nature methods

Article

https://doi.org/10.1038/s41592-026-03007-y

DECODE: deep learning-based common deconvolution framework for various omics data

Received: 11 February 2025

Accepted: 21 January 2026

Published online: 2 March 2026

Check for updates

Tianyi Zhao[1,2,8], Renjie Liu[2,3,8], Yuzhi Sun[3,8], Bingtian Wang[3], Liyuan Zhang[3], Qiuhao Chen[2], Ruibang Luo[4], Zhiyuan Yuan[5], Guohua Wang[3], Liang Cheng[6,7] ✉ & Yadong Wang[3] ✉

Deconvolution algorithms estimate cell-type abundances from tissue-level data, enabling systematic cellular analysis of large cohorts. However, most deconvolution algorithms are specifically designed for single-omics data, thereby limiting their generalizability and scalability for various omics data from different cohorts. Here we present DECODE, a universal deconvolution framework for both cell types and cell states that can be applied to transcriptomic, proteomic and metabolomic data, and that seamlessly integrates diverse multiomics tissue datasets at the cellular level. DECODE fills the gap in metabolomics deconvolution and significantly outperformed state-of-the-art methods on different omics data across donors, disease conditions, healthy states, datasets and measurement platforms. In addition, DECODE exhibits high robustness in scenarios that are closer to real applications so it can accurately deconvolve known cell types even when the reference single-cell data are incomplete. DECODE will serve as a powerful tool for the fully extending multiomics cohort data into cellular level.

Cell abundance represents the proportional distribution of various cell types within a tissue, and its dynamic changes are pivotal in biological processes, including organ development[1], genetic regulation[2] and disease treatment[3]. Single-cell technologies of various omics have become the most direct tools for investigating cellular heterogeneity. However, their high costs remain major barriers to their widespread application in large-scale cohort studies. Moreover, the extensive number of preserved tissue samples cannot be processed into single-cell suspensions, further limiting the feasibility of single-cell approaches. Therefore, the generalization of tissue-level multiomics cohort data to the cellular level has become the key to explore cellular heterogeneity. Deconvolution methods, which leverage single-cell data as reference to infer cellular composition of tissue-level data, offer a cost-effective alternative for studying cellular heterogeneity in complex biological systems[4].

At present, deconvolution algorithms in the multiomics field have followed a specialized development paradigm. At the transcriptomic level, representative tools for tissue data such as MuSiC[5] and CIBERSORTx[6] have been widely adopted, while methods such as RCTD[7] and SPOTlight[8] specifically address the analytical needs of spatial transcriptomics. In proteomics, specialized approaches such as scpDeconv[9] have emerged. Although these tools perform effectively

[1]School of Medicine and Health, Harbin Institute of Technology, Harbin, China. [2]Harbin Institute of Technology Zhengzhou Research Institute, Zhengzhou, China. [3]Faculty of Computing, Harbin Institute of Technology, Harbin, China. [4]Department of Computer Science, The University of Hong Kong, Hong Kong, China. [5]Institute of Science and Technology for Brain-Inspired Intelligence; MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence; MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China. [6]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China. [7]National Health Commission (NHC) Key Laboratory of Molecular Probes and Targeted Diagnosis and Therapy, Harbin Medical University, Harbin, China. [8]These authors contributed equally: Tianyi Zhao, Renjie Liu, Yuzhi Sun. ✉e-mail: liangcheng@hrbmu.edu.cn; ydwang@hit.edu.cn

within their respective domains, their applicability on other omics remains highly uncertain. Most existing deconvolution methods for transcriptomic data are based on assumed distributions, such as the Poisson or negative binomial distribution, which are considered unsuitable for other omics data[9]. Furthermore, spatial deconvolution methods, while proficient at capturing microenvironmental heterogeneity[7], may misattribute biological variability in nonspatial contexts to spatial effects, potentially leading to distorted interpretations. This omics-specific methodological landscape presents two major challenges. First, when researchers aim to compare cell-type abundances across different omics layers and cohorts, methodological heterogeneity introduces unquantifiable systematic biases, undermining the reliability of integrative analyses. Second, multiomics studies demand substantial effort in cross-method parameter tuning and result calibration, leading to fragmented workflows that increase analytical complexity and time costs. These issues collectively pose a significant technical bottleneck to the scalability of large-scale multiomics research. Notably, dedicated deconvolution tools for metabolomic data are still lacking, despite metabolomics exhibiting the highest correlation with clinical phenotypes among all omics types[10,11]. This highlights the urgent need for a unified computational framework capable of accommodating the diverse data characteristics of different omics modalities: an essential step toward advancing the goals of precision medicine.

Developing a universal deconvolution framework for diverse omics data faces major technical hurdles. Different omics modalities vary in scale, distribution, sparsity and feature dimensionality, demanding highly flexible, adaptable model architectures. Another key challenge is cell-type mismatch between single-cell and tissue data, since single-cell omics may not capture all cell types present in tissues[12]. Thus, the model must exhibit robustness to noise to reliably estimate the relative abundances of n known cell types in tissue-level data with $n + m$ cell types (where $m \geq 1$), even when the reference single-cell data include only n cell types. Another major challenge is strong batch effects: because single-cell reference and tissue-level data typically come from different donors, technologies and health states, physiological and technical variation introduces substantial batch effects[13] that obscure true biological signals and hinder accurate deconvolution. In summary, although existing methods work in specific settings, there is still no universal framework that simultaneously handles omics heterogeneity, incomplete single-cell references and severe batch effects, which limits the broader application of deconvolution to existing large amounts of multiomics tissue data.

To address these challenges, we developed DECODE, a universal deconvolution framework for cell types and states in transcriptomics, proteomics and metabolomics that integrates tissue datasets across omics at the cellular level. It achieves high accuracy and robustness in estimating cell-type proportions even with incomplete single-cell references, and its consistent performance across heterogeneous omics datasets makes DECODE a powerful tool for multiomics analysis, enabling integrated multimolecular insights into the cellular composition of complex biological systems.

Results

Overview of DECODE

DECODE integrates adversarial training and contrastive learning techniques into a sophisticated computational framework. This framework consists of four stages that collectively ensure precise and reliable deconvolution across diverse omics data.

Stage 1. Pseudotissue samples are first generated for model training (Fig. 1a). A random cell-type proportion vector for all target cell types is drawn from a uniform distribution. Given these proportions and a preset total cell count, cells are randomly sampled from the corresponding single-cell tissue data, and their molecular profiles are aggregated to form one pseudotissue sample. Repeating this procedure multiple times yields the DECODE training set.

Stage 2. The next step is addressing batch effects between pseudotissue and target data (Fig. 1b). In this stage, the encoder, discriminator and eDeconvolver modules are trained together, guided by a combination of L1 loss and binary cross-entropy loss. This cooperative training process is designed to use a specially designed loss function to force the discriminator to fail in identifying the origin of the features, thereby effectively mitigating batch effects between target-tissue and train-tissue while preserving the biological signals required for the deconvolution task. On completing this stage, encoder parameters are fixed and subsequently passed to stage 3.

Stage 3. This stage aims to improve the robustness in dealing with various noise and omics data (Fig. 1c). Noise not exceeding 10% is randomly added to each train-tissue sample twice, creating training pairs with the train-tissue samples. The training pairs are first passed through encoder with fixed parameters in stage 2 and mapped to a different latent space via a dimension enhancement module (DimExpander). Then, the noisy part of training pairs are processed by an attention-based[14] denoiser module (Fig. 1e), which separates embedding features into noise features and purified train-tissue features. The purified train-tissue features undergo dimensionality reduction through linear attention and deconvolver operations to yield the predicted labels for training samples. All modules in stage 3 are supervised during training by L1 loss and a contrastive learning strategy (Fig. 1f).

Stage 4. In the final step, DECODE takes the target-tissue as input and outputs a cell-abundance vector for each sample. This inference can follow two pathways (Fig. 1d): standard deconvolution for pure tissues and relative deconvolution for tissues with unknown cell types. When single-cell reference cell types do not fully match the target-tissue (the common real-world case), the pathway with the denoiser is used; otherwise, the alternative pathway is applied.

Superior performance over previous approaches in transcriptomic and proteomic deconvolution

To evaluate the deconvolution performance of DECODE, comparisons were made with state-of-the-art deconvolution methods, including TAPE[15], CIBERSORTx[6], MuSiC[5], scpDeconv[9], Scaden[16], RCTD[7], Seurat[17], SPOTlight[8], Tangram[18], ucdselect[19] and cell2location[20] using both transcriptomics and proteomics data. Performance was evaluated via three metrics: Lin's concordance correlation coefficient (CCC), root mean square error (r.m.s.e.) and Pearson's correlation coefficient (Pearson's r).

This study conducted experiments across 15 datasets, establishing 7 scenarios to evaluate performance under diverse conditions. Scenario 1 assesses cross-donor generalization using a human lung transcriptomic dataset[21], with training and testing sets derived from different donors in similar health states. Scenario 2 evaluates cross-disease generalization on a human breast transcriptomic dataset[22], with estrogen receptor-positive samples for training and estrogen receptor-negative samples for testing. Scenario 3 tests cross-health state performance using a human breast proteomic dataset[23], training on premenopausal samples and testing on postmenopausal samples. Scenario 4 examines cross-dataset robustness across two omics modalities: (1) mouse islet transcriptomic datasets from ref. 24 (training) and ref. 25 (testing) and (2) mouse cell line proteomic datasets from ref. 26 (training) and ref. 27 (testing). The proteomics data also enables evaluation of cross-platform generalizability. Scenario 5 focuses on spatial transcriptomics performance, using STARmap data[28] from the mouse visual cortex (focus on 12 cell types) and Slide-seqV2 hippocampal data[29] (14 cell types, loaded via Squidpy[30]). Scenario 6 uses a human retinal transcriptomic dataset[31] containing 17 cell types to evaluate scalability to a

https://doi.org/10.1038/s41592-026-03007-y

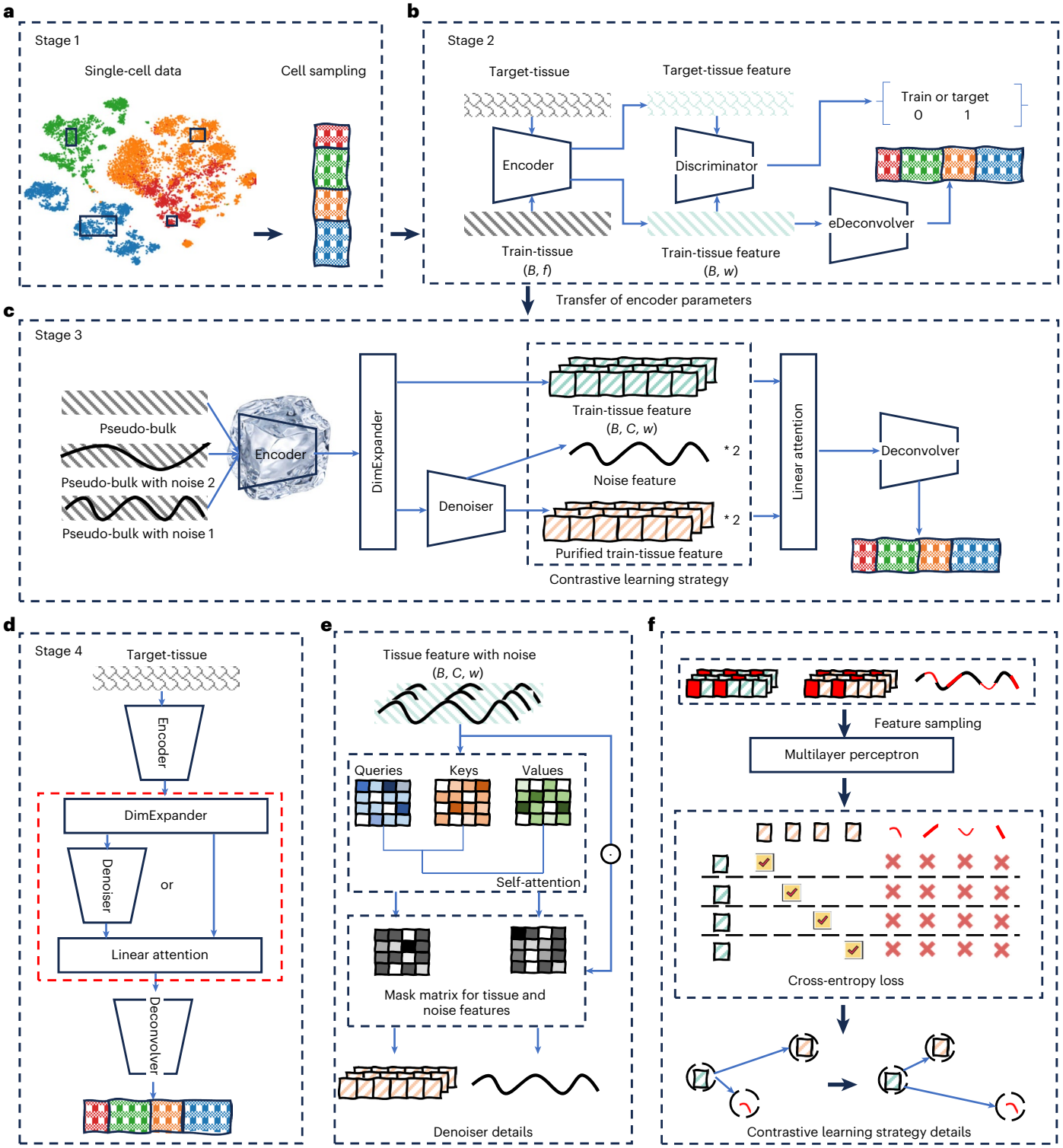

Fig. 1 | Framework of DECODE. The framework consists of four stages. **a**, Generating train-tissue data through random sampling from single-cell data. Different colors represent distinct cell types or states. **b**, Removing batch effects via adversarial training between target-tissue and train-tissue data. **c**, Performing feature enhancement and denoising through contrastive learning with paired train-tissue data (with and without noise). **d**, Implementing two inference pathways for tissue samples with and without unknown cell types. **e**, The denoiser module (from **c**) is presented. In this module, mask matrices are generated for purified train-tissue and noise features through a self-attention mechanism. These mask matrices are elementwise multiplied with the input features mixed with noise to obtain the purified train-tissue and noise features. **f**, The details of the contrastive learning loss strategy (in **c**). Train-tissue, purified train-tissue and noise features are randomly sampled. If the sampling positions are same for both train-tissue and purified train-tissue, the sampled features are treated as positive pairs, and the train-tissue and any noise features at arbitrary positions are treated as negative pairs. Through the cross-entropy loss, the distance between positive pairs is minimized, whereas the distance between negative pairs is maximized. The green pixel blocks represent the sampling of train-tissue features, the orange pixel blocks represent the sampling of purified train-tissue features and the red lines represent the sampling of noise features. Checkmarks are placed on the diagonal of the sampling matrix between train-tissue features and purified train-tissue features, indicating that they are sampled from the same location and represent a pair of positive samples. In the sampling matrix between train-tissue features and noise features, all samples are negative and thus marked with crosses. B, batch size; f, number of train-tissue features; C, dimension of the projected features; w, number of features after dimension reduction.

https://doi.org/10.1038/s41592-026-03007-y

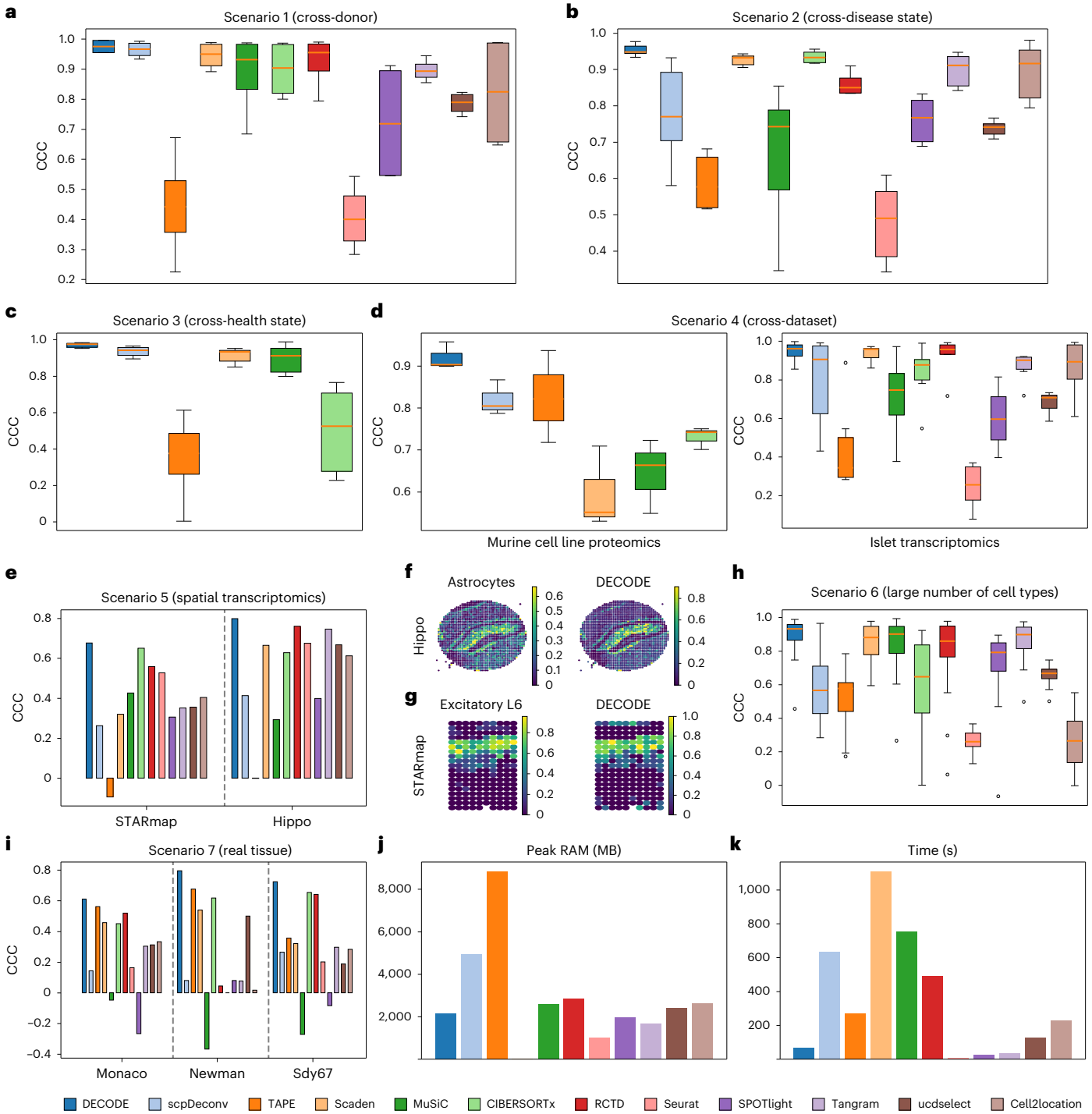

Fig. 2 | Overview of DECODE's transcriptomics and proteomics data deconvolution performance. a–**e**, Comparison of CCC across different methods under scenarios 1–5. **a**, Cross-donor evaluation on transcriptomic data ($n = 4$ cell types). **b**, Cross-disease state evaluation on transcriptomic data ($n = 6$ cell types). **c**, Cross-health state evaluation on proteomics data ($n = 6$ cell types). **d**, Cross-dataset evaluation on both proteomics ($n = 3$ cell types, left) and transcriptomic ($n = 6$ cell types, right) data. **e**, Evaluation on spatial transcriptomics data. **f**,**g**, Comparison between the ground-truth proportions of cell types (left) and the predictions of DECODE (right). **f**, Astrocytes in Hippo. **g**, Excitatory L6 in STARmap. **h**, CCC comparison across methods on transcriptomic data with a large number of cell types ($n = 17$ cell types). **i**, CCC comparison across methods in three real bulk datasets. **j**,**k**, Peak RAM usage (**j**) and runtime (**k**) of DECODE compared to other methods in scenario 1. The CCC values range from −1 to 1, with values closer to 1 indicating better performance. The r.m.s.e. and Pearson's r metrics for each scenario are presented in Supplementary Figs. 1 and 2, respectively. Box plots show the median (center line), the 25th–75th percentiles (bounds of the box) and the minimum and maximum values (whiskers); dots indicate outliers. The number of cell types represents how many data points are contained within each box in the box plot. Unless otherwise stated, this definition applies to all subsequent box plots. A heatmap comparing the predicted and true proportions of other major cell types in spatial transcriptomic data is shown in Supplementary Fig. 3. All metrics for the proteomics data across all comparison methods (including spatial transcriptomics methods) are presented in Supplementary Fig. 7b–d. For evaluation methods, see the section 'Calculation methods for evaluation indicators' in the Supplementary Information for details.

large number of cell types, with samples from different retinal regions used for training and testing. At the same time, scenario 5 also partly assesses performance with a large number of cell types. For scenarios 1–6, pseudotissue samples were generated by randomly mixing single cells at varying proportions. Scenario 7 evaluates these methods on real tissue data using three peripheral blood transcriptomic datasets with ground-truth cell-type compositions[32–34] (SDY67).

DECODE shows superior and stable performance across diverse challenging scenarios, achieving top CCC values in cross-donor (Fig. 2a), cross-disease (Fig. 2b), cross-health state (Fig. 2c), cross-dataset (Fig. 2d), spatial transcriptomics (Fig. 2e) and multi-cell-type deconvolution tasks (Fig. 2h). Although a few methods slightly exceed DECODE in Pearson correlation in scenarios 1, 2, 4 and 5 (Supplementary Figs. 1 and 2), DECODE consistently outperforms both omics-specific and spatial-specific methods in all other aspects of complex applications, including the original problem settings of those methods. In spatial transcriptomics, DECODE's predictions closely match ground-truth cell-type distributions across major populations (Fig. 2f,g and Supplementary Fig. 3), further confirming its accuracy. Together, these results demonstrate DECODE's robustness across tasks of varying difficulty, omics types, and technological platforms. In real tissue data (scenario 7), DECODE only shows a slightly higher r.m.s.e. on the Monaco dataset compared to TAPE[15] and Scaden[16] (Supplementary Fig. 2d), it surpasses all methods in other critical aspects. We also show true versus predicted scatter plots for the three datasets in scenario 7 (Supplementary Figs. 4–6). Because each dataset has few samples (approximately 10–20) and highly concentrated cell-type proportions, cell-type-wise metrics are unstable, so overall performance provides a more reliable assessment of each method.

To assess DECODE's practicality, we compared its peak memory usage and runtime with other methods on the scenario 1 dataset (Fig. 2j); it ranked fifth and fourth, respectively, indicating reasonable efficiency. Overall, these comparisons with state-of-the-art approaches show that DECODE is currently the most effective deconvolution method for both transcriptomics and proteomics.

Accurate and stable metabolomics deconvolution

A large amount of tissue-level metabolome data has been generated for clinical applications, since metabolites can swiftly reflect changes in cell and tissue functions. However, research on the deconvolution of metabolomic data is still missing. To evaluate the performance of DECODE using metabolomics data, three single-cell metabolomics datasets were obtained from mouse liver[35], mouse bone marrow[36] and human colorectal cancer[37], containing 244, 107 and 112 metabolite features, respectively. The liver dataset includes three cell types (hepatocytes, Kupffer cells, endothelial cells), the bone marrow dataset five (granulocyte-monocyte progenitors, B cells, T cells, myeloid cells, erythroid cells) and the colorectal cancer dataset five (cancer cells, fibroblasts, B cells, myeloid cells, T cells). Because donor information is unavailable for the liver and bone marrow data, each dataset was split equally into two groups: one for generating pseudotissue training data and the other for testing. For the colorectal cancer dataset, we trained on mismatch repair-deficient samples and tested on conventional colorectal cancer samples to assess deconvolution performance under complex metabolomic biological heterogeneity.

Deconvolution of metabolomics data is significantly more challenging than that of transcriptomics or proteomics. Single-cell transcriptomics can simultaneously measure the expression levels of tens of thousands of RNAs, while mass spectrometry-based single-cell proteomics approaches, such as SCoPE and nanoPOTS, can detect approximately 1,000 to 3,000 proteins per cell[27,38,39]. In contrast, although single-cell metabolomics has advanced rapidly toward higher throughput, the number of detectable metabolites remains in the range of hundreds[35–37]. As a result, metabolomics offers the smallest number of available features. In addition, we assessed feature similarity across different omics data[35,40,41] for the same cell types within liver, and found that metabolomic profiles exhibited the highest similarity across cell types compared to transcriptomic and proteomic profiles (Fig. 3a). To further illustrate this limitation, we investigated the differential metabolites among different cell types within three datasets (Supplementary Tables 1–3) and found that only granulocyte-monocyte progenitor cells displayed markedly distinct metabolites compared to the other cell types (Fig. 3b and Supplementary Fig. 9a,b). The limited number of features and the high similarity between cell types make it difficult to characterize different cell types through metabolite profiling. These challenges require DECODE to be able to capture subtle signals of cellular differences.

In the comparison of DECODE with other methods on the three datasets (the CIBERSORTx[6] protocol failed to process the liver and the colorectal cancer dataset, and the MuSiC[5] method failed to process the colorectal cancer data), it can be observed that DECODE is only slightly inferior to MuSiC[5] in terms of CCC on the mouse liver dataset, but clearly outperforms all other methods in all other metrics (Fig. 3c and Supplementary Figs. 8 and 9c). To examine how high similarity in metabolic features impacts deconvolution performance, we plotted true versus predicted cell proportions for each method (Supplementary Figs. 10–12). DECODE's points lie close to the 1:1 line, indicating accurate estimates, whereas other methods show x axis truncation for poorly characterized cell types, reflecting their inability to distinguish weak intercellular signals.

Overall, DECODE demonstrates superior performance in metabolomic deconvolution.

Accurate cell state deconvolution on three omics datasets

Cell states, such as differentiation, activation and apoptosis, offer valuable insights into cellular functions from a perspective distinct from that of cell types. Accurately quantifying cell state abundances is essential for understanding processes such as organ development, cell division and responses to external stimuli[42,43]. To evaluate DECODE's performance in reconstructing cell state abundances associated with pseudotime trajectories, cell cycle phases and drug response time points, we curated three datasets.

Dataset 1, derived from MeDuSA[44], is a monocyte case study with explicitly annotated pseudotime states. MeDuSA is an algorithm for recovering pseudotime state abundances from tissue transcriptomes. As shown by the uniform manifold approximation and projection visualization (Fig. 4a), pseudotime states are clearly separated in low-dimensional space. Continuous pseudotime values (0–1) were discretized into ten temporal labels, and cells in each state were randomly split into training and testing sets. Dataset 2 comes from ref. 45 that profiled proteins in monocytes and melanoma cells across cell cycle phases (G1, S, G2). The differentially expressed proteins across cell types but within the same phase (Fig. 4b) show that cell state yields highly consistent protein expression regardless of cell type. Here, melanoma cells were used for training and monocytes for testing. Dataset 3, from ref. 46, measured 20 proteins and metabolites involved in signaling, phenotypic, and metabolic regulation in melanoma cells over 5 days of drug treatment. Violin plots (Fig. 4c) show marked expression changes with increasing treatment duration. Cells at different time points were treated as distinct states and, within each state, cells were randomly split into training and testing sets.

DECODE consistently achieved the best performance across all datasets (Fig. 4d and Supplementary Fig. 13). This finding demonstrates DECODE's ability to recover cell state abundances related to pseudotime trajectories, cell division cycles and cell state changes induced by environmental variations. Notably, since MeDuSA[44] requires a reference single-cell matrix with continuous pseudotime labels, comparisons with MeDuSA[44] were conducted only using dataset 1. As the cell state labels in datasets 2 and 3 are discrete, they are unsuitable for MeDuSA[44]. In addition, the CIBERSORTx[6] protocol failed to process

https://doi.org/10.1038/s41592-026-03007-y

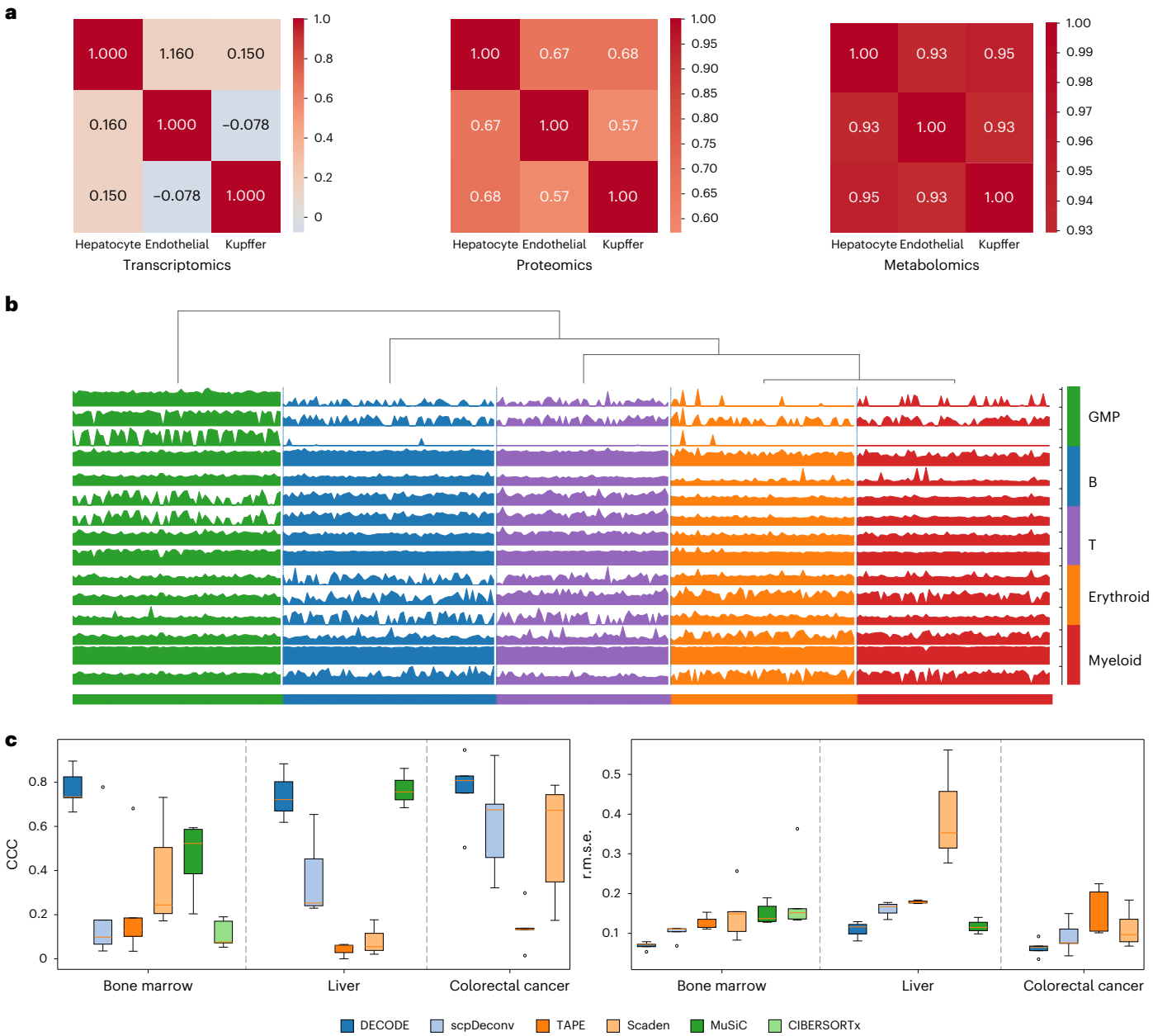

Fig. 3 | Overview of DECODE's metabolomics data deconvolution performance. a, Heatmaps showing the Kendall similarity of cell types in mouse liver based on single-cell transcriptomic, proteomic and metabolomic data, respectively. **b**, Abundance of differential metabolites in bone marrow among different cell types. The vertical axis represents the abundance of metabolites, with each row corresponding to a specific metabolite. The color scale on the right side represents the abundance of differentially expressed metabolites of different cell types. Abundance of differential metabolites in other two datasets are shown in Supplementary Fig. 9a,b. GMP: granulocyte-monocyte progenitor. **c**, CCC (left) and r.m.s.e. (right) values for DECODE, scpDeconv, TAPE, Scaden, MuSiC and CIBERSORTx using the bone marrow ($n = 5$ cell types), liver ($n = 3$ cell types) and colorectal cancer ($n = 5$ cell types) datasets. The Pearson's r evaluation metrics are shown in Supplementary Fig. 9c. The comparison results of all metrics with spatial transcriptomics methods are shown in Supplementary Fig. 8.

datasets 2 and 3, thus, the performance of CIBERSORTx[6] is not reported for these two datasets.

Overall, DECODE is a deconvolution method well-suited to multi-omics data across diverse cell states.

Precise deconvolution with incomplete single-cell profiles

Deconvolution performance depends strongly on the single-cell reference, yet building a comprehensive tissue reference is difficult. First, some cell populations are inevitably lost during single-cell library preparation. Second, cell dissociation can introduce perturbations—such as enzymatic digestion that particularly affects podocytes and intercalated cells—leading to the loss of specific cell types[47]. Finally, single-cell proteomics and metabolomics are mainly mass spectrometry-based and typically use in vitro cell lines, so rare or specialized cell types are often absent. These discrepancies in cell-type composition between single-cell references and tissue data demand tailored algorithmic solutions. Although methods such as Scaden[16] have examined how unknown cell types affect deconvolution stability, they do not resolve this problem. DECODE tackles incomplete cell-type coverage by incorporating a stage 3 denoiser and applying contrastive learning to separate noise.

In the following section, DECODE's deconvolution performance is systematically investigated by incrementally introducing unknown cell types into the test data. The datasets include the human lung transcriptomics dataset[21] (adding neutrophils for testing), the human breast

https://doi.org/10.1038/s41592-026-03007-y

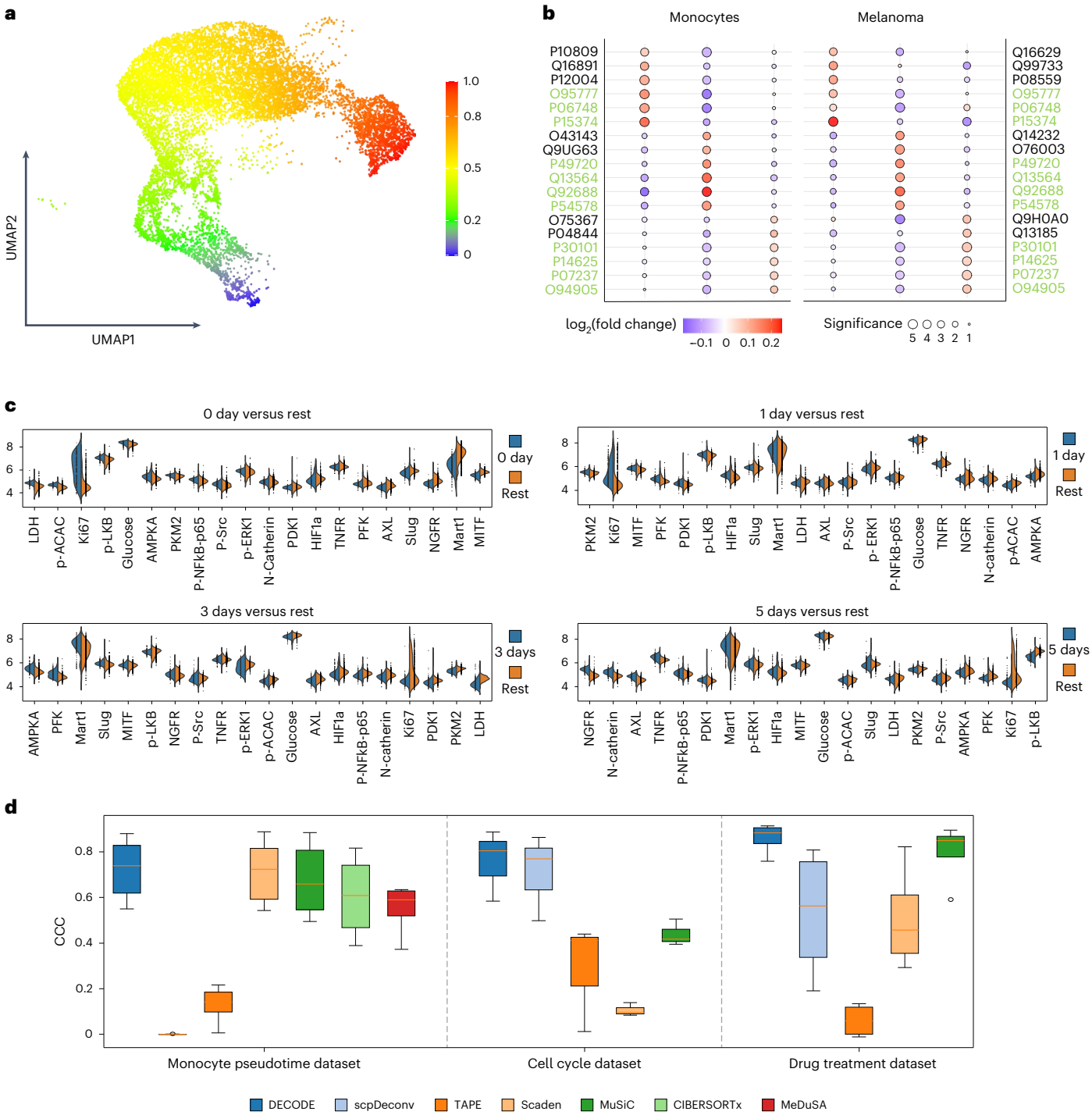

Fig. 4 | Overview of DECODE's cell state deconvolution performance across different omics data. a, Uniform manifold approximation and projection (UMAP) diagram of the pseudotime of monocytes. **b**, Results of differentially expressed proteins (left: monocytes; right: melanoma cells), showing the differences in $\log_2$-fold changes for the top six proteins across cell cycle phases and cell types, respectively. The proteins marked in green represent the differentially expressed proteins that overlap between the two cell types. The red and blue bubbles represent positive and negative $\log_2$-fold change values, respectively. The greater the absolute value of the $\log_2$-fold change is, the darker the color. The size of the bubbles indicates the log (*P* value), scaled between 1 and 5; larger bubbles represent more significant differences between the 2 groups. **c**, Violin plots of proteins and metabolites differences in cells at 0 days (top left), 1 day (top right), 3 days (bottom left) and 5 days (bottom right) compared with cells from other time points. The blue color represents the feature abundance for the current day, whereas the orange color represents the feature abundance for cells from other days. The *x* axis represents different metabolites and/or proteins, and the *y* axis represents the relative abundance of the metabolites and/or proteins. **d**, CCC values for DECODE, scpDeconv, TAPE, Scaden, CIBERSORTx and MeDuSA across the monocyte pseudotime dataset ($n = 10$ cell states, left), the cell cycle dataset ($n = 3$ cell states, center) and the drug treatment dataset ($n = 4$ cell states, right). The r.m.s.e. and Pearson's *r* evaluation metrics are shown in Supplementary Fig. 13. Statistical analysis for **b** was performed using a two-sided Wilcoxon rank-sum test, with multiple comparisons adjusted using the Benjamini–Hochberg method.

proteomics dataset[22] (without vascular lymphatic cells for training) and the mouse bone marrow metabolomics dataset[36] (adding hematopoietic stem cells for testing).

Moreover, we applied three types of perturbation to the test data to evaluate method performance under noise and missing values: (1) we multiplied a subset of features by random factors between 0.9 and 1.1, gradually increasing the proportion of perturbed features to mimic mild experimental fluctuations; (2) we multiplied some features by 0.8 or 1.2, progressively expanding the perturbed set to simulate larger systematic biases from instrument calibration and (3) we randomly set a fraction of features to zero, gradually increasing the proportion to mimic reduced detection sensitivity or feature loss. For proteomics and metabolomics, we progressively reduced the available features and retrained the models rather than simply zeroing test features, because these low-dimensional, continuous data are highly sensitive to missing values. In contrast, high-dimensional, sparse transcriptomic data are more tolerant to missingness. Thus, in scenarios mimicking failed detection of specific features, we removed those features and retrained the models to obtain a more realistic and robust evaluation.

This section evaluates the performance of DECODE under four perturbation scenarios (Fig. 5a,c,e and Supplementary Figs. 14 and 15a) using three types of omics dataset—transcriptomics, proteomics and metabolomics—across three evaluation metrics. The results demonstrate that DECODE outperforms other methods in most comparisons, particularly in terms of the CCC evaluation metric, with only a few cases of suboptimal performance observed in transcriptomic data. Notably, all methods except DECODE exhibit unusable performance on metabolomic data. This highlights, on one hand, the unique effectiveness of DECODE in metabolome deconvolution and, on the other hand, the intrinsic challenges of metabolomic data, which is more sensitive to noise due to its limited feature availability and low specificity to cell types, making deconvolution significantly more difficult. The results indicate that different perturbations within the same omics data show substantial variability, and that the same perturbation manifests differently across omics data. A detailed discussion of this phenomenon is provided in the section 'Interpretation of data performance in noisy environments' of the Supplementary Information.

To further assess stability, we calculated the coefficient of variation (CV) for each evaluation metric across datasets and perturbation scenarios (Fig. 5b,d and Supplementary Fig. 5b). Because stability is not informative when accuracy is very low, we restricted this analysis to DECODE and other competitive methods across omics types. In transcriptomics, Scaden, scpDeconv and RCTD perform comparably to DECODE and show lower CVs for all three metrics in all four scenarios (Fig. 5b). In proteomics, Scaden and scpDeconv are again competitive and usually have lower CVs than DECODE, except at noise levels 0.8 or 1.2 (Fig. 5d), yet DECODE still achieves the highest overall deconvolution accuracy in all proteomic scenarios. In metabolomics, only DECODE yields usable results, making further stability comparison meaningless (CVs shown in Supplementary Fig. 15b).

In summary, DECODE is well-suited for real-world deconvolution applications, as it robustly estimates cell proportions in tissue samples under various perturbations.

High consistency across different omics datasets

Applying deconvolution to cohort studies enables comparison of cell-type abundances across disease states. Many studies profile the same tissue with different omics to discover biomarkers under varying conditions[48–50]. However, using different deconvolution tools for each omics layer can produce inconsistent results due to method-specific error sensitivities. A unified deconvolution approach minimizes error accumulation when comparing cell abundances across omics and reduces the effort needed to evaluate multiple tools. Therefore, assessing the consistency of DECODE's performance across omics is essential for fully leveraging multiomics data in multi-cohort studies.

We used a peripheral blood mononuclear cell (PBMC) CITE-seq dataset from ref. 51 in this analysis. This dataset consists of transcriptomic and surface protein data from 43,791 single cells and includes five cell types: CD4 T cells, CD8 T cells, B cells, natural killer (NK) cells and myeloid cells. Transcriptomic and proteomic pseudocohorts were constructed using the data of donor HS5, whereas the training samples were sampled from donor HS1. The similarity and accuracy of DECODE's deconvolution results in these two pseudocohorts would reflect the consistency of its deconvolution ability for different omics data.

DECODE outperforms all other methods (Fig. 6a and Supplementary Fig. 17a,b); sample-wise CCC values across the two omics and a one-tailed Wilcoxon test with Bonferroni correction (Supplementary Table 5) confirm a significant advantage, with nearly identical performance between omics. Deconvolution results from 1,000 pseudocohort test samples are highly consistent (Supplementary Fig. 17c); for visualization, 50 samples are shown in Fig. 6b, where DECODE yields highly similar cell abundances across omics. Consistency between omics, quantified by KL divergence and Spearman correlation (Fig. 6c,d), shows predictions clustered at low KL and high correlation, indicating that DECODE provides consistent, robust cell-abundance estimates for cross-omics cohort integration.

Based on the above results, we applied DECODE to conduct a comprehensive analysis on multiomics cohorts of patients with human breast cancer and mouse liver. (1) A total of 238 samples covering transcriptomics and proteomics were collected from four studies[52–55]. After integration, the samples were divided into three phenotypic groups: nonmetastatic breast cancer tissue ($n = 99$), breast cancer tissue with brain metastasis ($n = 45$) and breast cancer brain metastatic lesions ($n = 94$). We focused on nine cell types, including B cells, cancer-associated fibroblasts, cancer epithelial cells, endothelial cells, myeloid cells, NK cells, perivascular-like (PVL) cells, Plasmablasts and T cells, using single-cell multiomics data from ref. 22 as a reference. (2) In 10 studies, a total of 285 samples covering transcriptomics[56–62], proteomics[63] and metabolomics[64] were collected. After integration, the samples were divided into five groups: standard diet (chow, $n = 208$), high-fat diet (HFD, $n = 63$), HFD with TLC-065 treatment (HFD-TLC, $n = 4$), nonalcoholic steatohepatitis (NASH, $n = 6$) and western diet with alcohol (WDA, $n = 4$). We focused on three cell types: parenchymal cells (hepatocytes) and nonparenchymal cells (endothelial cells and Kupffer cells). Transcriptomic, proteomic and metabolomic single-cell data from refs. 35,40,41 were used as references.

The three breast cancer groups show distinct cellular compositions (Fig. 6e). Nonmetastatic primary tumors have higher T cell and PVL cell abundances and lower B cell abundance than the other two groups: T cells are 1.14-fold higher than in metastatic tumors ($P = 1.17 \times 10^{-2}$) and 1.23-fold higher than in brain metastases ($P = 4.19 \times 10^{-6}$); PVL cells are 1.48-fold and 1.64-fold higher, respectively ($P = 3.38 \times 10^{-9}$, $P = 2.85 \times 10^{-20}$). B cells are 1.70-fold lower than in metastatic tumors ($P = 6.79 \times 10^{-21}$) and 1.47-fold lower than in brain metastases ($P = 3.28 \times 10^{-11}$). Plasmablasts are 1.14-fold higher than in metastatic tumors ($P = 1.18 \times 10^{-2}$) but 1.36-fold lower than in brain metastases ($P = 2.16 \times 10^{-5}$). These results indicate the protective roles for T cells and PVL cells in breast cancer and a possible association between elevated B cell abundance and metastatic progression.

Across multiple mouse liver cohorts, the recovery of cell abundance in the chow group using various omics data showed strong consistency and aligned well with established consensus estimates—namely, that hepatocytes account for nearly 70%[65] and Kupffer cells for 15%[66] of liver cells (Supplementary Fig. 19a). These findings indicate that the deconvolution results obtained using DECODE across multiomics datasets are highly reliable. Compared to mice on a healthy diet (Fig. 6f,g), Kupffer cells exhibited a gradual increasing trend: a 1.23-fold increase in the WDA group ($P = 1.38 \times 10^{-2}$) and a 1.61-fold increase in the NASH group ($P = 2.06 \times 10^{-13}$), while remaining stable in the HFD group. This suggests that the combination of sugar, cholesterol and

https://doi.org/10.1038/s41592-026-03007-y

Fig. 5 | Evaluation of robustness under four perturbation scenarios. a, The CCC line plot of the 12 methods on the lung transcriptomics dataset in four disturbance scenarios (unknown cells, random noise (0.9–1.1), random noise (0.8 or 1.2), and random dropout). **b**, The CV of CCC, r.m.s.e. and Pearson's *r* for 12 methods on the lung transcriptomics dataset under four disturbance scenarios. To facilitate the comparison of the CVs across different methods, we transform the CV to the range of [0.1, 1], where a higher value indicates better stability (detailed calculation methods are provided in the section 'Transformation of coefficient of variation' of the Supplementary Information; original CV values are listed in Supplementary Tables 6–8). **c**, The CCC line plot of six methods under four disturbance scenarios (unknown cells, random noise (0.9–1.1), random noise (0.8 or 1.2) and random feature deletion) on the breast proteomics dataset. **d**, The CV of CCC, r.m.s.e. and Pearson's *r* for the six methods on the breast proteomics dataset under four disturbance scenarios. **e**, The CCC line plot of six methods four disturbance scenarios (unknown cells, random noise (0.9–1.1), random noise (0.8 or 1.2) and random feature deletion) on the bone marrow metabolomics dataset.

https://doi.org/10.1038/s41592-026-03007-y

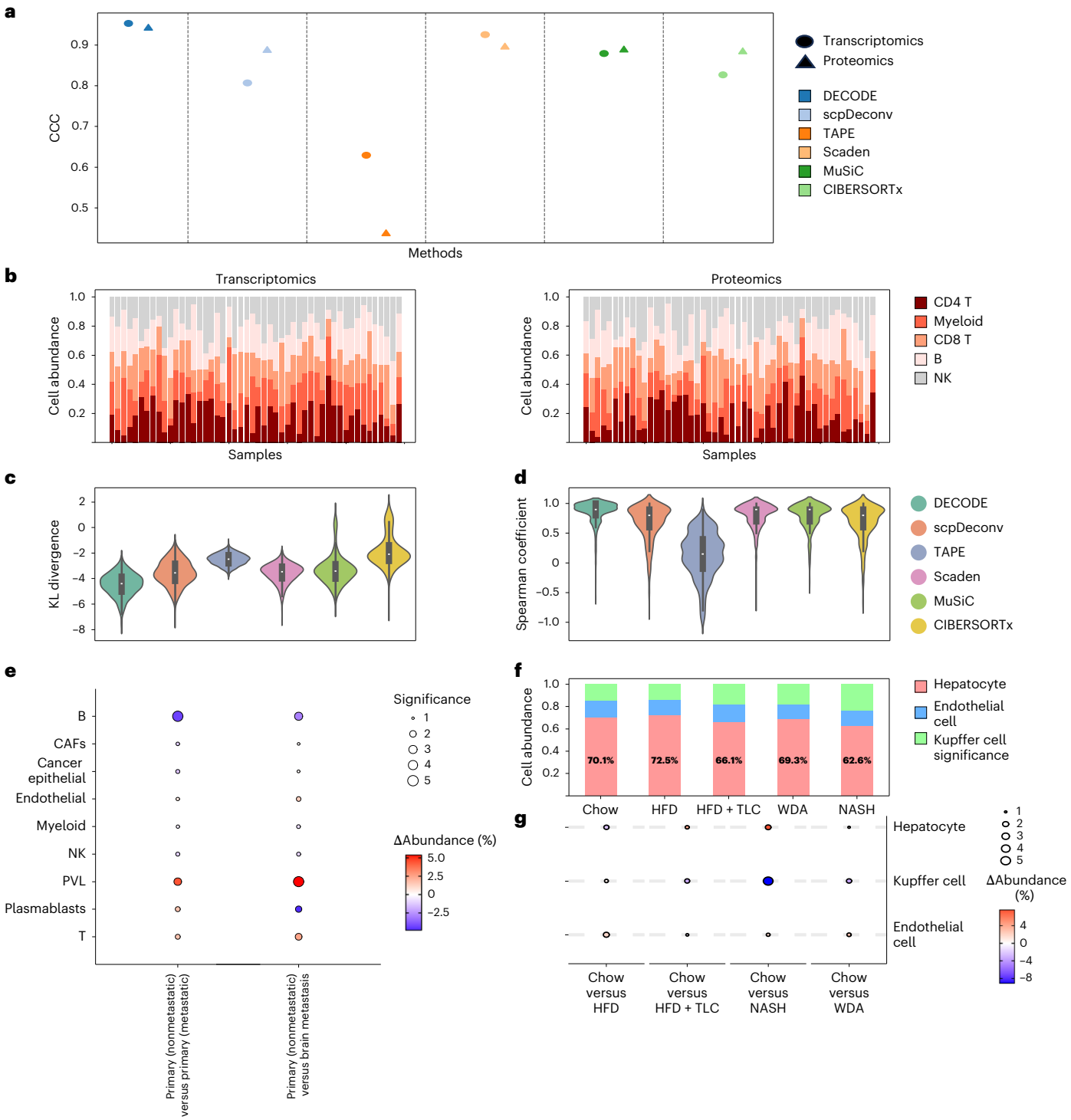

Fig. 6 | Application of DECODE to pseudo and real multiomics cohorts. **a**, CCC value of DECODE and other methods across two omics pseudocohorts. The vertical axis represents the CCC value, and the horizontal axis represents different methods. Triangles represent the deconvolution result of transcriptomics pseudocohort, and circles represent that of proteomics pseudocohort. The r.m.s.e. and Pearson's *r* results are shown in Supplementary Fig. 17. **b**, DECODE's deconvolution results for 50 test samples ($n = 5$ cell types). The vertical axis represents the cell abundance, whereas the horizontal axis represents the test sample. The left and right sides show the deconvolution results using transcriptomic and proteomic data, respectively. **c**,**d**, Violin plots ($n = 1{,}000$ test samples) of the log (KL divergence) (**c**) and Spearman coefficient (**d**) between the deconvolution results of all pseudocohort samples. A smaller log (KL divergence) indicates better consistency, while a Spearman coefficient closer to 1 indicates a stronger correlation. **e**, Bubble plot displaying the differences in cellular abundance and *P* values between different phenotypes of breast cancer. Red and blue indicate positive and negative cellular abundance differences, respectively. The size of the bubbles represents the $-\log_{10}(P\text{ value})$ scaled between 1 and 5, with larger bubbles indicating more significant differences between the two groups. CAFs: cancer-associated fibroblasts. **f**, Average cell abundance in different phenotypic mice. **g**, Bubble plot displaying the differences in cellular abundance and *P* values between different phenotypes of liver. Violin plots display the distribution of the values, with the width of each violin proportional to the data density. The inner box shows the median (center line) and the 25th–75th percentiles (bounds of the box). The number of test samples indicates how many data points contribute to each violin. Statistical analysis for **e** and **g** was performed using a two-sided unpaired *t*-test, and no adjustment for multiple comparisons was applied.

alcohol in WDA is more likely to provoke an inflammatory response and activate Kupffer cells than high-fat intake alone. Although hepatocyte abundance decreased in the HFD-TLC and WDA groups, these changes were not statistically significant. In contrast, the NASH group showed a significant 1.12-fold reduction ($P = 1.87 \times 10^{-3}$). Hepatocyte abundance slightly increased in the HFD group (1.03-fold, $P = 1.48 \times 10^{-2}$).

In summary, DECODE demonstrates high accuracy and consistency in deconvolution across different omics data, making it a powerful tool for studying cell proportion changes across various cohorts.

Discussion

DECODE is a deconvolution algorithm capable of dealing with transcriptomics, proteomics and metabolomics data, addressing a critical gap in metabolomics deconvolution. This advancement signifies an important milestone in multiomics data analysis. DECODE is a highly generalizable multiomics deconvolution method. Its stage 2 uses transferred adversarial training to align diverse omics data across platforms, health states and sample types, effectively removing batch effects in multiple settings (Supplementary Fig. 16a–d). Stage 3 combines contrastive learning and self-attention to correct measurement biases in tissue samples and reconcile perturbations between tissue data and single-cell references, enabling reconstruction of purified features from noisy inputs (section 'Noise removal section under the ablation experiment' in the Supplementary Information). These modules together confer strong robustness in recovering both cell types and cell states, even in metabolomics with subtle intercellular differences, and their individual contributions were validated by dividing DECODE into three components (Supplementary Table 9). Experiments on CITE-seq pseudocohorts further demonstrate consistent performance across omics, establishing DECODE as a powerful tool for multiomics analysis at the cellular level and for bridging translational and clinical applications.

DECODE revealed significant shifts in cell-type proportions across breast cancer stages using multiomics cohort data. Between nonmetastatic carcinoma in situ, metastatic carcinoma in situ and brain metastases, plasmablasts showed inconsistent trends but were generally enriched in nonmetastatic carcinoma in situ relative to metastatic lesions, suggesting an early protective immune role as precursors of plasma cells that rapidly produce antibodies during initial tumorigenesis[67]. However, DECODE also indicates plasmablast enrichment in brain metastases, consistent with ref. 68, that proposed this pattern may result from the brain's unique immune 'screening effect' and tumor-driven 'active regulation', although further validation is needed. Meanwhile, DECODE showed significantly increased T cells and markedly reduced B cells in nonmetastatic carcinoma in situ. Previous studies have consistently demonstrated that intratumoral T cell enrichment strongly correlates with favorable prognosis[69], which aligns with our observations. By contrast, B cells show pronounced heterogeneity in the tumor microenvironment, with roles that vary depending on subtype and context, ranging from anti-tumor activity to tumor-promoting effects[70]. To further dissect these patterns, we performed subtype-level deconvolution of T and B cells (Supplementary Fig. 18). Previous studies have shown that high infiltration of $CD8^+$ T cells is strongly associated with favorable prognosis, while $CD4^+$ T cells play an essential role in supporting $CD8^+$ activation, expansion and maintenance[71]. In line with these findings, our analysis revealed that $CD4^+$ T cells, $CD8^+$ T cells and cycling T cells all show increase trend in nonmetastatic carcinoma in situ. In contrast, the increase in B cells in metastatic carcinoma in situ and brain metastases was primarily driven by naive B cells, whereas memory B cells showed no significant enrichment. This is consistent with ref. 72 that reported patients with late-stage tumor frequently display elevated proportions of naive B cells, potentially reflecting impaired immune function and enhanced metastatic potential, although the causal relationship requires further experimental confirmation. Finally, DECODE also identified that PVL cells were enriched in nonmetastatic breast cancer tissues, in agreement with previous studies reporting that PVL deficiency is closely linked to metastasis and poor prognosis[73,74], a phenomenon that may result from immune evasion.

When DECODE was applied to the mice liver cohorts of three omics, results highly consistent with previous studies were obtained. With the progression of NASH, the number of hepatocytes gradually decreases; conversely, under HFD conditions, the number of hepatocytes increases. This may be due to hepatocyte apoptosis in the context of NASH[75], but a HFD can lead to oxidative stress. Higher levels of oxidative stress are associated with hepatocyte proliferation[76]. In addition, the number of Kupffer cells also increases under NASH and WDA conditions. This may be because strong inflammation is generated in NASH and Western alcohol diets, leading to the recruitment of more Kupffer cells in the liver[77,78]. One study[60] noted that female mice exhibit a stronger inflammatory response in NASH, which was validated in DECODE's analysis: the number of Kupffer cells in female NASH mice was higher than that in the control group and male NASH mice (Supplementary Fig. 19b). These findings further support the significant role of gender factors in the progression of NASH.

However, DECODE has several limitations. First, training requires introducing artificial noise cells into simulated tissues by combining features from different single-cell types, which incurs extra computational cost dependent on single-cell feature dimensionality (section 'Artificial noise cell generation' in the Supplementary Information). As this step is performed only once per deconvolution task, the fixed cost is unlikely to impede clinical use. Second, current single-cell metabolomics technologies yield limited datasets, preventing a comprehensive evaluation of DECODE's robustness, and while large blood metabolite cohorts exist, single-cell metabolomics has focused mainly on other tissues, constraining DECODE's applicability in cohort-based metabolomics. Future work could enhance DECODE by adding a dedicated spatial module to better exploit spatial transcriptomics data and extending it to additional omics layers, such as DNA methylation, to further improve its adaptability for integrative multiomics analysis.

In summary, DECODE serves as an effective tool for estimating proportion of cell types and states on three omics data. It provides a widely applicable framework that is able to fully exploit the vast amount of existing multiomics tissue-level data that offer new insights and methods for advancing biomedical research.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-026-03007-y.

References

1. Lu, H. et al. Dissecting the impact of maternal androgen exposure on developmental programming through targeting the androgen receptor. *Adv. Sci.* **11**, 2309429 (2024).
2. Cai, Y. et al. An atlas of genetic effects on cellular composition of the tumor microenvironment. *Nat. Immunol.* **25**, 1959–1975 (2024).
3. Kim-Hellmuth, S. et al. Cell type–specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
4. Garmire, L. X. et al. Challenges and perspectives in computational deconvolution of genomics data. *Nat. Methods* **21**, 391–400 (2024).
5. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
6. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

7. Cable, D. M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**, 517–526 (2022).
8. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **49**, e50 (2021).
9. Wang, F. et al. Deep domain adversarial neural network for the deconvolution of cell type mixtures in tissue proteome profiling. *Nat. Mach. Intell.* **5**, 1236–1249 (2023).
10. Schmidt, D. R. et al. Metabolomics in cancer research and emerging applications in clinical oncology. *CA Cancer J. Clin.* **71**, 333–358 (2021).
11. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).
12. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).
13. Goh, W. W. B., Wang, W. & Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* **35**, 498–507 (2017).
14. Vaswani, A. et al. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)* https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (2017).
15. Chen, Y. et al. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nat. Commun.* **13**, 6735 (2022).
16. Menden, K. et al. Deep learning–based cell composition analysis from tissue expression profiles. *Sci. Adv.* **6**, eaba2619 (2020).
17. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
18. Biancalani, T. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).
19. Charytonowicz, D., Brody, R. & Sebra, R. Interpretable and context-free deconvolution of multi-scale whole transcriptomic data with UniCell deconvolve. *Nat. Commun.* **14**, 1350 (2023).
20. Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).
21. Vieira Braga, F. A. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
22. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
23. Gray, G. K. et al. A human breast atlas integrating single-cell proteomics and transcriptomics. *Dev. Cell* **57**, 1400–1420 (2022).
24. Oppenländer, L. et al. Vertical sleeve gastrectomy triggers fast β-cell recovery upon overt diabetes. *Mol. Metab.* **54**, 101330 (2021).
25. Tritschler, S. et al. A transcriptional cross species map of pancreatic islet cells. *Mol. Metab.* **66**, 101595 (2022).
26. Woo, J. et al. High-throughput and high-efficiency sample preparation for single-cell proteomics using a nested nanowell chip. *Nat. Commun.* **12**, 6246 (2021).
27. Dou, M. et al. High-throughput single cell proteomics enabled by multiplex isobaric labeling in a nanodroplet sample preparation platform. *Anal. Chem.* **91**, 13119–13127 (2019).
28. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
29. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
30. Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
31. Cowan, C. S. et al. Cell types of the human retina and its organoids at single-cell resolution. *Cell* **182**, 1623–1640 (2020).
32. Monaco, G. et al. RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep.* **26**, 1627–1640 (2019).
33. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
34. Zimmermann, M. T. et al. System-wide associations between DNA-methylation, gene expression, and humoral immune response to influenza vaccination. *PLoS ONE* **11**, e0152034 (2016).
35. Yuan, Z. et al. SEAM is a spatial single nuclear metabolomics method for dissecting tissue microenvironment. *Nat. Methods* **18**, 1223–1232 (2021).
36. Cao, J. et al. Deciphering the metabolic heterogeneity of hematopoietic stem cells with single-cell resolution. *Cell Metab.* **36**, 209–221 (2024).
37. Nunes, J. B. et al. Integration of mass cytometry and mass spectrometry imaging for spatially resolved single-cell metabolic profiling. *Nat. Methods* **21**, 1796–1800 (2024).
38. Slavov, N. Unpicking the proteome in single cells. *Science* **367**, 512–513 (2020).
39. Petelski, A. A. et al. Multiplexed single-cell proteomics using SCoPE2. *Nat. Protoc.* **16**, 5398–5425 (2021).
40. Su, Q. et al. Single-cell RNA transcriptome landscape of hepatocytes and non-parenchymal cells in healthy and NAFLD mouse liver. *iScience* **24**, 103233 (2021).
41. Guilliams, M. et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell* **185**, 379–396 (2022).
42. van der Leun, A. M., Thommen, D. S. & Schumacher, T. N. CD8+ T cell states in human cancer: insights from single-cell analysis. *Nat. Rev. Cancer* **20**, 218–232 (2020).
43. Hara, T. et al. Interactions between cancer cells and immune cells drive transitions to mesenchymal-like states in glioblastoma. *Cancer Cell* **39**, 779–792 (2021).
44. Song, L., Sun, X., Qi, T. & Yang, J. Mixed model-based deconvolution of cell-state abundances (MeDuSA) along a one-dimensional trajectory. *Nat. Comput. Sci.* **3**, 630–643 (2023).
45. Leduc, A., Huffman, R. G., Cantlon, J., Khan, S. & Slavov, N. Exploring functional protein covariation across single cells using nPOP. *Genome Biol.* **23**, 261 (2022).
46. Su, Y. et al. Multi-omic single-cell snapshots reveal multiple independent trajectories to drug tolerance in a melanoma cell line. *Nat. Commun.* **11**, 2345 (2020).
47. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol.* **30**, 23–32 (2019).
48. Kaoutari, A. E. et al. Metabolomic profiling of pancreatic adenocarcinoma reveals key features driving clinical outcome and drug resistance. *eBioMed.* **66**, 103332 (2021).
49. Beck, L. et al. Clinical proteomics of metastatic melanoma reveals profiles of organ specificity and treatment resistance. *Clin. Cancer Res.* **27**, 2074–2086 (2021).
50. Brown, C. C. et al. Transcriptional basis of mouse and human dendritic cell heterogeneity. *Cell* **179**, 846–863 (2019).
51. He, J. Y. et al. Dysregulation of CD4(+) and CD8(+) resident memory T, myeloid, and stromal cells in steroid-experienced, checkpoint inhibitor colitis. *J. Immunother Cancer* **12**, e008628 (2024).
52. Krug, K. et al. Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* **183**, 1436–1456 (2020).
53. Mendoza-Valderrey, A. et al. Immunogenomics and spatial proteomic mapping highlight distinct neuro-immune architectures in melanoma vs. non-melanoma-derived brain metastasis. *BJC Rep.* **2**, 38 (2024).

54. Cosgrove, N. et al. Mapping molecular subtype specific alterations in breast cancer brain metastases identifies clinically relevant vulnerabilities. *Nat. Commun.* **13**, 514 (2022).
55. Menck, K. et al. WNT11/ROR2 signaling is associated with tumor invasion and poor survival in breast cancer. *J. Exp. Clin. Cancer Res.* **40**, 395 (2021).
56. Mononen, J. et al. Genetic variation is a key determinant of chromatin accessibility and drives differences in the regulatory landscape of C57BL/6J and 129S1/SvImJ mice. *Nucleic Acids Res.* **52**, 2904–2923 (2024).
57. Small, L. et al. Seasonal light hours modulate peripheral clocks and energy metabolism in mice. *Cell Metab.* **35**, 1722–1735.e1725 (2023).
58. Wang, C. et al. METTL3-mediated chromatin contacts promote stress granule phase separation through metabolic reprogramming during senescence. *Nat. Commun.* **15**, 5410 (2024).
59. Liu, Y. J. et al. ACMSD inhibition corrects fibrosis, inflammation, and DNA damage in MASLD/MASH. *J. Hepatol.* **82**, 174–188 (2025).
60. Arivazhagan, L. et al. Sex differences in murine MASH induced by a fructose-palmitate-cholesterol-enriched diet. *JHEP Rep.* **7**, 101222 (2025).
61. Choi, J., Smith, D. M., Scafidi, S., Riddle, R. C. & Wolfgang, M. J. Carnitine palmitoyltransferase 1 facilitates fatty acid oxidation in a non-cell-autonomous manner. *Cell Rep.* **43**, 115006 (2024).
62. Sasaki, K. et al. Kupffer cell diversity maintains liver function in alcohol-associated liver disease. *Hepatology* **81**, 870–887 (2025).
63. Puopolo, T. et al. Uncovering the anti-inflammatory mechanisms of phenolic-enriched maple syrup extract in lipopolysaccharide-induced peritonitis in mice: insights from data-independent acquisition proteomics analysis. *Food Funct.* **14**, 6690–6706 (2023).
64. Ghazalpour, A. et al. Genetic regulation of mouse liver metabolite levels. *Mol. Syst. Biol.* **10**, 730 (2014).
65. Meyer, J., Lacotte, S., Morel, P., Gonelle-Gispert, C. & Bühler, L. An optimized method for mouse liver sinusoidal endothelial cell isolation. *Exp. Cell Res.* **349**, 291–301 (2016).
66. Kolios, G., Valatas, V. & Kouroumalis, E. Role of Kupffer cells in the pathogenesis of liver disease. *World J. Gastroenterol.* **12**, 7413–7420 (2006).
67. Downs-Canner, S. M., Meier, J., Vincent, B. G. & Serody, J. S. B cell function in the tumor microenvironment. *Annu Rev. Immunol.* **40**, 169–193 (2022).
68. Lu, W. C. et al. Genomic landscape of the immune microenvironments of brain metastases in breast cancer. *J. Transl. Med.* **18**, 327 (2020).
69. Zebley, C. C., Zehn, D., Gottschalk, S. & Chi, H. T cell dysfunction and therapeutic intervention in cancer. *Nat. Immunol.* **25**, 1344–1354 (2024).
70. Sharonov, G. V., Serebrovskaya, E. O., Yuzhakova, D. V., Britanova, O. V. & Chudakov, D. M. B cells, plasma cells and antibody repertoires in the tumour microenvironment. *Nat. Rev. Immunol.* **20**, 294–307 (2020).
71. Speiser, D. E., Chijioke, O., Schaeuble, K. & Münz, C. CD4(+) T cells in cancer. *Nat. Cancer* **4**, 317–329 (2023).
72. Zhang, E. et al. Roles and mechanisms of tumour-infiltrating B cells in human cancer: a new force in immunotherapy. *Biomark. Res.* **11**, 28 (2023).
73. Hanahan, D. & Coussens, L. M. Accessories to the crime: functions of cells recruited to the tumor microenvironment. *Cancer Cell* **21**, 309–322 (2012).
74. Bichsel, C. A. et al. Increased PD-L1 expression and IL-6 secretion characterize human lung tumor-derived perivascular-like cells that promote vascular leakage in a perfusable microvasculature model. *Sci. Rep.* **7**, 10636 (2017).
75. Pantano, L. et al. Molecular characterization and cell type composition deconvolution of fibrosis in NAFLD. *Sci. Rep.* **11**, 18045 (2021).
76. Wang, Y., Ausman, L. M., Greenberg, A. S., Russell, R. M. & Wang, X. D. Nonalcoholic steatohepatitis induced by a high-fat diet promotes diethylnitrosamine-initiated early hepatocarcinogenesis in rats. *Int J. Cancer* **124**, 540–546 (2009).
77. Wang, X. et al. Characteristic gene expression in the liver monocyte-macrophage-DC system is associated with the progression of fibrosis in NASH. *Front Immunol.* **14**, 1098056 (2023).
78. Duly, A. M. P. et al. Effect of multiple binge alcohol on diet-induced liver injury in a mouse model of obesity. *Nutr. Diab.* **5**, e154 (2015).

Methods

Data preprocessing

First, single-cell matrices for transcriptomics, proteomics and metabolomics are generated from the raw data. The single-cell transcriptomics data undergo additional preprocessing. Following the Scanpy[79] workflow, genes expressed in fewer than three cells and cells with fewer than 200 genes or total counts are filtered out. In addition, cells with mitochondrial gene expression exceeding 20% are removed. Next, total counts per cell are normalized to 10,000 reads and log-transformed. Pseudotissue samples for all three omics types are then generated by sampling and aggregating cells from the processed single-cell matrices. Each pseudotissue dataset was normalized to the range [0, 1] by dividing all elements by its respective largest eigenvalue. The real bulk data are processed using the same normalization and scaling procedures as the pseudotissue data.

Datasets

This section introduces the specific datasets used in the DECODE study, along with some hyperparameter settings for each dataset in the experiments, which can be found in the Supplementary Table 4.

Data used in scenario 1

In scenario 1, one research group[21] constructed a lung single-cell dataset consisting of 12,971 cells. The authors provided a single-cell dataset in Scanpy[79] format, with detailed annotations of cell types and donor information. For each gene, the dataset also indicates whether it is a highly variable gene; only the highly variable genes (n = 3,346) were used in the DECODE experiments. Two donors, 296C (female) and 302C (male), with similar health conditions but different sexes, were selected. Data from donor 296C were used to generate pseudotissue data for training, whereas data from donor 302C were used to generate pseudotissue data for testing. The experiment primarily focused on type 2 alveolar cells, luminal macrophages, dendritic cells and fibroblasts.

Data used in scenario 2

Scenario 2 uses the human breast single-cell transcriptomic data containing 100,064 cells provided in ref. 22. Data from two donors, CID3586 (estrogen receptor-positive breast cancer) and CID3921 (estrogen receptor-negative breast cancer), were selected for training and testing, respectively. Highly variable genes (n = 2,525) for each donor were identified using Scanpy. The analysis focused on six overlapping cell types present in both donors: B cells, T cells, cancer-associated fibroblasts, myeloid cells, endothelial cells and PVL cells.

Data used in scenario 3

In scenario 3, the proteomics-based single-cell data of the breast generated in ref. 23 using cytometry by time of flight, comprising a total of 751,970 cells, were used. The dataset includes 34 high-quality protein features that effectively distinguish six cell types: alveolar, hormone-sensing, basal, fibroblast, vascular lymphatic, and immune cells. The data from donors in the dataset were stratified on the basis of menopausal status, with data from a premenopausal donor (B1H35) used for training and those from a postmenopausal donor (B1H32) used for testing.

Data used in scenario 4

In scenario 4, four datasets are used to form two experimental groups: (1) the groups in refs. 26 and [27] measured single-cell data from the same three murine cell lines—macrophage (RAW 264.7), airway epithelial (C10) and endothelial (SVEC)—using two measurement techniques and obtained 108 and 72 cells, respectively. The ref. 26 dataset contains 1,437 protein features, whereas the ref. 27 dataset contains 1,032 protein features, with 762 overlapping features. For the DECODE experiments, only the overlapping features were used. (2) The mouse islet datasets measured by refs. 24 and [25] were integrated into the GSE211799 dataset. The Scanpy-format data were obtained from this source using the Gene Expression Omnibus (GEO) accession number to locate the datasets provided by the two studies mentioned above. The data contain 31,706 available transcriptomic features, which were reduced to 2,558 transcriptomic features after highly variable gene selection. Among these, single-cell data from the VSG_MUC13634 donor were used to construct pseudotissues for mixed training, while single-cell data from the STZ_G1 donor were used to construct pseudotissues for mixed testing.

Data used in scenario 5

In scenario 5, two experimental groups are formed using three datasets: (1) the data come from the ref. 80 dataset 10, which includes spatial transcriptomics data from the primary visual cortex of mice obtained using STARmap technology and single-cell RNA sequencing data based on Smart-seq2 technology provided by the Allen Brain Atlas. The STARmap data contains 882 transcriptomic features from 1,523 cells, while the Smart-seq2 data includes 34,042 transcriptomic features from 14,249 cells. The STARmap data are divided into a 750 μm × 750 μm grid, resulting in 189 spots data used for testing, while the Smart-seq2 single-cell data are used to mix the training samples. (2) High-resolution mouse hippocampal data obtained using Slide-seqV2 technology, loaded using the Squidpy package[30] are used. The raw dataset contains 17,733 transcriptomic features from 41,786 cells. After processing with the ucdselect model framework (Interpretable and context-free deconvolution of multi-scale whole transcriptomic data with UniCell deconvolve), the data are gridded into a 100 μm × 100 μm grid, resulting in 1,892 spots with 17,733 transcriptomic features. The ungridded data serve as a reference for conducting DECODE experiments on the spot.

Data used in scenario 6

In scenario 6, the Fovea single-cell sequencing data provided by ref. 31 were used as the mixed training set, while the Periphery single-cell data served as the mixed test set. These two datasets contain 34,723 and 19,768 cells, respectively, focusing primarily on the following 17 cell types: retinal pigment epithelial cell, retinal ganglion cell, amacrine cell, retinal rod cell, retinal cone cell, retina horizontal cell, rod bipolar cell, OFF-bipolar cell, Mueller cell, astrocyte, pericyte, endothelial cell of vascular tree, microglial cell, fibroblast, monocyte, NK cell and T cell. The original data contained 33,796 transcriptomic features and, after screening for highly variable genes, 3,789 features were retained for subsequent analysis.

Data used in scenario 7

The data used in scenario 7 were obtained from TAPE[15] experiments, where the cell types were grouped into six major categories: monocytes, CD4 T cells, CD8 T cells, NK cells, B cells and unknown cells. The single-cell data were used to train CIBERSORTx[6] and MuSiC[5]. Methods based on deep learning were trained using the premixed pseudotissue data provided in TAPE. Across these datasets, a total of 11,328 shared RNA features were identified, and DECODE selected only the highly variable genes (982) as input features.

Data used in metabolomics

In the metabolic data scenario, DECODE uses three datasets sourced from refs. 35–37, which provide mouse liver metabolomics data, mouse bone marrow metabolomics data and human colorectal cancer metabolomics data, respectively. These datasets contain 724, 1,428 and 57,078 cells, along with 244, 107 and 112 metabolic features, respectively. For the mouse liver metabolomics dataset, we assess the deconvolution performance with three cell types: hepatocytes, endothelial cells and Kupffer cells. In the mouse bone marrow metabolomics dataset, we investigate the deconvolution performance with five cell types: granulocyte-monocyte progenitors, B cells, T cells, myeloid cells and erythroid cells. In the human colorectal cancer dataset, we investigate the deconvolution performance with five cell types: cancer cells,

fibroblasts, B cells, myeloid cells and T cells. It is important to note that in the mouse bone marrow and liver datasets, each cell type was evenly divided into training and testing sets. The human colorectal cancer dataset uses a mixed training dataset composed of mismatch repair-deficient colorectal cancer samples and a mixed test dataset consisting of normal colorectal cancer samples.

Monocyte transcriptomics data with cell state

One research group[81] generated a human monocyte transcriptome dataset, which was annotated and analyzed via pseudotime analysis (slingshot) via MeDuSA[44]. The dataset was available for download on the official MeDuSA[44] website under the example section. This dataset includes 10,846 cells and 20,238 features, among which 2,000 features were labeled as highly variable genes. DECODE experiments used only these highly variable genes. The pseudotemporal annotations assigned each cell a continuous value between 0 and 1, divided into 10 intervals, with the cells categorized into 10 pseudotemporal states (1 to 10) on the basis of their interval values, which served as cell state labels. The cells from each state were evenly split into training and testing sets.

Melanoma drug response metabolomics data

One research group[46] investigated 20 proteins and metabolites related to signaling pathways, phenotypic regulators and metabolic regulators, in melanoma cells with the $BRAF^{V600E}$ mutation during drug exposure. This dataset is available in the source data of their publication. Each cell was observed at 4 points, representing states on day 0 (no drug exposure), day 1, day 3 and day 5 of drug exposure. These time points were used as state labels for each cell. The cells from each state were evenly split into training and testing sets.

Melanoma and monocyte cell cycle proteomics data

One study[45] measured protein changes in monocytes and melanoma cells during different phases of the cell division cycle. The dataset includes 2,844 protein features corresponding to 1,573 cells and three states (G1, S and G2 phases). These data are available in the Supplementary Information. In the experiments, pseudotissues generated from melanoma cells were used for training, while those generated from monocytes were used for testing. Notably, to avoid interference from unrelated protein features during the deconvolution task, only 39 differential proteins (marked by Leduc et al.[45]) across the three phases were used as input features.

PBMC CITE-seq data

One research group[51] generated CITE-seq data from PBMCs, including both transcriptomic information and cell surface protein profiles for CD4 T cells, CD8 T cells, B cells, myeloid cells and NK cells. The dataset is available for download from the GEO database (accession no. GSE253721). It contains a total of 43,791 cells, with each cell comprising 20,568 transcriptomic features and 205 proteomic features. Highly variable genes ($n = 1101$) were identified from the transcriptomic data using the Scanpy[79] toolkit, and only these genes were used in the DECODE experiments. Single-cell data from donor HS1 were used to generate pseudotissues for training, whereas data from donor HS5 were used to generate pseudotissues for testing.

Model

A schematic diagram of the model structure is shown in Fig. 1. The proposed framework has four stages. Stage 1 involves constructing pseudotissues by randomly sampling single cells from the single-cell data according to predefined proportions and tissue numbers. In stage 2, an adversarial training framework is introduced to minimize the feature distance in the latent space between training pseudotissues and the tissue data designated for deconvolution, aiming to mitigate batch effects. Stage 3 incorporates contrastive learning to guide the denoiser module, enabling it to denoise the data and estimate more accurate relative proportions of target cell types. Finally, in stage 4, a single forward inference process is implemented. The optimal pathway between two computational pathways is selected on the basis of the available single-cell data.

Stage 1

For any tissue-level data, we need to get single-cell data of the tissues for training DECODE. Assuming that the collected single-cell data contains n distinct cell types. A set of n random numbers, $\{r_1, r_2, \cdots, r_n\}$, is generated from a uniform distribution. These random numbers are normalized to obtain a set of cell proportions for the tissue, $P = \{p_1, p_2, \cdots., p_n\}$, as shown in equation (1):

$$p_i = \frac{r_i}{\sum_{j=1}^{n} r_j}, \sum_{i=1}^{n} p_i = 1, p_i \geq 0 \tag{1}$$

To generate the data for the pseudotissue, the total number of single cells within the pseudotissue is predefined as m. On the basis of the initial proportions $P = \{p_1, p_2, \cdots., p_n\}$, the number of cells to be sampled for each cell type, n_i, is calculated using equation (2):

$$n_i = \text{round}\,(p_i \times m), \sum_{i=1}^{n} n_i \approx m \tag{2}$$

Here round$(\cdot)$ denotes the rounding operation.

For each cell type i, n_i cells are randomly sampled from the collected single-cell data. If the number of cell types is less than n_i, repeated sampling will be performed. Let each cell in the single-cell data be represented as $\mathbf{x} \in \mathbb{R}^f$, where f denotes the dimensionality of single-cell features. Thus, for cell type i, the representation of the sampled cell set is given by equation (3):

$$X_i = \left\{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{n_i}^{(i)}\right\}, \mathbf{x}_k^{(i)} \in \mathbb{R}^f \tag{3}$$

Since n_i is approximated using the rounding operation during the sampling process, the actual sampled cell proportions deviate from P. Therefore, the cell proportion vector $\mathbf{P}$ needs to be updated according to equation (4):

$$p_i = \frac{n_i}{\sum_{j=1}^{n} n_j}, \sum_{i=1}^{n} p_i = 1, p_i \geq 0 \tag{4}$$

The feature vector of the pseudotissue, $\mathbf{X}_{\text{Pseudo}}$, is defined as the summation of feature vectors from all sampled cells, as shown in equation (5):

$$\mathbf{X}_{\text{Pseudo}} = \sum_{i=1}^{n} \sum_{k=1}^{n_i} \mathbf{x}_k^{(i)} \tag{5}$$

The above process is repeated U times to generate the training set X_{Train}, which contains U feature vectors of dimension f, representing the pseudotissue features. Each sample is paired with an updated cell proportion vector $\mathbf{P}$, which serves as the cell-type proportion label for that sample. Stage 1 has two outputs: X_{Train} and P_{Train}. The rules for setting the cellular capacity m of an individual pseudotissue and the capacity U of X_{Train} are presented in the section 'Pseudotissue capacity and quantity' in the Supplementary Information and Supplementary Fig. 15.

Stage 2

The objective of stage 2 training is to eliminate batch effects between the train-tissue and the target-tissue through iterative adversarial training. The train-tissue dataset is denoted as $X_{\text{Train}} \in \mathbb{R}^{U*f}$, whereas the target-tissue dataset includes two types (pseudotissue data and real tissue data), which are referred to as $X_{\text{Target}} \in \mathbb{R}^{T*f}$, where T represents the number of target-tissue samples.

First, X_{Train} and X_{Target} are sequentially processed through an encoder block (composed of a linear layer, leaky rectified linear unit (ReLU) activation function, layer normalization and dropout layer), mapping the original features to hidden representations, resulting in E_{Train} and $E_{\text{Target}} \in \mathbb{R}^{B*w}$, B is the batch size and w is the dimensionality of the embedded features.

The hidden representation E_{Train} is fed into both the discriminator (composed of a linear layer, leaky ReLU activation function, dropout layer and sigmoid activation function) and a shallow deconvolution module, eDeconvolver (composed of a linear layer, leaky ReLU activation function, dropout layer and softmax activation function), while E_{Target} is fed into the discriminator. $\mathbf{D}'_{\text{Train}} \in \mathbb{R}^U$ and $\mathbf{D}'_{\text{Target}} \in \mathbb{R}^T$ represent the discriminator's predictions for E_{Train} and E_{Target}, respectively, while $\hat{P}$ denotes the deconvolution result of eDeconvolver for E_{Train} The discriminator is a module designed to determine whether the input E originates from train-tissue or target-tissue. The sample labels for train-tissue are denoted as $\mathbf{D}_{\text{Train}} \in \mathbb{R}^U$, a vector of length U with all the elements equal to 1. Similarly, the sample labels for target-tissue are denoted as $\mathbf{D}_{\text{Target}} \in \mathbb{R}^T$, a vector of length T with all the elements equal to 0. To ensure that the features generated by the encoder are suitable for downstream deconvolution tasks, the training process for the encoder, discriminator and eDeconvolver is jointly supervised using the L1 loss function and the binary cross-entropy (BCE) loss function, as described in equations (6)–(8):

$$\text{Loss}_1 = \text{Loss}_{di} + \text{Loss}_{eD} \tag{6}$$

$$\text{Loss}_{di} = \text{Loss}_{\text{BCE}}(\mathbf{D}_{\text{Train}}, \mathbf{D'}_{\text{Train}}) + \text{Loss}_{\text{BCE}}(\mathbf{D}_{\text{Target}}, \mathbf{D}'_{\text{Target}}) \tag{7}$$

$$\text{Loss}_{eD} = \text{Loss}_{L1}(P, \hat{P}) \tag{8}$$

After backpropagating the aforementioned loss functions to update the model parameters, the encoder and discriminator parameters are further updated using the loss function defined in equation (9):

$$\text{Loss}_2 = \text{Loss}_{\text{BCE}}(\mathbf{D}_{\text{Train}}, \mathbf{D'}_{\text{Target}}) + \text{Loss}_{\text{BCE}}(\mathbf{D}_{\text{Target}}, \mathbf{D}'_{\text{Train}}) \tag{9}$$

At the end of each batch, two stages of loss calculation and two parameter updates are performed. After the completion of stage 2 training, the encoder parameters are then frozen and transferred to the encoder in stage 3.

Stage 3

In stage 3, artificial impurity cells are generated and mixed with train-tissue to form sample pairs that are jointly input into the model for computation. The process for generating artificial impurity cells is as follows: first, one cell is randomly sampled from each of the n cell types in the single-cell data, forming $X_{\text{sample}} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$, $X_{\text{sample}} \in \mathbb{R}^{n*f}$; then, a weighted mixing feature matrix $M\sqrt{ab} \in \mathbb{R}^{n*f}$ is randomly generated, where each column of M contains elements greater than 0, and the elements in each column sum to 1. The dot product between the corresponding columns of X_{sample} and M produces an artificial impurity cell. This process is repeated multiple times to form a collection of artificial impurity cells. During mixing, a random subset of artificial impurity cells is sampled, with the number of selected cells not exceeding $m \times 0.1$, where m is the cell capacity of the train-tissue. These sampled artificial impurity cells are added to X_{Train} to form $X_{(\text{Train\&noise1})}$ and $X_{(\text{Train\&noise2})}$, which together constitute the sample pair (X_{Train}, $X_{(\text{Train\&noise1})}$, $X_{\text{Train\& noise2}}$).

The sample pair is then fed into the encoder, and the obtained E_{Train}, $E_{(\text{Train\&noise1})}$ and $E_{\text{Train\& noise2}} \in \mathbb{R}^{B\times w}$ are processed by a function called DimExpander (composed of a linear layer and the reshape operation in Torch), which is designed to increase their dimensionality. This process yields $\text{Dim}_{\text{Train}}$, $\text{Dim}_{(\text{Train\&noise1})}$ and $\text{Dim}_{\text{Train\&noise2}} \in \mathbb{R}^{B\times C\times w}$, where C represents the newly expanded dimensionality. The purpose of this dimensionality enhancement is to project the original features to different solution spaces, enabling better learning and representation of complex feature relationships. The noise-mixed pair ($\text{Dim}_{\text{Train\&noise1}}$, $\text{Dim}_{\text{Train\&noise2}}$) is subsequently fed into the denoiser module built with a self-attention mechanism (Fig. 1e). After passing through the denoiser block, the purified train-tissue features (De_1, De_2) $\in \mathbb{R}^{B\times C\times w}$ and the artificial impurity cell features to be removed (No_1, No_2) $\in \mathbb{R}^{B\times C\times w}$ are obtained. Subsequently, (De_1, De_2) and $\text{Dim}_{\text{Train}}$ are fed into a linear attention layer (composed of a linear layer and the transpose operation in Torch), where the previously expanded dimensions are weighted and summed to integrate features from different solution spaces. Finally, the fused features are input into the deconvolution module (composed of a linear layer, LayerNorm, leaky ReLU activation function and softmax activation function) to obtain the final deconvolution results $\hat{P}_1$, $\hat{P}_2$ and $\hat{P}_3$. In this stage, the loss function 'Loss' (equations (10)–(12)) is used to supervise the entire training process.

$$\text{Loss} = \text{Loss}_1 + \text{Loss}_2 + \text{Loss}_3 + \text{Loss}_{\text{Denoiser}} \tag{10}$$

$$\text{Loss}_i = \text{Loss}_{L1}(P, \hat{P}_i) \tag{11}$$

$$\text{Loss}_{\text{Denoiser}} = \text{Loss}_{\text{NCE}}\left(\text{Dim}_{\text{Train}}^{\text{sample}}, (\text{De}_1^{\text{sample}}, \text{De}_2^{\text{sample}}), (\text{No}_1^{\text{sample}}, \text{No}_2^{\text{sample}})\right) \tag{12}$$

The purified train-tissue features (De_1, De_2) are derived from train-tissue data and mixed with artificial impurity cells processed through the encoder, DimExpander and denoiser. Theoretically, in the feature space, these features should be very close to $\text{Dim}_{\text{Train}}$. In contrast, the extracted artificial impurity cell mixture features (No_1, No_2) should theoretically be far from $\text{Dim}_{\text{Train}}$ in the feature space. To obtain a deconvolution model with denoising capabilities, a noise contrastive estimation loss function based on contrastive learning is adopted. This loss function minimizes the distance between positive samples while maximizing the distance between negative samples (Fig. 1f). Here, $\text{Dim}_{\text{Train}}^{\text{sample}}$, ($\text{De}_1^{\text{sample}}$, $\text{De}_2^{\text{sample}}$) and ($\text{No}_1^{\text{sample}}$, $\text{No}_2^{\text{sample}}$) $\in \mathbb{R}^{B*S*e}$ represent the sampled results of ($\text{Dim}_{\text{Train}}$, ($\text{De}_1$, De_2), No_1, No_2), respectively, with the sampling process completed by the Contrastive(·) function.

$$\text{Contrastive} = \text{MLP}\left(\text{Sample}\,(\cdot)\right) \tag{13}$$

Here Sample (·) represents randomly selecting S points from the same positions in the feature sets ($\text{Dim}_{\text{Train}}$, ($\text{De}_1$, De_2), No_1, No_2). MLP(·) denotes a dimensional transformation function applied to the sampled points, where e represents the dimensionality after the transformation. Let ($\text{Dim}_{\text{Train}}^j$, De_1^j), $j \in [0, S]$ represent one sampling point from the S samples and No_1^S represent the S sampled points. Here ($\text{Dim}_{\text{Train}}^j$, De_1^j) represents a positive sample pair, whereas $\text{Dim}_{\text{Train}}^j$ and all S sampled points from No_1^S are used to form negative samples. This setup defines a multiclass classification task, which is supervised via the noise contrastive estimation loss. This loss function draws positive samples closer while pushing negative samples farther apart. This sampling strategy also increases the numbers of both positive and negative samples, helping the model learn more diverse and comprehensive features.

Stage 4

Stage 4 outlines the inference process of the model, which involves two computational paths. If the cell types in the single-cell data cannot fully cover the potential cell types present in the tissue, the data must be passed through the denoiser block after the encoder block for

processing. Otherwise, no denoiser processing is needed. The specific computational workflow is outlined in equations (14)–(17):

$$E_{\text{Target}} = \text{Encoder}\left(X_{\text{Target}}\right) \tag{14}$$

$$\text{Dim}_{\text{Target}} = \text{DimExpander}(E_{\text{Target}}) \tag{15}$$

$$D = \begin{cases} \& \text{Dim}_{\text{Target}} & \text{if TissueCellType} \subseteq \text{SingleCellDataCellType} \\ \& \text{Denoiser}(\text{Dim}_{\text{Target}}) & \text{otherwise} \end{cases} \tag{16}$$

$$\hat{P} = \text{Deconvolver}\left(\text{LinearAttention}\left(D\right)\right) \tag{17}$$

Method comparison

In the DECODE experiment, a comparison was made with 12 state-of-the-art methods, which are TAPE, CIBERSORTx, MuSiC, scpDeconv, Scaden, RCTD, Seurat, SPOTlight, Tangram, ucdselect and cell2location. Their specific descriptions and usage methods can be found in the section 'Details of comparison methods' in the Supplementary Supplementary Information. Code examples for running the aforementioned methods can be found in Zenodo (https://doi.org/10.5281/zenodo.15682763)[82].

Evaluation metrics

To evaluate the linear consistency between the true proportions of cell types or states and the model's predicted results, we adopted the following three evaluation metrics: CCC, r.m.s.e. and r. The specific formulas are shown in equations (18)–(20).

$$\text{CCC}(y,\hat{y}) = \frac{2r\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \tag{18}$$

$$\text{r.m.s.e.}(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

$$r(y,\hat{y}) = \frac{\text{cov}(y,\hat{y})}{\sigma_y\sigma_{\hat{y}}} \tag{20}$$

Here y represents the true proportions of cell types (or states), and $\hat{y}$ represents the predicted proportions of cell types (or states) produced by the model. σ_y and $\sigma_{\hat{y}}$ denote the standard deviations of y and $\hat{y}$, respectively. $\text{cov}(y,\hat{y})$ indicates the covariance between y and $\hat{y}$. σ_y and $\sigma_{\hat{y}}$ are the means of y and $\hat{y}$, respectively, and n represents the sample size. These metrics were collectively used to evaluate the consistency and error magnitude between the predicted results and the ground truth.

During the data analysis process, Kendall correlation (τ), CV, KL divergence (D_{KL}) and Spearman correlation coefficient were also used. The specific formulas are shown in equations (21)–(24).

$$\tau = \frac{(C-D)}{\frac{1}{2}n(n-1)} \tag{21}$$

$$\text{CV} = \frac{\sigma}{\mu} \tag{22}$$

$$D_{\text{KL}}(P \parallel Q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \tag{23}$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{24}$$

Here τ is a statistical measure used to assess the correlation between two variables, primarily evaluating the consistency of their rankings. C is the number of concordant pairs of observations, D is the number of discordant pairs of observations and n is the sample size. σ is the standard deviation of the sample or population and μ is the mean of the sample or population. **P** and **Q** represent probability distribution vectors, where p_i and q_i denote the probabilities of the two vectors at the ith element, respectively. ρ represents Spearman's rank correlation coefficient. Assume there are n samples, where x_i and y_i represent the values of the ith sample for two variables. Ranks $R(x_i)$ and $R(y_i)$ represent the ascending order of samples x_i and y_i, respectively. The rank difference for the ith sample between the two variables is represented by the formula $d_i = R(x_i) - R(y_i)$. We used three metrics: cosine similarity, Kendall similarity and Spearman similarity, due to their applicability under different data characteristics; cosine similarity is suitable for high-dimensional and sparse data and is used to evaluate the similarity of transcriptomic cell types, Kendall similarity is robust to outliers and sensitive to subtle differences, making it suitable for comparing highly similar cell types, while Spearman similarity is better for assessing trend consistency and is therefore used to analyze the consistency of DECODE's predictions across pseudotissue datasets.

System configuration

In this study, the configuration is as follows: the operating system is Ubuntu v.20.04, Python v.3.8, the deep learning framework is PyTorch v.2.0.0 and the CUDA v.11.8. In terms of hardware, the setup includes an NVIDIA RTX 4090 GPU (24GB) and an Intel(R) Xeon(R) Gold 6430 CPU with 16 vCPUs.

Data analysis

Highly variable genes, differential proteins and differential metabolites were calculated using the Scanpy[79] program. Similarity calculations, Spearman correlation and KL divergence were performed using tools from SciPy[83]. We used the ttest_ind and wilcoxon functions from the stats module in SciPy to conduct the t-test and Wilcoxon test, respectively. We applied the Bonferroni method for P value correction.

Reporting summary

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

Data availability

The data used in this study are all publicly available, and their usage methods are discussed in detail in the Methods. The scenario 1 dataset is available via the Human Cell Atlas at https://www.covid19cellatlas.org/index.healthy.html#publication. The scenario 2 dataset is available via the GEO database at Dataset ID GSE176078. The scenario 3 dataset is available via Mendeley at https://doi.org/10.17632/vs8m5gkyfn.1. The scenario 4 (murine cell line) datasets, including the datasets from ref. 26 (accession no. MSV000086809) and ref. 27 (accession no. MSV000084110), were both downloaded from the MassIVE data repository. The scenario 4 (islet) dataset includes the datasets from refs. 24 and [25], both of which can be downloaded from GSE211799. Scenario 5 datasets include those from refs. 30,80. The spatial dataset from ref. 80 can be obtained via GitHub at https://github.com/QuKunLab/SpatialBenchmarking/tree/main/FigureData/Figure4/Dataset10_STARmap/Rawdata, while the single-cell data can be downloaded via their Google cloud folder at https://drive.google.com/drive/folders/1pHmE9cg_tMcouV1LFJFtbyBJNp7oQo9J. The dataset from ref. 30 can be downloaded using the squidpy.datasets.slideseqv2() function. The dataset used in scenario 6 is available via cellxgene at https://cellxgene.cziscience.com/collections/2f4c738f-e2f3-4553-9db2-0582a38ea4dc. The dataset used in scenario 7 comes from the tutorials available via TAPE at https://sctape.readthedocs.io/datasets/#pbmc-datasets (ref. 15). Bone marrow single-cell metabolomics data are available via the Metabolomics Workbench at https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Project&ProjectID=PR001858

(project ID PR001858). The liver single-cell metabolomics data are available via the project's official GitHub repository at https://github.com/yuanzhiyuan/SEAM/tree/master/SEAM/data/raw_tar. Monocyte transcriptomics data are available via Zenodo at https://doi.org/10.5281/zenodo.15682763 (ref. 82). Melanoma drug response metabolomics data are available via *Nature* in the source data for ref. 46 at https://www.nature.com/articles/s41467-020-15956-9#Sec32. Melanoma and monocyte cell cycle proteomics data are available via the Slavov Laboratory at https://scp.slavovlab.net/Leduc_et_al_2022 (ref. 45). The dataset of PBMC CITE-seq data can be downloaded from the GEO database (Dataset ID GSE253721). Tissue-level proteomics data for breast cancer multiomics analysis are available in the Data Commons under accession number PDC000120. The tissue-level transcriptomics data used for the analysis are available via the GEO database (GSE245467, GSE184869, GSE161865). For the mouse liver multiomics analysis, the metabolomics data are available via ref. 64, the proteomics data are available via the MassIVE database (MSV000092153) and the transcriptomics data are available via the GEO database (GSE196941, GSE200356, GSE222550, GSE243906, GSE253217, GSE256501, GSE267916, GSE269058). Source data are provided with this paper.

Code availability

The complete user manual for DECODE has been published on GitHub[84], and the code for reproducing the DECODE experiments is available on DECODE Experiment Records[85].

References

79. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
80. Li, B. et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods* **19**, 662–670 (2022).
81. Oetjen, K. A. et al. Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *JCI Insight* **3**, e124928 (2018).
82. Liu, R. DECODE comparative model usage demo. *Zenodo* https://doi.org/10.5281/zenodo.15682763 (2025).
83. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
84. forceworker. DECODE. *GitHub* https://github.com/forceworker/DECODE (2025).
85. renjie, liu. DECODE Experiment Records. *Zenodo* https://doi.org/10.5281/zenodo.17045605 (2025).

Acknowledgements

We thank X. Li for the support and academic exchanges provided during the preparation of this article. We also thank all personnel involved in generating and maintaining the datasets used in the DECODE experiments; their careful data preservation has been essential for enabling more meaningful methodological advances. This work was supported by the National Natural Science Foundation of China (grant no. 62572148) to T.Z. and the National Natural Science Foundation of China (No. 62222104) to L.C. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper. T.Z. was supported by the National Natural Science Foundation of China (grant no. 62572148). L.C. was supported by the National Natural Science Foundation of China (grant no. 62222104).

Author contributions

T.Z., R. Liu and Y.S. conceived and designed the project. R. Liu and Y.S. performed data processing, model development and result evaluation. L.C. and Y.W. supervised the execution. The liver metabolic data were collected by Q.C. B.W. assisted the lead author in completing the performance testing of the spatial transcriptomics method. L.Z., G.W., R. Luo and Z.Y. helped revise the paper. All authors have read and approved the paper.

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/s41592-026-03007-y.

Correspondence and requests for materials should be addressed to Liang Cheng or Yadong Wang.

Peer review information *Nature Methods* thanks Jiangning Song and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Arunima Singh, in collaboration with the *Nature Methods* team.

nature portfolio

Corresponding author(s): Yadong Wang

Last updated by author(s): Nov 19, 2025

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	scanpy==1.9.8
Data analysis	https://github.com/forceworker/DECODE https://zenodo.org/records/17045605 https://zenodo.org/records/15682763 Public software:scipy==1.10.1,torch==2.0.0+cu118,numpy==1.24.2,scanpy==1.9.8,scikit-learn==1.3.2,pandas==2.0.3

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

The data used in this study are all publicly available, and their usage methods are discussed in detail in the "Methods" section. The dataset of Scenario 1 can be

downloaded from https://www.covid19cellatlas.org/index.healthy.html#publication. The dataset of Scenario 2 can be downloaded from the GEO database (Dataset ID: GSE176078). The dataset of Scenario 3 can be found at Brugge_Breast_Atlas_CyTOF_Primary_Processed - Mendeley Data. The Scenario 4 (murine cell line) datasets, including the Woo et al. dataset (accession no. MSV000086809) and the Dou et al. dataset (accession no. MSV000084110), were both downloaded from the MassIVE data repository. The Scene 4 (islet) dataset includes the Oppenländer et al. dataset and the Tritschler et al. dataset, both of which can be downloaded from GSE211799. Scene 5 data sets include those from Li et al. and Palla et al. The spatial data set from Li et al. can be obtained from their GitHub open-source platform, while the single-cell data can be downloaded from their provided cloud storage. The data set from Palla et al. can be downloaded using the squidpy.datasets.slideseqv2() function. The dataset used in Scene 6 can be found on cellxgene. The dataset used in Scene 7 can be downloaded from the tutorials provided by TAPE12. Bone marrow single-cell metabolomics data is available on the Metabolomics Workbench (project ID PR001858). The liver single-cell metabolomics data can be found in the project's official GitHub repository. Monocyte transcriptomics data is available for download from Zenodo (15686257). Melanoma drug response metabolomics data can be found in the source data of their paper. Melanoma and monocyte cell cycle proteomics data can be found at Leduc et al, 2022 | Single-cell proteomics by mass spectrometry | Slavov Laboratory. The dataset of PBMC CITE-seq data can be downloaded from the GEO database (Dataset ID: GSE253721). Tissue-level proteomics data for breast cancer multiomics analysis are available in the Proteomic Data Commons under accession number PDC000120. The tissue-level transcriptomics data used for the analysis can be found in the GEO database (GSE245467, GSE184869, GSE161865). For the mouse liver multiomics analysis, the metabolomics data are available in the supplementary files of the article, the proteomics data are in the MassIVE database (MSV000092153), and the transcriptomics data are in the GEO database (GSE196941, GSE200356, GSE222550, GSE243906, GSE253217, GSE256501, GSE267916, GSE269058).

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	Not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

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	In this study, we investigated the performance of multiple single-cell transcriptomics and proteomics datasets, including the lung single-cell transcriptomics dataset with 12,971 cells and 3,346 high-variability genes, the breast single-cell transcriptomics dataset with 100,064 cells and 2,525 high-variability genes, the breast single-cell proteomics dataset with 751,970 cells and 34 protein features, the mouse single-cell proteomics dataset with 108 cells featuring 1,437 proteins from Woo and 1,032 from Dou, with 762 overlapping features, the mouse islet datasets measured by Oppenländer et al. and Tritschler et al. contain 31,706 available transcriptomic, the STARmap data contains 882 transcriptomic features from 1,523 cells, while the Smart-seq2 data includes 34,042 transcriptomic features from 14,249 cells, the hippocampal dataset contains 17,733 transcriptomic features from 41,786 cells, the Fovea and Periphery datasets contain 34,723 and 19,768 cells, respectively, focusing primarily on the following 17 cell types, the real tissue dataset with 11,328 RNA features, the marrow bone, liver, and human colorectal cancer contain 1,428, 724, and 57,078 cells, along with 107, 244, and 112 metabolic features, the monocyte transcriptomics dataset with 10,846 cells and 2,000 high-variability genes, the melanoma drug response metabolomics dataset with 674 cells, 20 metabolites, and 4 cell states, the melanoma and monocyte cell cycle proteomics dataset with 1,573 cells and 2,844 protein features, and the PBMC CITE-seq dataset with 43,791 cells and 1,101 high-variability genes.
Data exclusions	In some datasets, we only used donors with specific phenotypes, which is clearly explained in the text.
Replication	This study did not involve new wet-lab experiments; all analyses were performed on existing datasets, for which biological and technical replication were defined by the original studies. Computational analyses can be reproduced using the released code, trained model parameters and prediction results available on GitHub.
Randomization	Randomization of participants or animals to experimental groups was not applicable, because this study only re-analyses observational omics datasets. For model training, pseudo-tissue samples were generated by random sampling of single cells as described in the Methods.

Blinding	Blinding of investigators to group allocation was not applicable. All outcomes (for example, prediction errors and correlation metrics) were computed automatically by software from de-identified data without manual scoring.

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	Not applicable.
Novel plant genotypes	Not applicable.
Authentication	Not applicable.