishing the true signal from noise is challenging. Techniques such as parameter optimization, regularization, and external validation can mitigate overfitting, but they cannot eliminate this risk entirely. Therefore, we recommend users employ these strategies to enhance the robustness of their clustering results. Overfitting concerns are common across all machine learning methods, emphasizing the need for careful validation to ensure models generalize beyond the training data.

GAUDI's performance highlights the importance of developing integration methods that not only provide statistical robustness but also maintain biological interpretability. By achieving accuracy in clustering and showing a strong association with clinical outcomes, GAUDI has proven to be a useful tool for researchers and clinicians, creating new opportunities for discovery in precision medicine.

Future work should focus on extending GAUDI's application to other complex diseases and exploring its potential to guide therapeutic decision-making. Integrating GAUDI with other machine learning models may also improve its classification power for patient stratification and treatment response, offering a more comprehensive framework for interpreting complex datasets.

In conclusion, this study emphasizes the need for advanced multi-omics integration tools like GAUDI, which combines the strengths of UMAP with methodological improvements, to keep pace with the evolving landscape of biomedical data. Our benchmarking supports the adoption of GAUDI in future multi-omics studies and demonstrates its advantages over existing methods. Its ability to shed light on complex biological phenomena enhances our understanding of diseases and informs the development of precision therapies, helping to drive progress in biomedical research.

Methods

UMAP embedding integration for optimizing GAUDI

To optimize the UMAP embedding integration strategy for our GAUDI method, we systematically assessed various approaches to improve data segregation into biologically relevant clusters. This analysis was performed in Python using a custom script, "umap_integration_methods.py", which is available in our source code repository for reproducibility.

For the benchmark, we used gene expression and miRNA expression datasets from DepMap, leveraging cell line lineage information to evaluate the resulting clusters. UMAP was first applied to each omics dataset to generate individual embeddings. We then explored several integration techniques, including intersection, union, subtraction, and concatenation of these embeddings. Additionally, we performed UMAP on the concatenated raw omics datasets and applied joint matrix factorization to the UMAP embeddings, testing the integration techniques on both the concatenated UMAP and the shared matrix embeddings.

Each integration (a total of 10) was clustered using HDBSCAN, and the effectiveness was evaluated based on cluster purity and silhouette scores (Supplementary Fig. 4). This process identified the concatenation of individual UMAP embeddings followed by a combined UMAP as the most effective technique for our GAUDI method.

Integration bias mitigation through dimensional standardization

GAUDI addresses potential integration biases through a two-stage UMAP application process. First, each omics layer is independently embedded using UMAP, creating fuzzy topological representations based on local connectivity patterns rather than raw dimensionality[12]. This initial step effectively normalizes each layer's contribution regardless of its original feature space size. The subsequent concatenation and second UMAP application ensure equal weighting of information from all omics layers through dimensional standardization. We validated this approach using SHAP values derived from XGBoost models predicting UMAP coordinates in our DepMap analysis, which revealed that lower-dimensional omics layers maintained strong contributions to the final embedding. Specifically, metabolomics features showed the highest enrichment scores (19.50 and 17.64 for dimensions 1 and 2, respectively), followed by miRNA data (5.98 for dimension 1), demonstrating that our integration strategy effectively preserves signals from lower-dimensional omics layers without bias from high-dimensional data types.

Parameter selection and implementation details

GAUDI provides comprehensive parameter customization capabilities while offering empirically-derived defaults that enable robust analysis across diverse multi-omics contexts. For the initial dimensionality reduction, GAUDI uses UMAP's established default parameters with *n_neighbors* set to 15, *metric* set to euclidean, and *min_dist* set to 0.01. Our dimensionality reduction follows a two-stage approach. We first reduce each individual omics dataset to 4 dimensions—a choice based on typical dimensionality requirements for capturing primary biological variance components in omics data. This is followed by a final reduction to 2 dimensions for visualization and interpretation of biological relationships. For clustering, if not defined by the user, HDBSCAN's minimum cluster size is automatically computed as 3% of the sample count with a minimum threshold of 2 samples - values selected to provide reasonable cluster sizes across the range of typical multi-omics sample counts while preventing singleton clusters. For downstream analysis of feature importance, we support two approaches. XGBoost parameters are tuned for feature selection in biological data: lambda set to 0 to avoid over-regularization of sparse biological signals, eta set to 0.5 for moderate learning rate, gamma set to 50 to encourage sparse trees, *max_depth* set to 10 to capture complex feature interactions, and subsample set to 0.95 to maintain stable feature selection. Alternatively, Random Forest can be used for broader feature inclusion. While XGBoost performs efficient feature selection, Random Forest tends to capture more features at higher computational cost, allowing researchers to choose based on their analytical needs.

Like most machine learning approaches, while these parameters provide robust performance across typical multi-omics datasets, users are encouraged to optimize them for their specific biological context through standard cross-validation and benchmarking procedures, particularly when analyzing datasets with unique characteristics or investigating novel biological phenomena.

Multi-omics data simulation

The multi-omics datasets in our study were simulated using the *InterSIM* package, available on CRAN[19]. This package creates interrelated datasets that emulate the complexity of real-world data, based on DNA methylation, mRNA gene expression, and protein expression from TCGA ovarian cancer data. For our analysis, we generated 500 simulated samples, increasing the sample size of 100 used in Cantini et al.'s paper. As described in Cantini et al. benchmarking workflow, we created five, ten, and fifteen clusters of equal and randomly varying sizes.

Latent space clustering

For those benchmarked methods that do not inherently generate clusters (excluding intNMF and GAUDI), we employed k-means clustering on the latent space matrix (i.e., factor matrix), aligning with the approach outlined by Cantini et al. Given the stochastic nature of k-means clustering, we executed the clustering process 1000 times. Subsequently, we derived a consensus clustering by determining the most common sample-to-cluster associations across these iterations. This approach aligns with the clustering strategy employed by intNMF, which utilize Euclidean distance-based k-means clustering.

Clustering quality metrics

The Jaccard Index is a statistic used for gauging the similarity and diversity of sample sets. The index measures the similarity between sample sets by dividing the number of observations common to both sets by the total number of distinct observations in both sets. Its value ranges from 0 to 1, where 0 signifies no similarity and 1 denotes that the sets are identical.

The C-index is a clustering evaluation metric that assesses the goodness of a clustering result in relation to the within-cluster variation and the separation between clusters. A lower C-index value indicates a better clustering structure, i.e., small intra-cluster distances and large inter-cluster distances.

The Adjusted Rand Index (ARI) quantifies the congruence between two clustering solutions, taking into account the possibility of chance groupings. It accounts for both true positives and false positives within the clustering process. An ARI of 1 signifies an impeccable correspondence between the two clustering solutions, whereas an ARI of 0 indicates that the clustering is no better than a random assignment.

Cluster purity is a simple and transparent evaluation measure. It assesses the homogeneity of clusters with respect to the class distribution (e.g., lineage). The purity of a cluster is achieved when most of its members belong to the same class, with a purity score ranging from 0 (no purity) to 1 (complete purity).

To evaluate the clusters in the DepMap benchmark, we defined composite scores as the average of ARI and cluster purity, collectively assessing the accuracy and homogeneity of the lineage-specific clusters.

Finally, the silhouette score is used to measure how similar a sample is to its own cluster compared to other clusters. The silhouette score for a sample is calculated based on how much closer it is to members of its own cluster versus members of the nearest cluster to which it does not belong. The goal is to have a high silhouette score (with 1 being the highest value), which indicates that the sample is well matched to its own cluster and poorly matched to neighboring clusters.

Each of these metrics contributes to a comprehensive evaluation framework, providing us with a multifaceted understanding of the clustering quality and the effectiveness of each method in grouping multi-omics data.

Selectivity score

The selectivity score, as defined by Cantini et al., quantifies the specificity of the associations between latent factors and clinical annotations. This score achieves a maximum of 1 when each latent factor associates with a unique clinical or biological annotation, reflecting a highly specific one-to-one relationship. Conversely, the score approaches a minimum of 0 when latent factors are broadly and non-specifically associated with multiple annotations. An ideal method would aim to maximize the number of meaningful, unique factor-annotation associations while ensuring the selectivity score remains high, thereby achieving specificity without excessive generalization[9].

Cluster gene differential expression analysis in single-cell data

To explore molecular differences between clusters 2 and 3 identified by GAUDI, within the K562 cell line, we performed a Limma differential expression analysis using the POMA R package[24,27]. Our analysis sought to detect genes exhibiting significant expression differences between the clusters, considering an adjusted p-value of 0.05 as the threshold for significance. A volcano plot was also created to illustrate these differences, with a fold-change cutoff of 1.5 and the same p-value threshold to emphasize the most significant changes.

After this, we conducted an enrichment analysis using the R package clusterProfiler to understand the biological implications of the differentially expressed genes, focusing on KEGG pathways[28]. This analysis, employing 10,000 permutations and an FDR-adjusted p-value of 0.05, helped to contextualize the gene expression patterns within broader biological pathways.

For an in-depth understanding of the benchmarking framework employed in our study, we recommend consulting the detailed exposition provided in the paper by Cantini et al.

Sample matching for single-cell Hi-C and RNA-seq integration

To integrate chromatin conformation with gene expression at single-cell resolution, we developed a computational approach to match cells between separate scHi-C and scRNA-seq datasets. The data consisted of 4993 cells with scHi-C profiles (59,639 features) and 4781 cells with scRNA-seq measurements (27,306 features). After removing constant features (45 from scHi-C and 9732 from scRNA-seq), we standardized both datasets using z-score normalization and performed dimensionality reduction via PCA, selecting 2445 components that explained ~80% of variance in the scRNA-seq data (79.4%) and the scHi-C data (70.5%).

Since the scRNA-seq dataset contained fewer cells, we employed a greedy algorithm to identify the optimal subset of scHi-C cells that best matched the scRNA-seq cells in the reduced-dimension space. The matching quality was assessed using Procrustes analysis (final alignment disparity: 0.67). The resulting integrated dataset comprised 4781 cells with matched scHi-C and scRNA-seq measurements, enabling subsequent multi-modal analyses of chromatin structure and gene expression relationships.

Computational efficiency and scalability

GAUDI demonstrates efficient performance across varying dataset sizes. Our scalability analysis across 60 scenarios (varying from 50 to 1000 samples, 1000–25,000 features, and 2–5 omics layers) shows approximately linear scaling with dataset size (Supplementary Fig. 5). Execution times range from seconds for smaller datasets to under 10 minutes for the most complex configuration (1000 samples ×

25,000 features × 5 omics layers) on standard hardware. Memory usage also scales linearly with total dataset size, while maintaining clustering quality (silhouette scores with $Q1 = 0.51$, $\text{median} = 0.61$, $Q3 = 0.68$) regardless of data complexity. This makes GAUDI suitable for routine analysis of typical multi-omics datasets without requiring specialized computing resources.

Reporting summary

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

Data availability

The simulated data are produced using the R package *InterSIM* and can be reproduced using our quarto document *"01_comparisons_on_simulated_data.qmd"*. The cancer TCGA data were downloaded from http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html. The single-cell data are available in the *"data/"* folder of our GitHub repository (https://github.com/hirscheylab/umap_multiomics_integration). Biological annotations were obtained from Cantini et al. GitHub repository and can be found in our GitHub repository as well. The Cancer Dependency Map datasets were obtained from https://depmap.org/portal/download/custom. Finally, the scHi-C and scRNA-seq datasets were obtained from the GEO database under accession code GSE148793. Source data are provided with this paper.

Code availability

The source code necessary to reproduce the analyses and figures presented in this paper is fully available to the scientific community at our GitHub repository under MIT license, https://github.com/hirscheylab/umap_multiomics_integration, ensuring transparency and ease of access. The specific version of the code associated with this publication is archived in Zenodo and is accessible via https://doi.org/10.5281/zenodo.15442172[29]. In addition to the source code, we have developed a dedicated R package for the method presented in this paper, GAUDI. This package is designed to facilitate the easy implementation of our method by other researchers. The package includes comprehensive documentation and examples to assist users in implementing our method in their own research. We encourage researchers to use these resources for their work and welcome any contributions or feedback to improve the utility and functionality of this method.

References

1. Turanli, B., Yildirim, E., Gulfidan, G., Arga, K. Y. & Sinha, R. Current state of "omics" biomarkers in pancreatic cancer. *J. Pers. Med.* **11**, 127 (2021).
2. Krassowski, M., Das, V., Sahu, S. K. & Misra, B. B. State of the field in multi-omics research: from computational needs to data mining and sharing. *Front. Genet.* **11**, 610798 (2020).
3. Tenenhaus, M., Tenenhaus, A. & Groenen, P. J. Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika* **82**, 737–777 (2017).
4. Bady, P., Dolédec, S., Dumont, B. & Fruget, J. F. Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities. *Comptes Rendus Biol.* **327**, 29–36 (2004).
5. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 1–17 (2020).
6. Chalise, P. & Fridley, B. L. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PloS one* **12**, e0176278 (2017).
7. Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7**, 523 (2013).
8. Teschendorff, A. E., Jing, H., Paul, D. S., Virta, J. & Nordhausen, K. Tensorial blind source separation for improved analysis of multi-omic data. *Genome Biol.* **19**, 1–18 (2018).
9. Cantini, L. et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).
10. Makrodimitris, S., Pronk, B., Abdelaal, T. & Reinders, M. An in-depth comparison of linear and non-linear joint embedding methods for bulk and single-cell multi-omics. *Brief. Bioinforma.* **25**, bbad416 (2024).
11. Vahabi, N. & Michailidis, G. Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. *Front. Genet.* **13**, 854752 (2022).
12. McInnes et al. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
13. ElKarami, B., Alkhateeb, A., Qattous, H., Alshomali, L. & Shahrrava, B. Multi-omics data integration model based on UMAP embedding and convolutional neural network. *Cancer Informatics*. **21**, 11769351221124205 (2022).
14. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
15. Huang, H., Wang, Y., Rudin, C. & Browne, E. P. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Commun. Biol.* **5**, 719 (2022).
16. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**, 205 (2017).
17. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). (2016).
18. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. (2017).
19. Chalise, P., Raghavan, R. & Fridley, B. L. InterSIM: simulation tool for multiple integrative 'omic datasets'. *Comput. Methods Prog. Biomed.* **128**, 69–74 (2016).
20. Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids Res.* **46**, 10546–10562 (2018).
21. Goldman, S. L. et al. The impact of heterogeneity on single-cell sequencing. *Front. Genet.* **8**. (2019).
22. Poirion, O., Zhu, X., Ching, T. & Garmire, L. X. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. *Nat. Commun.* **9**, 4892 (2018).
23. Liu, L. et al. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.* **10**, 470 (2019).
24. Castellano-Escuder, P., González-Domínguez, R., Carmona-Pontaque, F., Andrés-Lacueva, C. & Sánchez-Pla, A. POMAShiny: a user-friendly web-based workflow for metabolomics and proteomics data analysis. *PLOS Comput. Biol.* **17**, e1009148 (2021).
25. Zhou, B. et al. Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.* **29**, 472–484 (2019).
26. Rappoport, N., Chomsky, E., Nagano, T. & Tanay, A. Single cell Hi-C identifies plastic chromosome conformations underlying the gastrulation enhancer landscape. *Nat. Commun.* **14**, 3844 (2023).
27. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
28. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* **16**, 284–287 (2012).
29. Castellano-Escuder, P., Zachman, D. K., Han, K. & Hirschey, M. D. GAUDI: Interpretable multi-omics integration with UMAP

embeddings and density-based clustering. Zenodo, https://doi.org/10.5281/zenodo.15442172, (2025).

Acknowledgements

We would like to thank the entire Hirschey lab for thoughtful feedback. We would also like to acknowledge funding support from the NIAID 1R38AI40297 (DKZ), 1K38CA282963 (DKZ), and from developmental funds of the Duke Cancer Institute as part of the Cancer Center Support Grant P30 CA014236 (MDH), from the National Institutes of Health and the NIA R01AG045351 (MDH) and 1R21AG080334 (MDH).

Author contributions

P.C.E., D.K.Z., and M.D.H. conceived and designed the study. P.C.E. developed the GAUDI method and implemented the computational framework. P.C.E. performed the computational analyses and benchmarking. P.C.E. and K.H. created the visualizations. P.C.E., D.K.Z., and M.D.H. interpreted the results and analyzed the biological implications. P.C.E. and K.H. developed the R package. P.C.E. and M.D.H. wrote the manuscript with contributions from all authors. M.D.H. supervised the project and secured funding. All authors read, revised, and approved the final manuscript.

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/s41467-025-60822-1.

Correspondence and requests for materials should be addressed to Matthey D. Hirschey.

Peer review information *Nature Communications* thanks Abedalrhman Alkhateeb and the other anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

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